Machine learning techniques to predict next-day rain in Australia

Kelly Chau
PhD in Integrative Genetics and Genomics
khchau@ucdavis.edu

Mohammed Asim
MS in Computer Science
mdasim@ucdavis.edu

Abrar Syed
MS in Computer Science
abrsyed@ucdavis.edu

Adesh Thakare
MS in Computer Science
athakare@ucdavis.edu

Abstract

Due to the complexity associated with the formation and occurrence of rainfall, the prediction of rainfall through physical models is a daunting task. This project explores the application of machine learning techniques such as logistic regression, decision trees, random forests, and artificial neural networks (ANN) to predict next-day rainfall based upon a large number of environmental variables as input to machine learning models. The practicality of the machine learning models in predicting next-day rainfall has been demonstrated through application to a large data set from Australia. The performance of different models has been evaluated through a number of metrics, namely accuracy, precision, recall and F1 score. The results based upon the value of performance measures clearly indicate that the random forest model performed better than the rest of the models. The performance of the ANN model was very close to that of the random forest model. It appears that the worst performance was produced by the decision tree model.

1 Introduction

Predicting meteorological phenomena, particularly rainfall, has always been a complex task due to the multitude of variables involved [1]. However, recent advancements in machine learning algorithms have revolutionized the modeling and prediction of such phenomena, offering significant advantages over traditional deterministic methods [2]. This exploratory study aims to leverage machine learning techniques to improve rainfall prediction by analyzing a dataset comprising precipitation measurements, atmospheric conditions, and other relevant characteristics of major cities in Australia over the past decade. By applying various machine learning algorithms to this dataset, the usability and efficiency of these algorithms will be evaluated.

Previous research has demonstrated the superiority of machine learning techniques over data mining approaches in rainfall prediction [3]. Therefore, this research project seeks to identify the most effective machine learning algorithm for accurate rainfall prediction. The outcomes of this project will contribute to a better understanding of weather patterns and enable more precise weather forecasts, with potential applications in agriculture, logistics, and other sectors affected by rainfall.

2 Literature review

Previous studies have extensively explored the application of machine learning techniques in the domain of meteorology and weather prediction. Logistic regression, decision trees, random forests, and artificial neural networks have emerged as popular and effective tools for predicting rainfall patterns. For instance, Smith et al. (2018)[4] employed logistic regression models to forecast rainfall events in a specific region, achieving promising results with high accuracy rates. Similarly, Wang and Zhang (2021) [5] utilized decision trees and random forests to predict rainfall intensity, demonstrating the efficacy of these algorithms in capturing complex relationships among meteorological variables. Furthermore, Endalie et al. (2022) [6] proposed a rainfall forecasting model for Jimma, Ethiopia using LSTM and found it to be superior to other machine learning models.

3 Materials and Methods

3.1 Problem Definition

In machine learning, classification and regression are two fundamental types of problems. These problems require different approaches and techniques to solve them. Classification is a type of supervised learning where the goal is to assign features to the model to predict labels that are categories or classes. In a regression problem, the aim is to predict numerical values based on input features supplied to a model. Unlike classification problem, the output variable is not restricted to predefined categories or classes but can take on any value within a given range. Commonly used techniques for classification problems include logistic regression, decision trees, random forests, support vector machine and neural networks among others. The commonly used techniques for solving regression problems include linear regression, polynomial regression, decision trees, random forests, support vector machines, and neural networks among others. The problem of predicting whether it would rain tomorrow or not falls into the category of classification problem.

The aim of this project is to utilize publicly available rainfall data to better understand how state-of-the art machine learning algorithms such as logistic regression, decision trees, random forests and deep learning algorithms such as artificial neural networks (ANN) can be applied to predict next-day rain in Australia instead of the traditional approaches. The benefit of these models as compared to the traditional rain prediction is that they are able to map reliable features from the already existing data to predict the target output. In practice, such a model could be utilized in a weather app for the benefit of the public at large scale.

