**Rainfall prediction using orange**

**Julius Chan**

Table of Contents

[**1.0 Rainfall Prediction using Orange (Part B)** 4](#_Toc167915779)

[**1.1 Introduction to Orange** 4](#_Toc167915780)

[**1.1.1 Technical Features** . 4](#_Toc167915782)

[**1.1.2 Usage for Machine Learning** 4](#_Toc167915783)

[**1.2: Rainfall Prediction using Neural Network** 5](#_Toc167915784)

[**1.2.1 Objective** . 5](#_Toc167915785)

[**1.2.2 Expected Outcomes** 5](#_Toc167915786)

[**1.2.3 Dataset: weatherAUS** 5](#_Toc167915787)

[**1.2.4 Data-Preprocessing** 6](#_Toc167915788)

[**1.2.5 Machine Learning Algorithm Overview** 8](#_Toc167915789)

[**1.2.6 Selected Machine Learning Algorithms** 9](#_Toc167915790)

[**1.2.6 Performance and Results** 10](#_Toc167915791)

[**1.2.7 Conclusion** 12](#_Toc167915792)

[**2.0 List of References** 13](#_Toc167915793)

**Table of Figures**

[Figure 1: 'WeatherAUS' Data Types & Labels (Julius Chan, 2024) 5](#_Toc167915823)

[Figure 2: Ranking of Features (Julius Chan, 2024) 5](#_Toc167915824)

[Figure 3: Feature Selection for Data Pre-Processing (Julius Chan, 2024) 6](#_Toc167915825)

[Figure 4: Impute Widget (Julius Chan,2024) 6](#_Toc167915826)

[Figure: 5: Data Post-Processed Data 7](#_Toc167915827)

[Figure 6: Construction of Machine Learning Program (Julius Chan, 2024) 7](#_Toc167915828)

[Figure 7: Test and Score of Machine Learning Algorithms (Julius Chan,2024) 9](#_Toc167915829)

[Figure 8: Confusion Matrix of Machine Learning Algorithms (Julius Chan,2024) 10](#_Toc167915830)

[Figure 9: ROC Analysis of Models (Julius Chan, 2024) 10](#_Toc167915831)

# **1.0 Rainfall Prediction using Orange**

## **1.1 Introduction to Orange**

## In 1996, the Bioinformatics Laboratory of the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, developed Orange in cooperation with an open-source community. (Meta Brown, 2016). Orange has been selected as the machine learning tool as it is a strong, user-friendly open-source machine learning and data visualization platform that has grown popular within the data science community appealing to beginners and experts.

### **1.1.1 Technical Features Visual Programming**: No coding is required. Orange's visual programming interface allows users to create data pipelines by dragging and dropping widgets. **Interactive Data Visualization**: The widgets accept data from the input and send out filtered or processed data, models, or whatever the widget performs on the output.

### **1.1.2 Usage for Machine Learning**

**Exploratory Data Analysis:** With Orange's visualization capabilities, users may quickly discover insights and trends in their data.

**Model Training and Results:** Orange facilitates the training and evaluation of predictive models by allowing users to experiment with multiple algorithms and parameter settings to see which models perform best.

## **1.2: Rainfall Prediction using Neural Network**

### **1.2.1 Objective**The objective is to predict rainfall accurately for the following day. This prediction is vital for planning daily activities and mitigating the impact of unexpected weather events. Precise predictive models can be developed with orange to improve forecast accuracy, enabling better decision-making and preparation for weather-related challenges.

### **1.2.2 Expected Outcomes** **Binary Classification:** The forecast is us binary classification issue, with the goal variable indicating whether it will rain tomorrow (RainTomorrow). **Performance indicators**: Key performance indicators for assessing the model may include model accuracy. **Relevant Features**: Beyond prediction, the model may reveal which features are most significant in forecasting rainfall.

### **1.2.3 Dataset: weatherAUS**

**Source**: The dataset is sourced from the Australian Bureau of Meteorology and comprises around ten years of daily weather observations from various places throughout Australia. It is accessible and acquired on [Kaggle](This%20dataset%20is%20sourced%20from%20the%20Australian%20Bureau%20of%20Meteorology%20and%20comprises%20around%20ten%20years%20of%20daily%20weather%20observations%20from%20various%20places%20throughout%20Australia.%20It%20is%20accessible%20and%20acquired%20on%20Kaggle:%20https:/www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package.).

**Data Structures:** The dataset is a two-dimensional table with each row representing a daily weather observation from a specific place and each column representing an individual component of the weather data, such as numerical and categorical values.

**Data Labels & Types**: A breakdown of data labels and types, as shown in Figure 21

A screenshot of a data

Description automatically generated

Figure : 'WeatherAUS' Data Types & Labels (Julius Chan, 2024)

### **1.2.4 Data-Preprocessing**

A screenshot of a computer

Description automatically generated

Figure : Ranking of Features (Julius Chan, 2024)

The Rank Widget shown in Figure 22 helps to identify the best 10 features based on Information Gain and Gain Ratio Score.

Information Gain and Gain Ratio are important metrics. Information Gain values indicate the value of a feature when forecasting (RainTomorrow), whereas Gain Ratio normalizes this statistic by incorporating related information.

A screenshot of a computer

Description automatically generated

Figure : Feature Selection for Data Pre-Processing (Julius Chan, 2024)

Figure 23 shows a pre-processing stage. The purpose is to choose the highly correlated features, shown in Figure 22 as these features improve the model's accuracy and efficiency.

