Investigating the Comparative Effectiveness of Popular Machine Learning Models at Binary Classification

Research Question: To what extent are artificial neural networks more effective than a random forest model at classifying days with rainfall in Sydney, Australia?

Subject: Computer Science

Word Count: 3980

Table of Contents

1. Introduction

Rainfall prediction is a task that bears immense global significance, and its importance cannot be overemphasized (Goswami & Srividya, 1996). Every year floods and droughts are encountered globally, causing severe damage specifically to the world's crucial food-providing industries. Even with advancements in modern technology, annual agriculture production is still at the mercy of weather and climate (Ahmed et al., 2019). The complexity and dynamic nature of the atmosphere has inhibited typical statistical techniques from providing accurate forecasts for many years. Machine learning models that can effectively predict rainfall based on related atmospheric measurements have had huge implications for worldwide agriculture, minimizing the damage of natural disasters, and climate activism. After living in The United Kingdom which receives around 1,220mm of rain annually (The World Bank, 2023), as well as Saudi Arabia which can receive as little as 59mm of annual rainfall (The World Bank, 2023), I grew curious about the effectiveness of machine learning methods for the binary-classification of rainfall in a unique climate.

The guiding question of my research was: **to what extent are artificial neural networks more effective than a random forest model at classifying days with rainfall in Sydney, Australia?** Sydney was selected due to its moderate annual rainfall, typically between 800 mm (31.50 in) to 1,100 mm (43.31 in), which would provide a relatively balanced dataset for days with and without rain. Sydney, and Australia in general, is also known for its distinct rainfall season. However, the time of year that this period takes place tends to vary throughout the year making rainfall relatively unpredictable, particularly in the El Niño period. Also contributing to the choice of Sydney as the location being studied, is the high-quality weather data provided by Sydney Airport with relatively low amounts of null values. Additionally, Australia suffers one of

the highest risks of natural disasters such as storms, floods, and droughts according to the United Nations World Risk Index (World Risk Report 2023, 2023). Sydney's unique climate and abundant data make it an efficient choice for comparing the predictive abilities of machine learning models.

The approach taken focused on the development of binary classification models and comparing the results of various metrics to judge their individual benefits and drawbacks. Various features including pressure, humidity, and temperature from previous years will be used to train the models and complete this classification which focuses on whether rainfall above a certain threshold (1mm) was seen for a given day.

2. Background Information

2.1. Artificial Neural Networks

Artificial Neural Networks are machine learning models based on the real, biological neural networks in the human brain. They use connected nodes (or neurons or perceptrons) to transfer data from an input to an output layer in a process called forward propagation. The connections between artificial neurons are called edges which have associated weights that adapt throughout the learning process.

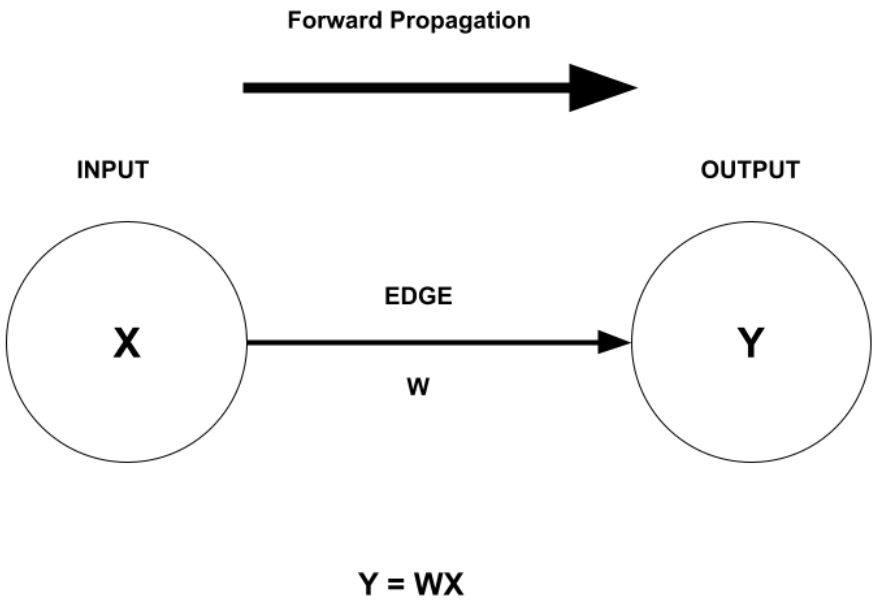

Figure 1: Simple Artificial Neural Network

The ANN seen in Figure 1 above, containing one input and one output neuron, is equivalent to a linear regression model.

Initially, the weights of the edges are random and require tuning by comparing the predicted output with the actual output given by a human for the same input. Each output within the structure is calculated by summing the activation of the input nodes multiplied by their respective weights. For the input layer, the weights represent the importance of each input variable's value with a higher weight bearing more significance on outputs. The summation of

the input and weight products is equal to the dot product of the row vectors of the input and weight variables which are $x = [x_1, x_2, \ldots, x_n]$ and $w = [w_1, w_2, \ldots, w_n]$ respectively.

$$x \cdot w = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

Additionally, a bias, also known as weight 0, is added to this sum to shift the activation function along the x-axis before the output is calculated. Letting z be the sum of the dot product of row vectors x and w and bias b, we get the following equation.

$$z = x \cdot w + b$$

The resulting z-value is input to an activation function (g) which is a function that calculates output, and in the case of classification defines a threshold. This gives us our output which is known as the predicted value denoted by y.

$$y = g(z)$$

This process is visualized in Figure 2 below.