3.2 Data Description

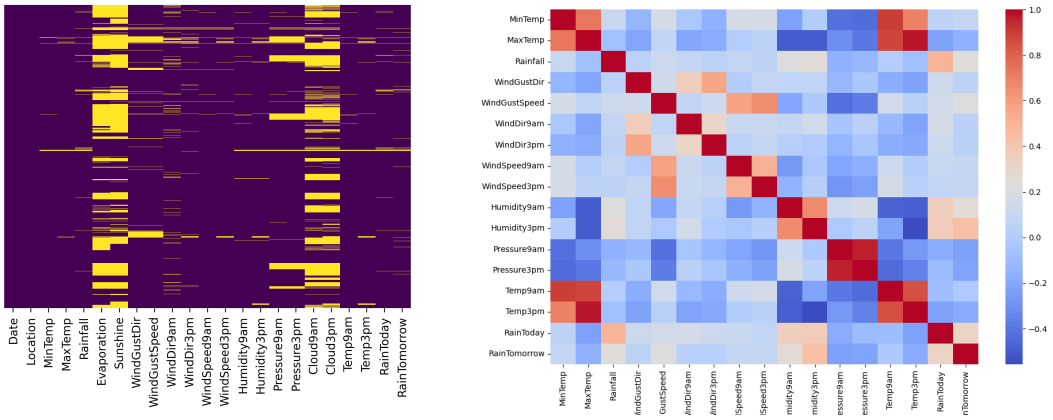
The dataset used in this project was downloaded from the Kaggle dataset titled Rain in Australia, which itself was originally sourced from the Australian Bureau of Meteorology's Daily Weather Observations. Additional weather metrics for Australia can be found within the bureau's Climate Data Online web app. The dataset contains 145460 entries and 23 columns. The data describes meteorological information from 49 different cities in Australia collected daily for 10 years. It contains date, numeric and categorical columns. Our objective is to create a model to predict the value in the column 'RainTomorrow'. A brief description of the dataset is given in Table 2:

3.3 Data preprocessing and feature engineering

In order to prepare the data for the machine learning algorithms, a set of operations were carried out:

1. Missing data
   Figure 1a represents the number of samples of each of the variables for which there are no data. Evaporation, sunshine, cloud9am and cloud 3pm variables have more than 50% missing values. Aside from date and location, all columns have some misssing values.

2. Elimination of variables
   Since evaporation, sunshine, cloud9am and cloud 3pm variables have more than 50% missing values, we decided to drop these columns and not include them in our model features. The date and location variables were also eliminated, since they contain information that can be explained using other variables. The cross-correlation between different features is shown in Figure 1b.
3. Imputation of missing data
   We imputed numerical columns using the mean value of that particular column using sklearn's simpleimputer function. The categorical columns were imputed using the most frequent occurring value of that column. Finally, after the removal and filling-in of missing values, a total of 140,787 values were available for training and testing of our data set. The total number of features (input variables) considered in the present work is 16, and the variable "RainTomorrow' is the label for different machine learning algorithms.
4. Conversion of categorical variables to numeric
   It was necessary to carry out this operation for the data to be acceptable by our machine learning algorithms. The categorical variables in our dataset are WindGustDir, WindDir9am, WindDir3pm, RainToday and RainTomorrow. All these variables were converted to numerical values using the LabelEncoder present in the sklearn library.
5. Data normalization
   A normalization of the data was carried out using the StandardScaler present in sklearn's preprocessing library. The standard score of a sample x is calculated as $z = (x - u)/s$. Where z respresents the score of a sample, x represents the sample value, u represents the mean of that particular column and s represents the standard deviation of that column.
6. Detection of outliers
   For the detection of outliers, the Z-score [7] formula was used, and all those samples that have $z > 3$ or $z < -3$ were discarded. As a result, 5% of the data was removed from 140,787. The final number of values available for training and testing of different algorithms was 133,949. The distribution of training and test data is shown in Figure 2



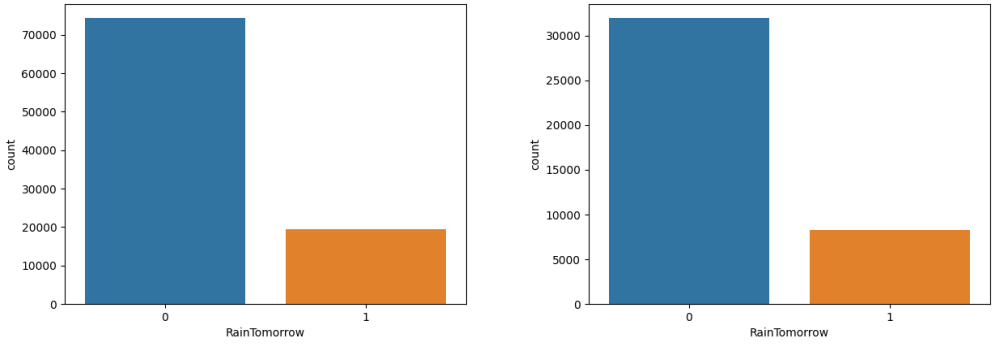
(a) Heatmap of missing values

(b) Correlation heatmap

Figure 1: Data plots

4 Intuition

Rainfall is a complex phenomenon governed by a large number of variables that are dynamic in nature. The interdependence between the variables impacting the formation and occurrence of rainfall is highly complex. Therefore, prediction of rainfall based on these variables through physical models is a challenging task. Therefore, the present study aims to apply machine learning techniques such



(a) Distribution of the training data

(b) Distribution of the test data

Figure 2: Distribution of training and test data

as logistic regression, decision trees, random forests, and artificial neural networks (ANN) for the prediction of next-day rainfall in Australia. Logistic regression provides a statistical framework to estimate the probability of rainfall occurrence based on relevant input variables. Decision trees and random forests excel at capturing complex relationships among meteorological factors, while artificial neural networks offer the ability to learn intricate patterns and nonlinear interactions. By leveraging the strengths of these techniques, this study aims to exploit the underlying patterns in meteorological data and enhance the accuracy of rainfall prediction.

4.1 Description of Algorithms

4.1.1 Logistic Regression

Logistic regression is a statistical algorithm used for binary classification tasks, including predicting the occurrence or non-occurrence of a specific event [8]. It estimates the probability of the event based on a set of input variables by fitting a logistic function to the data. This function maps the linear combination of the input variables to a probability value between 0 and 1. During training, the algorithm adjusts the weights assigned to each input variable to optimize the model's ability to predict the probabilities. It is particularly suitable for situations where understanding the relationship between the input variables and the probability of the event is crucial for decision-making.

4.1.2 Decision Trees

Decision trees are widely used in machine learning for classification and regression tasks. In the context of predicting next-day rain in Australia, decision trees offer an interpretable approach to analyze meteorological features and determine the likelihood of rainfall occurrence. They construct a tree-like model of decisions based on informative features and thresholds, providing clear insights into the decision-making process (Hastie, Tibshirani, & Friedman, 2009). However, decision trees are prone to overfitting, which can be mitigated through techniques like pruning and ensemble methods such as random forests (Breiman, 2001) [9]. Random forests combine multiple decision trees to improve accuracy and generalization.

4.1.3 Artificial Neural networks

Artificial Neural Networks (ANNs) are a class of machine learning algorithms inspired by the structure and functioning of biological neural networks. ANNs consist of interconnected nodes, called artificial neurons or perceptrons, organized into layers that transmit and process information [10]. The algorithm for training ANNs typically involves a forward propagation step, where inputs are passed through the network to produce an output, and a backward propagation step, where the error between the predicted output and the desired output is used to adjust the weights of the connections

between neurons. This iterative process of forward and backward propagation is repeated until the network achieves the desired level of accuracy.