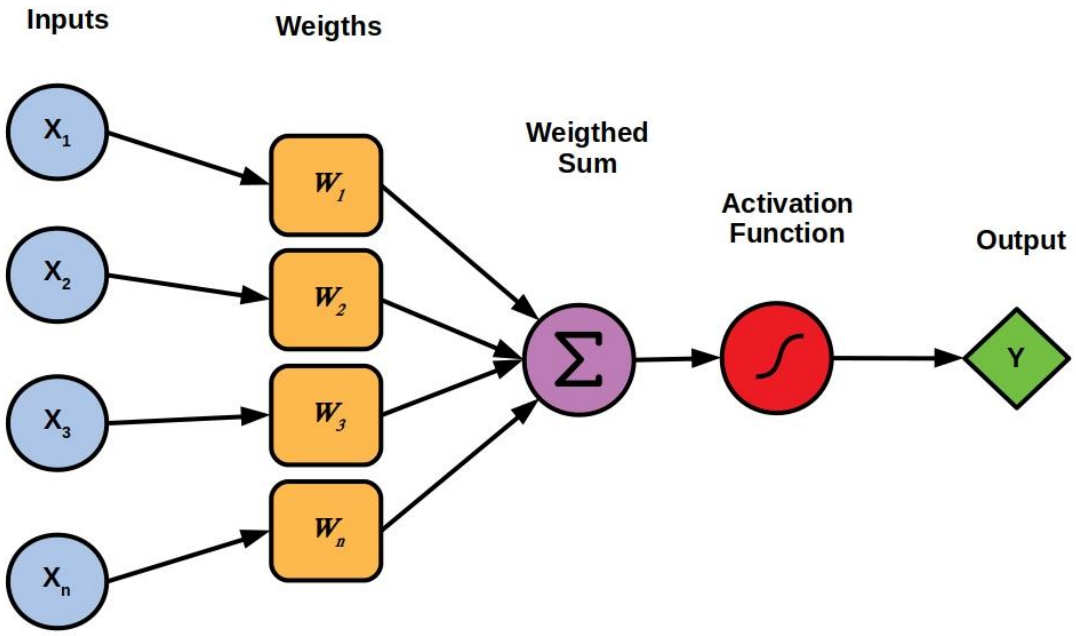

Figure 2: Forward Propagation Visualization (Bendaoud et al., 2022)

All neurons in a layer are connected to all neurons in the previous and following layers. The output given in the above equation is an input for each neuron in the next layer. The layers between the input and output neurons are called the hidden layers and an ANN with multiple hidden layers is called a deep neural network.

The parameters of an artificial neural network include the weights and biases which will be adjusted throughout the learning process. This process consists of backpropagation and optimization where a cost or loss function is used to calculate the error of a given prediction. This error is backpropagated to nodes in previous layers to estimate the error of each node based on their respective weights. With this information, we can update the weights connecting the

nodes to reduce the output error and better match the predicted output with the actual output. The actual output is provided by a human as a part of the supervised learning process. Through this process, artificial neural networks can model complex datasets and make predictions.

2.2. Decision Trees

Decision trees are non-parametric machine-learning models that can be used for classification and regression. They are comprised of nodes and branches (or edges). Each node evaluates one data feature, starting at the root node and ending at a leaf node. To determine which feature will be the root node, the Gini impurity of each feature can be calculated based on the ability of that sole feature to predict the desired output. Letting n be the number of classes (output categories) and p_i be the probability that class i is the output, the formula for the gini impurity of a leaf is as follows:

$$GI = 1 - \sum_{i=1}^{n} (p_i)^2$$

The use of Gini Impurity to determine the root node is visualized in Figure 3 below.

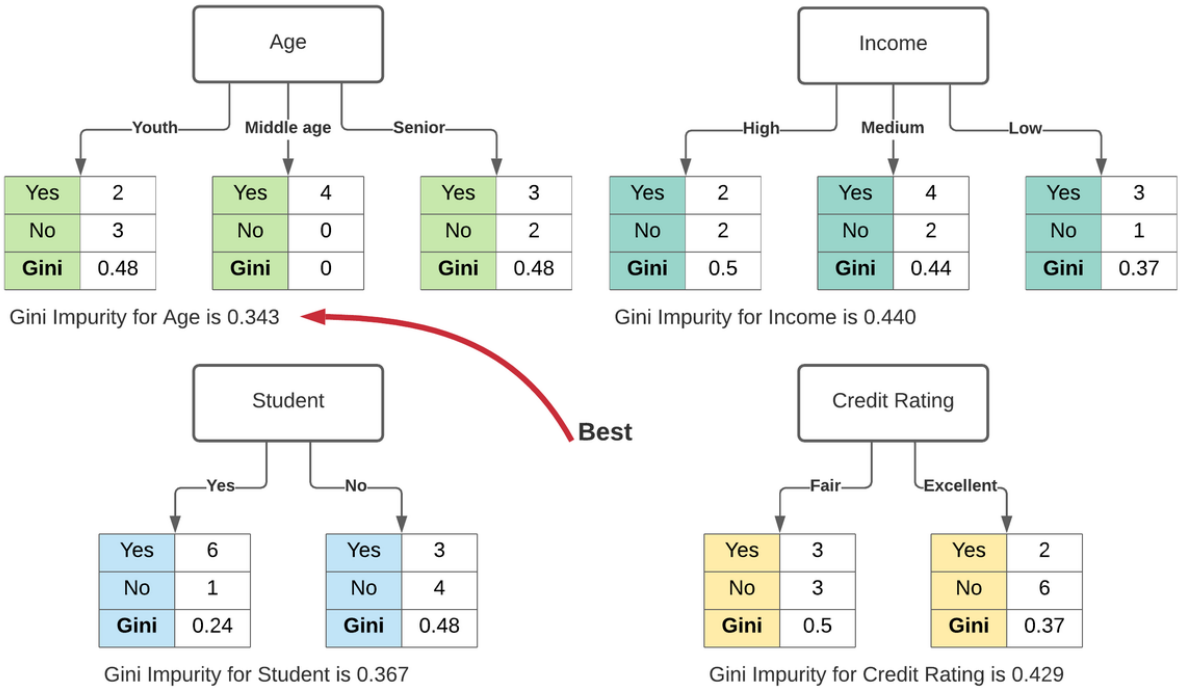

Figure 3: Example of using Gini Impurity to select a root node (Karabiber, n.d.)