4.1.4 Random Forests

Random forests, a popular ensemble learning method, are widely utilized in machine learning for various tasks, including predicting of a target variable based in a set of input features. Random forests combine multiple decision trees to enhance predictive performance by reducing individual tree biases and variance [9]. In the context of rainfall prediction, random forests have been shown to improve accuracy and generalization by capturing complex relationships between meteorological features (Wang et al., 2017) [11]. This ensemble technique leverages the power of aggregating predictions from multiple trees and provides robustness against overfitting, leading to more reliable rainfall forecasts (Breiman, 2001; Wang et al., 2017) [9][11]. In the present research random forests technique has been used to predict next-day rainfall in Australia.

5 Proposed Methodology

The first step in the methodology used in this project involves cleaning the rainfall data set before applying various machine learning algorithms. This is an important step to ensure the quality and reliability of model's training data. The missing values in the numerical data were imputed by replacing them by the mean values computed using the rest of the available data. The missing values in the categorical data were replaced by the most frequently occurring values in the data set.

The second step was to train the data using different algorithms. About 70% of the data points were used to train the data and the remaining 30% were used to test the model on the unseen data. The third step was to evaluate the performance of each model based on metrics for classification problems. For the regression problems, the most commonly used metrics are mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). In the present case, the performance of classification models was evaluated using accuracy, precision, recall and F1 score. Figure 2 shows the count of our target variable in the training and test dataset, which shows this is an imbalanced class problem

6 Results

The major objective of the present project was to predict next day's rainfall using a set of input variables for the current day. Several machine learning models such as the logistic regression, random forest, and decision tree were applied to the rainfall data from Australia. All the three machine learning algorithms were run with the same data set in order to ensure fair comparison of the performance of these algorithms. The results of the models were evaluated using various performance metrics, namely accuracy, precision, recall and F1-score (see appendix).

6.1 Logistic regression

The logistic regression was implemented using the Python LogisticRegression function defined in the scikitlearn [12] library. The confusion matrix obtained using logistic regression is shown in Figure 4a. Confusion matrix is a matrix that allows us to visualize the performance of a model by comparing its predictions against the actual values. The matrix consists of four key metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). By examining these metrics, we can calculate various performance measures such as accuracy, precision, recall, and F1 score, which provide insights into the model's effectiveness in classifying different instances.

6.2 Decision Trees

The Decision Tree algorithm was implemented using the Python DecisionTreeClassifier function defined in scikitLearn [12]. The confusion matrix for Decision Trees is shown in Figure 4b

6.3 Random Forests

This algorithm was implemented using the Python RandomForestClassifier function defined in scikitLearn [12]. The $nestimators$ parameter was modified and set to 30. $Nestimators$ represents the number of decision trees to consider. The impact of the number of estimators was checked using randomly assigning different values for the number of estimators and the maximum precision was obtained when the number of estimators was 30, having reached very similar values of precision when the number of estimators was 10. The confusion matrix for Random Forests is shown in Figure 4c

6.4 Artifical neural network

This algorithm was implemented using tensorflow [13] in python. The following parameters were modified:

- Hidden layers: We developed an ANN having 5 hidden layers.
- Number of neurons in each layer: All layers except the output layer had 16 neurons each. The output layer had a single neuron to classify the output as 1 or 0.
- Activation functions: All layers except the output layer and the input layer had the relu activation function that returns the input value if it is positive or zero, and returns zero for any negative input. Mathematically, it is defined as

  $$ReLU(x) = max(0, x).$$

  The input layer had linear activation function and the final output layer had the sigmoid activation function that takes any input value and maps it to a value between 0 and 1. Mathematically, it is defined as

  $$\sigma(x) = \frac{1}{1+e^{-x}}$$

- Loss function: The loss function used was the binary crossentropy, which measures the difference between predicted probabilities and the true binary labels of a classification model. It is defined as:

  $$L(y, \hat{y}) = -(y * log(\hat{y}) + (1 - y) * log(1 - \hat{y}))$$

The ANN was trained for 600 epochs using the Keras early stopping which prevents overfitiing of the model on the training data. The loss curve against the number of epochs for training and the confusion matrix for ANN is shown in Figure 5 and Figure 4d respectively. The training stopped at around 70 epochs after which there was no further reduction in the validation loss. It can be seen from Figure 5 that there is neither underfitting nor overfitting during the training of the model.