The feature with the lowest Gini impurity is put as the root node. A similar process is repeated to determine the sequence that features follow the root node until a leaf is reached. The output value of a leaf node is whichever class has the majority.

2.3. Random Forest

Random forest is a supervised learning algorithm that uses decision trees for classification. It employs the ensemble learning technique where several decision trees are created, and their predictions are averaged. Firstly, a bootstrapped dataset is created by selecting and combining random samples/days from the original dataset. The selection of these samples is completely random, and the same sample can be selected more than once. A decision tree is created from this bootstrapped dataset by randomly selecting a set number of its features as

potential root nodes. The root node will be the feature that best separates the samples, and the unselected features will be removed from the bootstrapped dataset. Nodes following the root node are selected in a similar process of random selection and evaluation of features until all features are used.

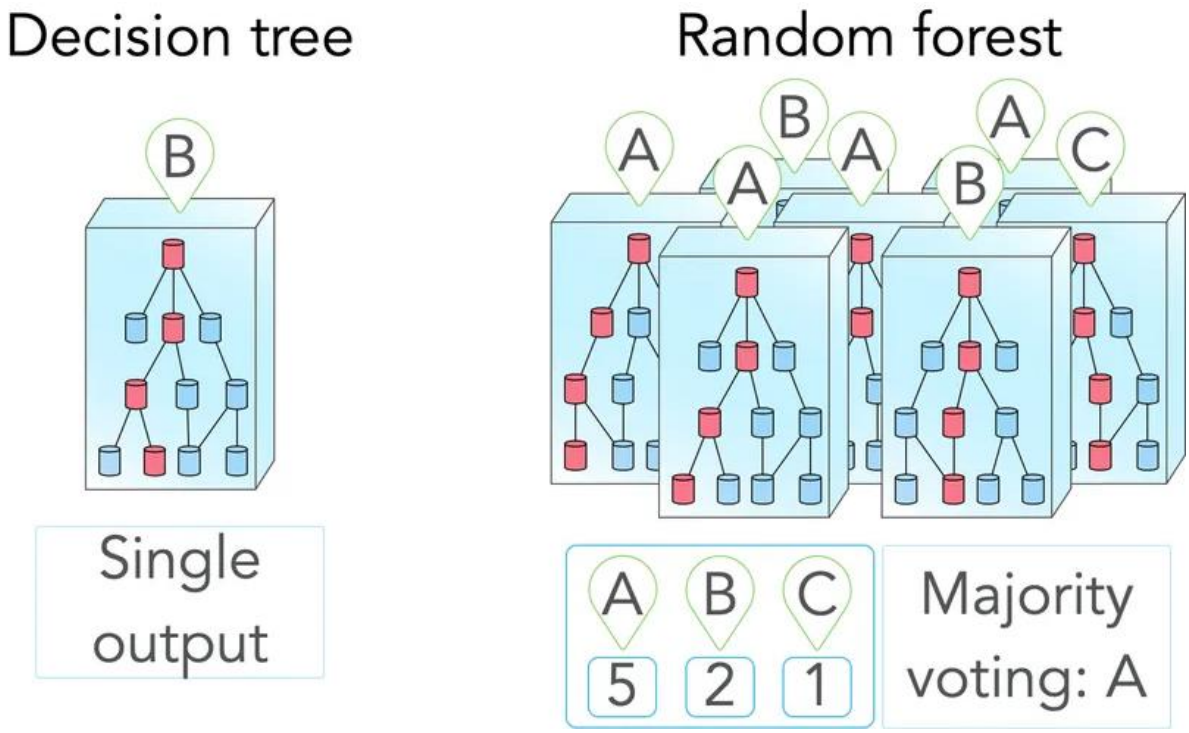

Figure 4: Random Forest Visualization (Rustemov, 2022)

As seen in Figure 4, this process is repeated several times for a wide variety of randomized decision trees. Predictions are made with this model by running the data through each tree and tallying the output. The class with the most predictions will be the final output of the model.

3. Experiment Methodology

3.1. Experimental Procedure

1. The data was pre-processed to fill in null values, remove unrelated features, and add useful features.
2. The data was split into a training and testing dataset at a certain date. This split had to be done so that only previous weather data would be used to predict future rainfall.
3. Several architectures were tested for each model and the one with the best MCC was chosen.
4. The final models were fit to the training data.
5. The models were used to make predictions on the testing data, and the results were recorded with the confusion matrix and several metrics.
6. The results were analyzed and compared.

3.2. The Dataset Used

The dataset used contains around 3009 days worth of Australian weather data with 19 recorded features. It was provided by Sydney Airport and extracted from the Australian Government's Bureau of Meteorology website. It features 10 years' worth of historical climate data from December 1st of 2008 to June 25th of 2017.

3.3. Dependent Variables (features)

Overall, the input variables for each model consisted of 31 features consisting of the selected features after pre-processing and rolling averages.

The included features for each day within the dataset were the minimum temperature (MinTemp, °C), the maximum temperature (MaxTemp, °C), the amount of precipitation in mm (Rainfall, mm), the amount of evaporation (Evaporation, mm), the number of hours of sunshine (Sunshine), the direction and speed of the strongest wind gust (WindGustDir, WindGustSpeed, km/h), the wind direction at 9 am and 3 pm (WindDir9am, WindDir3pm), the wind speed at 9 am and 3 pm (WindSpeed9am, WindSpeed3pm), the humidity at 9 am and 3 pm (Humidity9am, Humidity3pm, %), the atmospheric pressure at 9 am and 3 pm (Pressure9am, Pressure3pm, hPa), the cloud coverage at 9 am and 3 pm (Cloud9am, Cloud3pm, oktas), and the temperature at 9 am and 3 pm (Temp9am, Temp3pm, °C).

The target feature of whether or not there was rain for a given day was added to the data based on whether or not the recorded rainfall feature was over 1mm.

Other features were added to improve the accuracy of the models comprised of rolling means for all features used as input variables. A rolling mean is the average of a span of previous days that can be used to better forecast the current target feature. Through testing, it was established that the models worked best when the rolling means only observed the feature values of the previous day, so the "averages" consisted of one value being the last day's value. Understand that rainfall wasn't an input feature to predict a day's rain but the rolling mean of rainfall from the previous day was. This method improves the models' effectiveness as rain from yesterday can be valuable in predicting rain today.

3.4. Pre-processing

A pre-processing procedure was required to transform the raw data into a more applicable form for training and testing. In this procedure the percentage of null values for each feature was calculated, showing no feature had null values comprising more than 10% of the data. To further

limit the data and select the most relevant features, a correlation matrix was employed. This matrix, constructed with the seaborn library, is seen below.

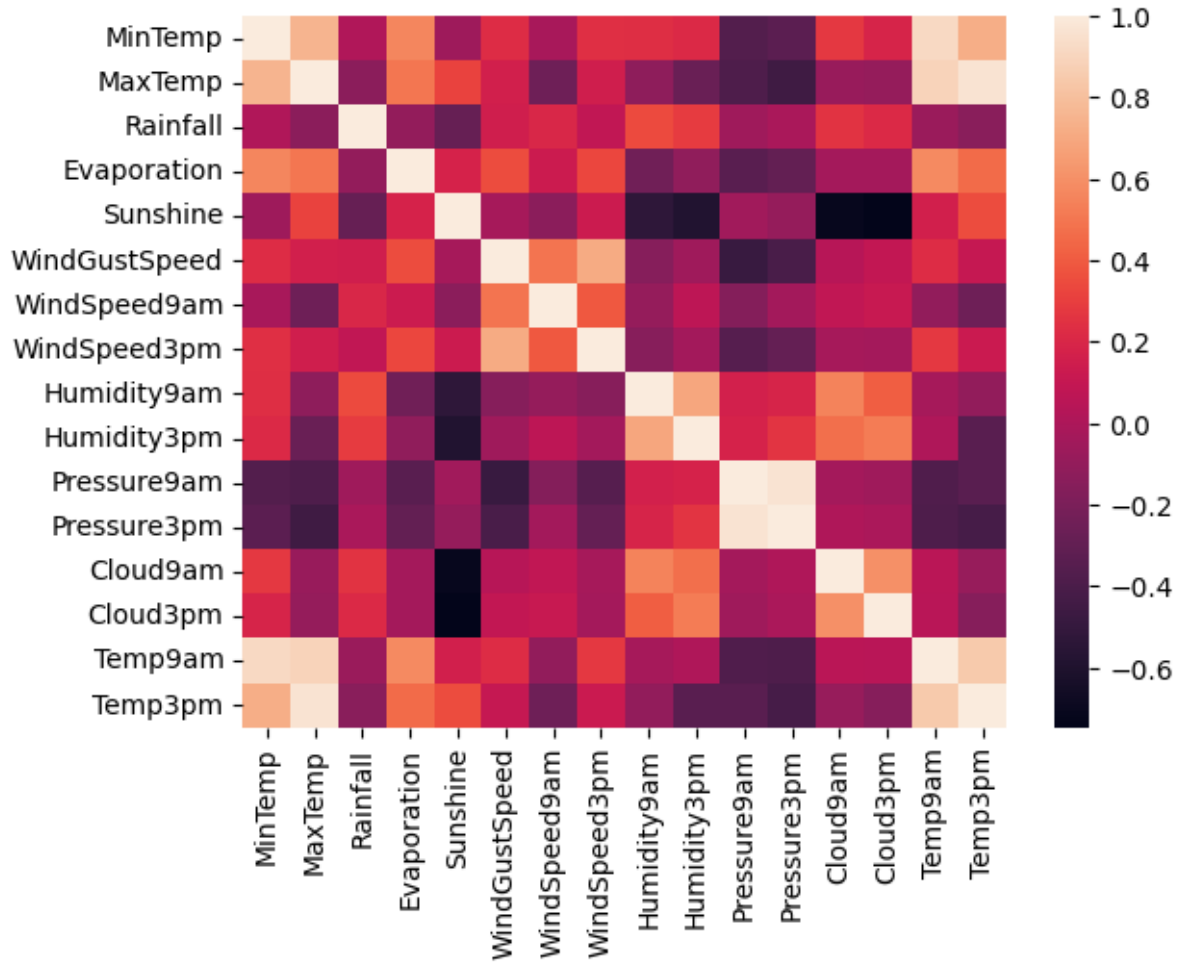

Figure 5: correlation matrix of feature variables

From the data used in Figure 5, we see that MaxTemp and Temp3pm have a correlation coefficient greater than 0.95. Since MaxTemp has no missing values while Temp3pm does, MaxTemp was the feature selected to represent the data. Similarly, Pressure9am and Pressure3pm had a correlation coefficient greater than 0.95, so Pressure9am could be removed due to having a larger percentage of null values between the two features.