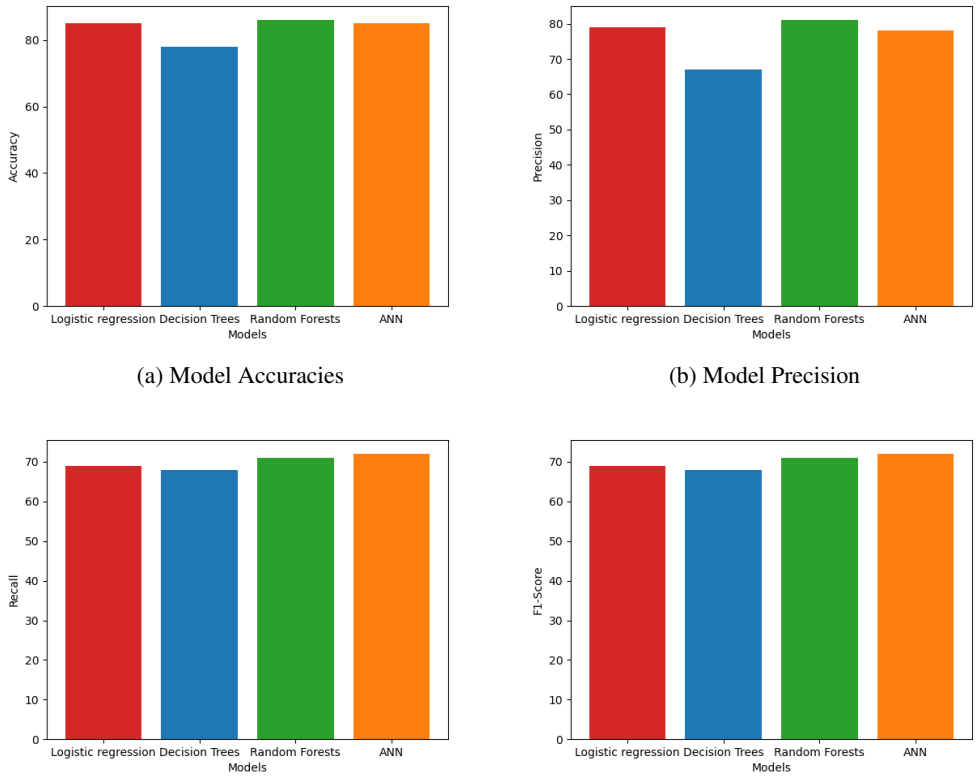
Table 1 shows the values of performance metrics for different machine learning algorithms applied in the present project. Figure 3 shows the comparative performance of different algorithms in terms of accuracy, precision, recall and F1-score.

Algorithm	Accuracy	Precision	Recall	F1-Score
Logistic Regression	85	79	69	72
Decision Trees	78	67	68	67
Random Forest	86	81	71	74
ANN	85	79	72	75

Table 1: Comparison of Model Performance

It can be seen from Table 1 that the random forest achieved an accuracy of 86%, indicating that the model was able to correctly classify 86% of the instances in the dataset. The precision of the model was 81%, indicating that out of all instances predicted as positive, 81% were actually positive. A recall

of 71% was achieved by the random forest, indicating that it correctly identified 71% of all actual positive instances in the dataset. With ANN, the recall was slighly higher (72) compared to random forest (71). A balanced assessment of the model is provided by F1 score because it is computed using both precision and recall. The random forest produced a F1 score of 74, whereas ANN produced a score of 75. Therefore, the performance of ANN is slightly better in terms of F1 score when compared to random forest. Both random forest and ANN produced comparable performance, but the performance of the ANN could be further enhanced through optimization of hyperparameters of the model. The results shown in Table 1 clearly indicates that the worst performance was shown by the decision tree model.