Figure 6 below visualizes the typically normal distribution of the features, with Humidity3pm and Pressure3pm as key examples.

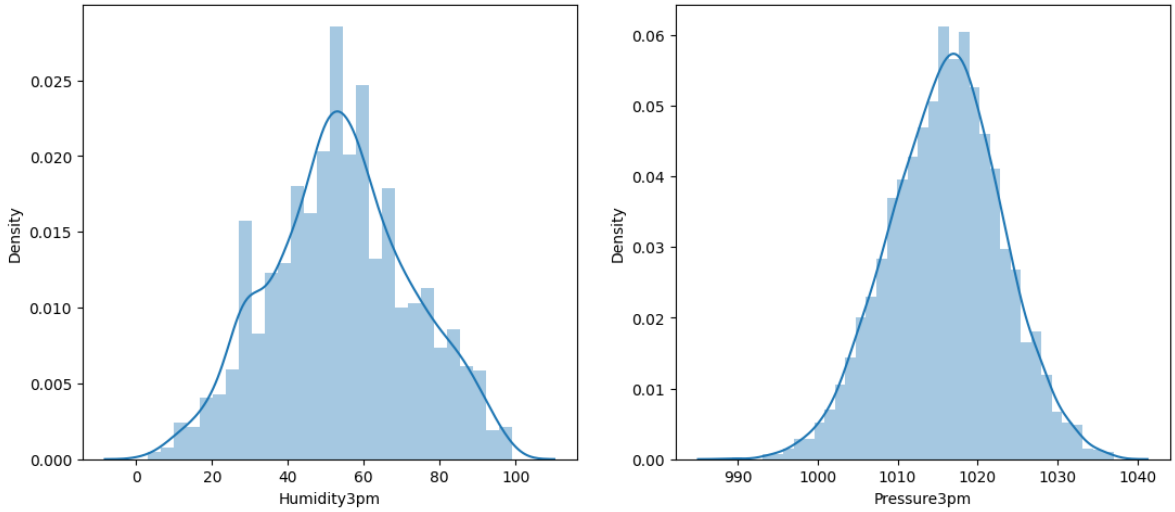

Figure 6: Normal distribution of Humidity3pm and Pressire2pm

Null values were imputed with the mean of the data due to the generally normal distribution of features. Categorical features such as wind direction had their data converted into numeric data, where a value like NS (north-south) would be assigned an integer number. Then missing values are filled in for this categorical data using a K-Nearest Neighbors (KNN) Imputer.

Null values for the target feature, RainToday, however, had to be filled in under the assumption that a lack of data suggests 0 rainfall that day. The days devoid of rainfall values couldn't be deleted because continuous day-by-day data for each feature was necessary as input for time-series forecasting.

3.5. Model Architectures

For this binary classification experiment, a feedforward neural network was employed. Unlike the random forest algorithm, which has only 1 or 2 hyperparameters to tune, deep neural networks have several possible numbers of hidden layers and nodes within those layers that require tuning for a fair comparison between the models.

After a lengthy trial and error process, gathering Matthew's correlation coefficient values for several combinations of hidden layers and numbers of nodes, the selected architecture consisted of 3 hidden layers each with 15 nodes. This structure is illustrated below.

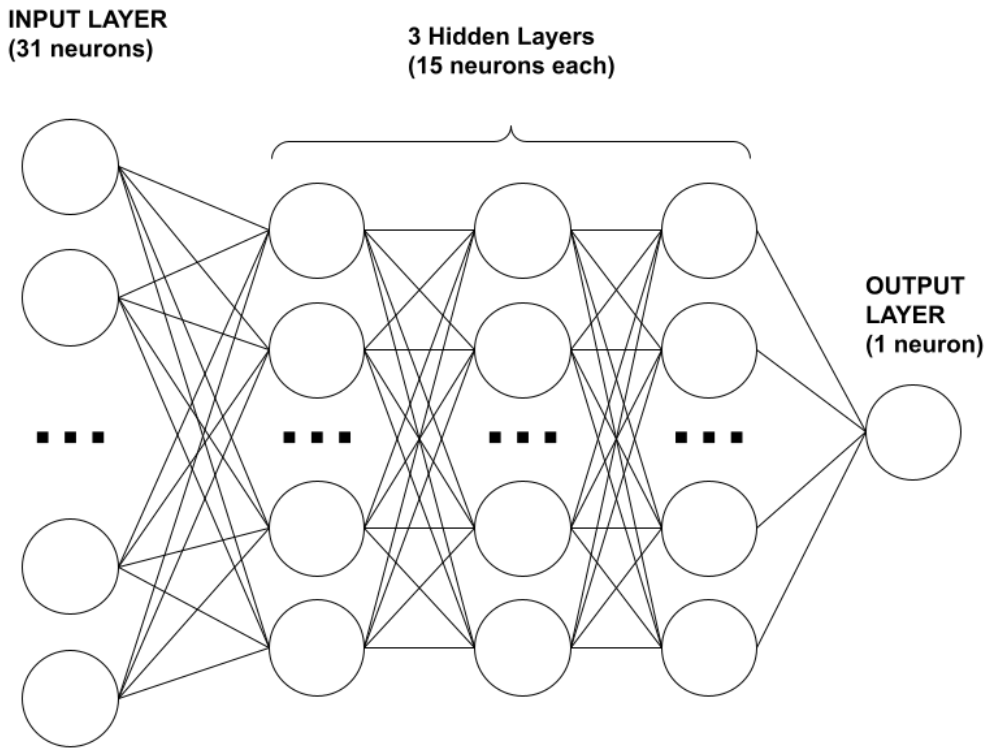

Figure 7: Artificial Neural Network Architecture

This architecture presented the highest MCC value with a relatively low resource usage. Each hidden layer uses the ReLu (rectified linear unit) activation function, while the output layer uses Sigmoid.

For the random forest, the two hyperparameters that had to be tuned were the number of decision trees and the minimum number of samples required to split an internal node. These were determined, through trial and error, to be 73 and 10 respectively.

4. Results

The results of the predictions for each model on the testing set were recorded in the confusion matrices seen below, which display the amount of correct and incorrect guesses for each class.

4.1. Random Forest

4.1.1. Confusion Matrix

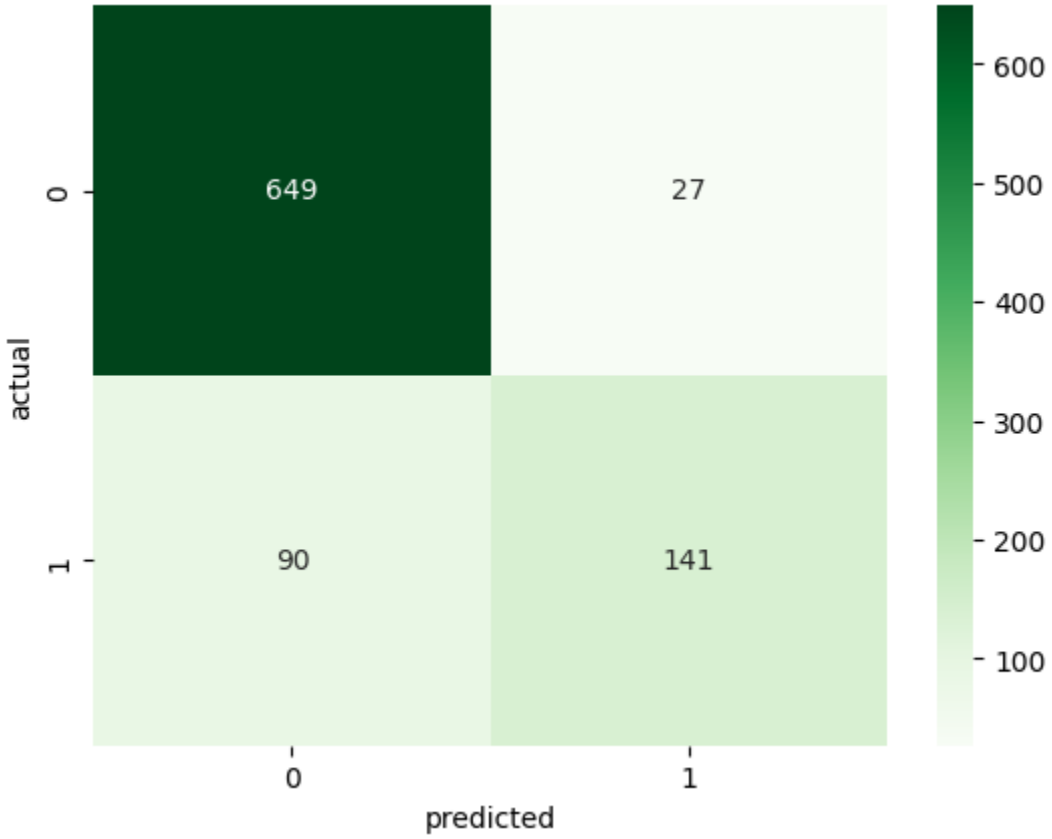

Figure 8: Confusion matrix for the random forest model

4.1.2. Metrics

From the confusion matrix seen in Figure 8 above, we can calculate the accuracy, precision, recall, and Matthew's correlation coefficient of the predictions made by the random forest.