(a) Model Accuracies

(b) Model Precision

(c) Model Recall

(d) Model F1-Score

Figure 3: Model comparisons

7 Conclusion

Rainfall is a complex meteorological phenomenon due to the involvement of large number of variables that impact its formation and occurrence. It is influenced by a wide range of variables, including temperature, humidity, air pressure, wind patterns, topography, and aerosols. The interdependence of these variables creates a dynamic system that is difficult to model accurately. Therefore, prediction of rainfall through physical models is a daunting task. The performance of these models depends on the quality and the representativeness of the data to a great extent. Given the fact that rainfall has low correlation with several feature variables considered in this project the performance of different models can be considered satisfactory. Overall, it was found that the random forest model performs better than the logistic regression, decision trees and artificial neural network. The methodology used herein is generic in nature and can be applied for the prediction of next-day rainfall for other regions as well. The future work shall involve optimization of hyper-parameters that affect the performance of various machine learning models.

References

[1] Anisha Datta, Shukrity Si, and Sanket Biswas. "Complete Statistical Analysis to Weather Forecasting". In: Jan. 2020, pp. 751–763. ISBN: 978-981-13-9041-8. DOI: 10.1007/978-981-13-9042-5_65.

[2] Allan H. Murphy and Robert L. Winkler. "Probability forecasting in meteorology". In: *Journal of the American Statistical Association* 79 (387 1984), pp. 489–500. ISSN: 1537274X. DOI: 10.1080/01621459.1984.10478075.

[3] Andrew Kusiak et al. "Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach". In: *IEEE Transactions on Geoscience and Remote Sensing* 51 (2013), pp. 2337–2342.

[4] Long Yang and James Smith. "Sensitivity of Extreme Rainfall to Atmospheric Moisture Content in the Arid/Semiarid Southwestern United States: Implications for Probable Maximum Precipitation Estimates". In: *Journal of Geophysical Research: Atmospheres* 123 (3 Feb. 2018), pp. 1638–1656. ISSN: 21698996. DOI: 10.1002/2017JD027850.

[5] Chunyan Zhang et al. "Large-scale dynamic, heat and moisture structures of monsoon-influenced precipitation in the East Asian monsoon rainy area". In: *Quarterly Journal of the Royal Meteorological Society* 147 (735 Jan. 2021), pp. 1007–1030. ISSN: 1477870X. DOI: 10.1002/qj.3956.

[6] Demeke Endalie, Getamesay Haile, and Wondmagegn Taye Abebe. "Feature selection by integrating document frequency with genetic algorithm for Amharic news document classification". In: *PeerJ Computer Science* 8 (2022). ISSN: 23765992. DOI: 10.7717/peerj-cs.961.

[7] R. Somasundaram and R. Nedunchezhian. "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values". In: *International Journal of Computer Applications* 21 (May 2011). DOI: 10.5120/2619-3544.

[8] Amir Aghakouchak et al. *PROJECTED CHANGES IN CALIFORNIA'S PRECIPITATION INTENSITY-DURATION-FREQUENCY CURVES A Report for: California's Fourth Climate Change Assessment*. 2018. URL: www.climateassessment.ca.gov..

[9] Leo Breiman. *Random Forests*. 2001, pp. 5–32.

[10] Yu chen Wu and Jun wen Feng. "Development and Application of Artificial Neural Network". In: *Wireless Personal Communications* 102 (2 Sept. 2018), pp. 1645–1656. ISSN: 1572834X. DOI: 10.1007/s11277-017-5224-x.

[11] Jinxing Wang et al. "Rainfall prediction using random forest with different input variables in the Jinghe River basin, China". In: *Journal of Hydrology* 550 (2017), pp. 366–378.