Class	Accuracy	Precision	Recall	MCC

Rain	**0.87100**	**0.83929**	**0.61039**	**0.63976**

Table 1: Evaluation metrics for Figure 8

4.2 Neural Network

4.2.1. Confusion Matrix

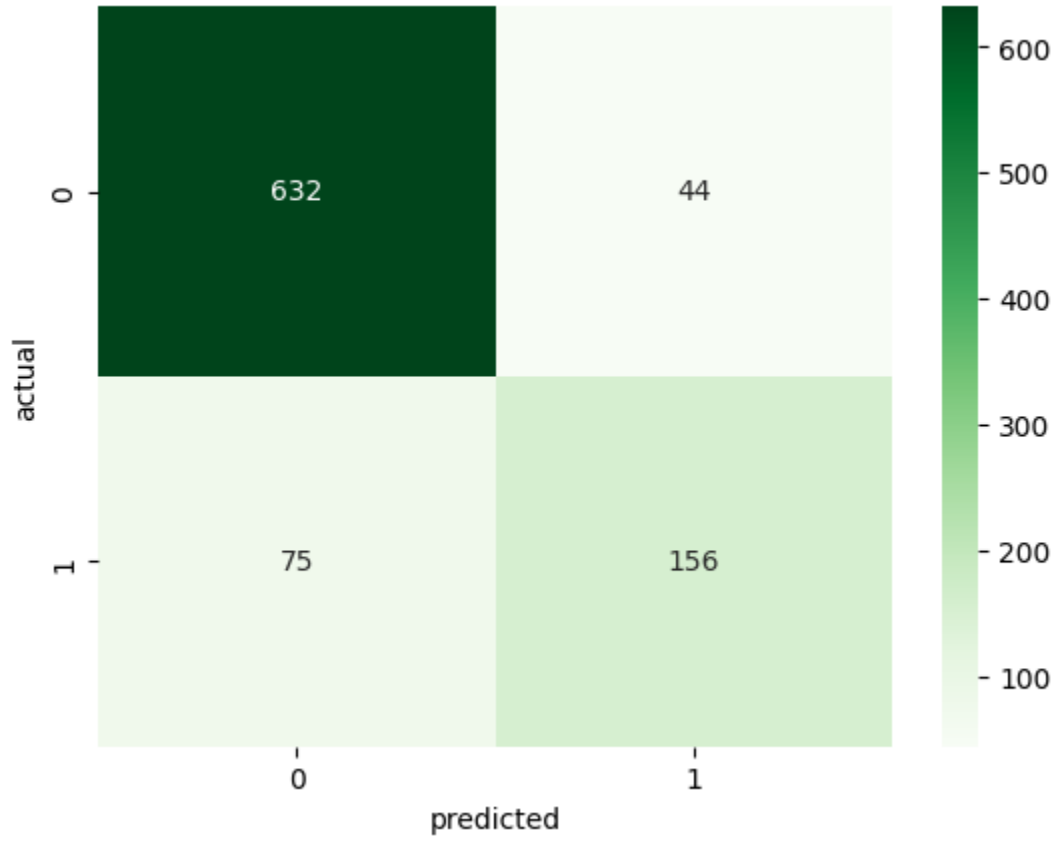

Figure 9: Confusion matrix for the ANN model

4.2.2. Metrics

From the confusion matrix seen in Figure 4 above, we can calculate the accuracy, precision, recall, and Matthew's correlation coefficient of the predictions made by the neural network**.**

Class	Accuracy	Precision	Recall	MCC
Rain	**0.86880**	**0.78000**	**0.67532**	**0.64129**

Table 2: Evaluation metrics for Figure 9

4.3. Result Explanation

A predicted value of 1 means that the prediction was that the given day had rainfall. A predicted value of 0 means the opposite: no rainfall. The same is true for the actual values where 1 indicates a day that historically did have rainfall while 0 indicates a day that did not.

A true positive prediction is a prediction that there was rainfall on a day that did have rainfall. Positive refers to any prediction of rainfall and can be associated with the 1 column on the predicted axis. A false positive prediction is a positive prediction (of 1) that is incorrect, implying that the actual value was 0 (no rain).

Negative predictions correspond to forecasting no rainfall for a given day and are linked to the 0 column on the predicted axis. A true negative prediction is a correct prediction of no rain, implying that the actual value was 0. A false negative prediction is a prediction of no rain for a day that had rain or a day with 1 as the actual value.

The total count of the predictions can be seen as the intersection of the rows and columns in the confusion matrices above.

Accuracy represents the number of correct predictions as a fraction of all predictions. In terms of true and false negatives and positives the following equation can represent it:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision is the fraction of predictions for a given class (rain or no rain) that were correct. Precision is calculated through the following formula where the number of true positive predictions is divided by the sum of true positive and false positive predictions:

$$Precision = \frac{TP}{TP + FP}$$

Precision can be viewed as how likely any prediction is to be correct.

Recall is the fraction of all days for a given class (rain or no rain) that were predicted correctly. Recall is calculated through the following formula, where the number of true positives is divided by the sum of true positive and false negative predictions:

$$Recall = \frac{TP}{TP + FN}$$

Remember that false negatives are predictions for rainy days that were incorrectly determined to have none. Recall can be viewed as the percentage of rainy days the model can identify across the 10 years.

Matthew's correlation coefficient (MCC), also known as the phi coefficient, was introduced into machine learning by biochemist Brian W. Matthews in 1975 (Matthews, 1975). It incorporates all true, false, positive, and negative counts into a number that gives a general representation of the effectiveness of a binary classification model. 1 represents entirely perfect

predictions, and 0 represents total disagreement between predictions and reality. This formula represents the relationship between the confusion matrix and the coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

5. Analysis

5.1. Analysis of results

From the results in Table 1 and Table 2 we can see that overall, the random forest had a higher accuracy than the artificial neural network. While this does mean that the random forest had a higher fraction of correct predictions overall, we know that this metric is unreliable due to its bias within imbalanced datasets (Boughorbel et al., 2017). Since there are more days without rainfall than with rainfall in this arid climate, a model that guesses every day having no rainfall would be decently accurate but a poor predictor of rainfall. This is where precision comes into play, indicating the percentage of days identified as having rainfall that experienced rainfall. In this metric, the random forest again proves superior with a precision of 0.83929 compared to the ANN's precision of only 0.78000. This is contrasted, however, by the ANN's superior recall of 0.67532 as opposed to the random forest's recall of only 0.61039. From this, one could conclude that the predictions of rain made by the random forest model are more likely to be correct, while the artificial neural network is better at identifying a larger portion of days with rainfall. Both models have higher precisions than recalls, showing that the percentage of their predictions of

rainy days that were correct was higher than the percentage of all rainy days that were predicted to have rain.

Despite providing more insight than accuracy, precision and recall both have their flaws. For example, if a model were to make one prediction for a day with rain and that prediction happened to be correct then the model would have a precision of 1 or 100%. However, this model is still a poor predictor of rainfall as it could only predict one of the several days with rain over the given time frame. Conversely, if a model predicted all days to have rainfall, then every day with rain will have been correctly identified and therefore the model has a recall value of 1 or 100%. Similarly, to the model with a precision of 100%, this model is still a poor predictor of rainfall as each prediction for a day with rainfall is highly unreliable.

The accuracy metric suffers from imbalanced data and is unable to effectively criticize classification models that are biased towards the majority class. The precision and recall metrics only analyze 2 to 3 of the 4 possible results (TP, FP, TN, FN) and therefore cannot provide an overview of the model's effectiveness. This is why metrics such as the F1 score and Matthew's correlation coefficient were invented. Research has shown that the MCC "gives a better summary of the performance of classification algorithms," than the F1 score because the F1 score doesn't incorporate the count of true negatives while the MCC incorporates all 4 scenarios (Boughorbel et al., 2017)

Acknowledging this, we recognize that the artificial neural network has a slightly higher MCC than the random forest. This difference, however, is only around 0.00153 and within the range of random deviation due to a neural network's randomly initialized weights. Hence, we cannot use this difference to conclude one model's predictive ability over the other.