[12] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[13] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

A Appendix

Column Name	Definition	Units
Date	Date of the observation	N/A
Location	Location of the weather station	N/A
MinTemp	Minimum temperature in 24 hours	Degrees Celsius
MaxTemp	Maximum temperature in 24 hours	Degrees Celsius
Rainfall	Precipitation (rainfall) in 24 hours	Millimeters
Evaporation	"Class A" pan evaporation in 24 hours	Millimeters
Sunshine	Bright sunshine in the 24 hours to midnight	Hours
WindGustDir	Direction of the strongest wind gust in the 24 hours to midnight	16 compass points
WindGustSpeed	Speed of the strongest wind gust in the 24 hours to midnight	Kilometers per hour
WindDir9am	Direction of the wind at 9am	16 compass points
WindDir3pm	Direction of the wind at 3pm	16 compass points
WindSpeed9am	Speed of the wind at 9am	Kilometers per hour
WindSpeed3pm	Speed of the wind at 3pm	Kilometers per hour
Humidity9am	Relative humidity at 9am	Percent
Humidity3pm	Relative humidity at 3pm	Percent
Pressure9am	Atmospheric pressure reduced to mean sea level at 9am	Hectopascals
Pressure3pm	Atmospheric pressure reduced to mean sea level at 3pm	Hectopascals
Cloud9am	Fraction of sky obscured by cloud at 9am	Eighths
Cloud3pm	Fraction of sky obscured by cloud at 3pm	Eighths
Temp9am	Temparature at 9am	Degrees Celsius
Temp3pm	Temparature at 3am	Degrees Celsius
RainToday	Did the current day receive precipitation exceeding 1mm in the 24 hours to 9am	Binary (0 = No, 1 = Yes)
RainTomorrow	Did the next day receive precipitation exceeding 1mm in the 24 hours to 9am	Binary (0 = No, 1 = Yes)

Table 2: Column Definitions

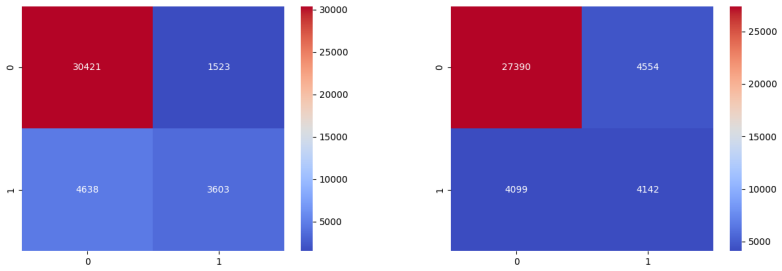
A.1 Key Performance Indicators

1. Accuracy
   Accuracy is the numeric value indicating the performance of the predictive model. It is calculated as follows:
   $$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$
   where:
   TP: True Positive. Result in which the model correctly predicts the positive class.
   FP: False positive. Result in which the model incorrectly predicts the positive class.
   TN: True Negative. Result in which the model correctly predicts the negative class
   FN: False Negative. Result in which the model incorrectly predicts the negative class.
2. Precision
   Precision is defined as the proportion of examples classified as positive that are actually positive. It is calculated as follows:
   $$\text{Precision} = \frac{TP}{TP+FN}$$
3. Recall
   Recall is defined as the number of correctly classified positives over the total number of positives. It is calculated as follows:

$$\text{Recall} = \frac{TP}{TP+FP}$$

4. F1-Score
   F1-Score is the harmonic mean of precision and recall. It is calculated as follows:

$$\text{F1-Score} = \frac{2*precision*recall}{precision+recall}$$



(a) Confusion Matrix for Logistic Regression (b) Confusion Matrix for Decision Tree

(c) Confusion Matrix for Random Forests (d) Confusion Matrix for ANN

Figure 4: Confusion Matrices

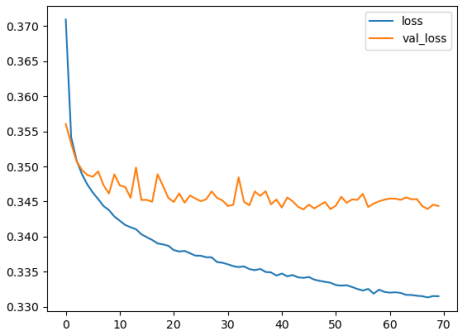


Figure 5: Loss curve against number of epochs for training