5.2. Analysis of Limitations

It is crucial to understand the nature of this investigation and some key limitations of my methodology. Weather prediction is an extensively researched and extremely difficult task, typically built upon mathematical models far more complex than I intended to explore here. My work investigates the fundamental advantages and disadvantages of well-known machine learning models that provide an insightful foundation into the nature of this problem. For practical application in meteorology, the methodology used here would be accompanied by several other models that aren't the focus of this exploration. For example, more extensive time series forecasting methods using recurrent neural networks can be used to classify days with rainfall further into the future. Additionally, these classification models can be used to impute missing historical data on rainfall given that other factors such as temperature were recorded. This can improve the data processing used for more advanced models.

Also present, was the option for a regression approach that aims to predict not just if a day will receive rainfall, but how much rainfall will be received. This can provide more useful details for agriculture and disaster damage mitigation. This method, accompanied by hourly measurements for the feature variables, could have been explored if not for the limited resources available to me. Training neural networks can require extensive time and computational power, which I lacked.

It is also important not to assume that the models were fairly compared on all accounts. The hyperparameters for both models had to be tuned through exhaustive trial and error to compare the best results. However, thousands of viable architectures, particularly for neural networks, could not be explored and may provide more state-of-the-art results.

The data was carefully selected from a reliable source to minimize the presence of null values. The issue of imbalanced data was slight and there was not a striking lack of data. Nevertheless, interpolation was employed throughout the pre-processing step under various assumptions that may provide inaccurate data. This issue was nearly unavoidable considering that removing days that lacked data would impact the effectiveness of the time-series forecasting. There was the potential for more extensive interpolation methods than using local averages or assuming 0, but we must also acknowledge that the ability to fill in missing values is one of the original goals of this classification task.

In summary, while the methods employed may not have been avant-garde, valuable insight is found in evaluating these different, viable models. For future extensions of this research, I intend to expand the dataset and spend more resources on training the models. This can be accomplished by gathering data as well as data augmentation, which is the process of artificially expanding a dataset similar to how I added the rolling variable features.

6. Conclusion

In conclusion, this investigation utilized 10 years of climate data from Sydney, Australia to train a random forest and an artificial neural network model. The data was initially pre-processed to fill in null values, remove redundant features, and create additional useful features. The new data was input into both models and their resulting confusion matrices were compared through various metrics.

The results showed that neither model proved to be vastly inferior in predictive performance, as indicated by the similar MCC values. However, there are noticeable advantages and disadvantages to each model as showcased by the precision and recall values. The decision

between which model is more effective depends on the priorities and values of someone desiring to predict rainfall. If an organization desires a few highly accurate predictions of days that will have rainfall, perhaps for establishing events that require rain, then the random forest model may be favored due to its higher precision for the rain class. Conversely, if an organization desires a broad overview of roughly how many days it will rain in an upcoming year, the artificial neural network may be favored due to its higher recall for the rain class.

These models might seem competitive for long-term agricultural purposes, but we must also consider the greater number of resources required for the neural network's predictions. In the case of climate-based natural disasters such as heavy storms and flash floods, accurate predictions may need to be determined within a few days or hours. In these time-restricted scenarios, the random forest model may be the more effective option due to the extensive learning time of a neural network.

Despite the significant popularity of deep neural networks in recent years, the simpler random forest model still proves comparable in this case of binary classification. It is important to recognize the potential of overengineering a problem, where the more complex model might not be necessary.

Word Count: 3980

7. References

Ahmed, A., Deb, D., & Mondal, S. (2019, September 10). Assessment of rainfall variability and its impact on groundnut yield in Bundelkhand region of India. *Current Science, 117*(5), 794-803, Retrieved August 28, 2023, from https://www.jstor.org/stable/27138343

Bendaoud, M., Wolfgang, B., & El Fathi, A. (Eds.). (2022). *The Proceedings of the International Conference on Electrical Systems & Automation: Recent Advances in Renewable Energy—Volume 1*. Springer Nature Singapore. Retrieved August 10, 2023, from https://doi.org/10.1007/978-981-19-0035-8_5

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017, June 2). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE*, *12*(6). e0177678, Retrieved August 8, 2023, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0177678

Goswami, P., & Srividya. (1996, March 25). A novel neural network design for long range prediction of rainfall pattern. *Current Science,*, *70*(6), 447-457, Retrieved August 25, 2023, from http://www.jstor.org/stable/24097412

Karabiber, F. (n.d.). *Gini Impurity*. LearnDataSci. Retrieved August 12, 2023, from https://www.learndatasci.com/glossary/gini-impurity/

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451, Retrieved August 13, 2023, from https://doi.org/10.1016/0005-2795(75)90109-9

Rustemov, I. (2022, December 11). *Random Forest from scratch*. Medium. Retrieved August 12, 2023, from https://medium.com/@ibrahimrustemov121/random-forest-from-scratch-3aa1396bb44

The World Bank. (2023, June 16). *Average precipitation in depth (mm per year)*. The World Bank. Retrieved August 25, 2023, from https://data.worldbank.org/indicator/AG.LND.PRCP.MM?most_recent_value_desc=tr

World Risk Report 2023. (2023). Bündnis Entwicklung Hilft, Ruhr University Bochum – Institute for International Law of Peace and Conflict 2023. Retrieved August 25, 2023, from https://weltrisikobericht.de/en.