A screenshot of a computer

Description automatically generated

Figure : Impute Widget (Julius Chan,2024)

The Impute widget in Figure 24 is a pre-processing step that manages missing data in a dataset. The "Remove instances with unknown values" option only eliminates data rows that have missing or unknown values.

A screenshot of a computer

Description automatically generated

Figure: : Data Post-Processed Data

After performing the Impute Function, no missing or unknown values are identified, This ensures the dataset used for model training and testing is cleaned, eliminating any concerns with missing data that could hinder model performance.

### **1.2.5 Machine Learning Algorithm Overview**

**A diagram of a diagram

Description automatically generated**

Figure : Construction of Machine Learning Program (Julius Chan, 2024)

Figure 26 shows an overview of the program construction using Orange and illustrates the complete rainfall prediction workflow, including data analysis , feature selection, pre-processing, model training, and assessment. The selected machine learning algorithm consists of Logistic Regression, Random Forest, and Neural Network.

### **1.2.6 Selected Machine Learning Algorithms**

**Random Forest:** A commonly used machine learning algorithm that combines the output of multiple Decision Trees to achieve a single result. (GFG, 2022).   
  
It works well with both numerical and categorical variables since it detects key features for rainfall prediction, making the weatherAUS dataset a good match due to its resistance to overfitting.

Rainfall is predicted differently by each tree using weatherAUS dataset features. The algorithm creates various decision trees during training by combining subsets of data.  
  
The final prediction for input is based on the average or weighted average of all the individual trees' predictions. (AnalytixLabs, 2023)

**Neural Network**: The widget uses sklearn's Multi-layer Perceptron algorithm. A Multi-Layer Perceptron (MLP) is a sort of artificial neural network that has multiple layers of neurons and is frequently used for different machine-learning tasks, including classification and regression. (GFG, 2023).  
  
Neural networks excel at weather prediction because of the ability to capture complicated, non-linear correlations in data and the potential for high predictive performance on large and complex datasets.

During training, the network learns complicated correlations in the weatherAUS dataset by passing input features through its layers and using non-linear activation functions which help to capture the relationship between features and rainfall outcome.

**Logistic Regression:** Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. (Vijay Kanade, 2022)

The algorithm simplifies and interprets the relationship between features and rainfall probability which is well-suited to binary classification like predicting rain tomorrow. It also provides a baseline for comparing complex models such as Random Forest and Neural Networks.

During training, Logistic Regression defines the parameters of a logistic function, which converts input features to a probability score ranging from 0 to 1 for rainfall prediction.

### **1.2.6 Performance and Results**

Figure 26 shows the selected performance matrix. Test & Score displays an overall model evaluation. ROC Analysis assesses the model's ability to distinguish between rainy and non-rainy days at various threshold levels. Confusion Matrix gives detailed error insights for refinement. This approach provides comprehensive evaluation and effective model improvement.

A screenshot of a computer

Description automatically generated

Figure : Test and Score of Machine Learning Algorithms (Julius Chan,2024)

**Accuracy:** Neural Network is the most accurate (0.892), outperforming Random Forest (0.891) and Logistic Regression (0.879).   
  
**Precision and Recall:** Random Forest and Neural Network both have excellent accuracy and recall, which allows them to detect real positives while limiting false positives successfully. Logistic regression produces somewhat more incorrect predictions.

**F1:** The Neural Network has the greatest F1 score (0.850), suggesting the best balance of precision and recall, which is critical for reducing prediction mistakes.   
**MCC:** Neural Network has the greatest MCC (0.555), suggesting the best-balanced performance across all classes and providing a trustworthy measure of model quality.

**A screenshot of a computer

Description automatically generated**

Figure : Confusion Matrix of Machine Learning Algorithms (Julius Chan,2024)

**Neural Network:** The best balance, with the fewest false negatives (11.9%) and the highest true positive rate (88.1%), suggests a good rainy day forecast.   
  
**Random Forest**: Similar to Neural Network, but with a larger false positive rate (25.3%), suggesting a somewhat better likelihood of forecasting rain when it does not fall.

**Logistic regression:** This has a slightly greater false negative rate (12.5%), which might lead to more missed rainy day forecasts.

A graph with a line

Description automatically generated

Figure : ROC Analysis of Models (Julius Chan, 2024)

The ROC curve for the 'Yes' target variable on RainTomorrow was evaluated and showed high ROC values for the selected models. This demonstrates greater performance in distinguishing positives from negatives, establishing Neural Network as the most effective for rainfall prediction.

### **1.2.7 Conclusion**

A rainfall forecast model with good predictability was developed through Orange’s classification algorithms and performance assessment using precise metrics making it applicable to real-world data sets.

A key learning outcome is to focus on finding the best model for predicting rainfall, rather than just tuning models to improve their accuracy for predictive analysis. This approach is highly applicable in the AI world.

# **2.0 List of References**

AnalytixLabs (2023) “Random Forest Regression — How It Helps in Predictive Analytics?” , https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics 01

GeeksforGeeks (2023) “Classification Using Sklearn Multi-Layer Perceptron.” , www.geeksforgeeks.org/classification-using-sklearn-multi-layer-perceptron/

d.o.o, Arctur (n.d.) “ORANGE: Faculty of Computer and Information Science.” *,* www.inzenirji-bomo.si/en/kvizum/2021090213234141/orange-faculty-of-computer-and-information-science/.

GeeksforGeeks (2022) “Differences between Random Forest and AdaBoost.” www.geeksforgeeks.org/differences-between-random-forest-and-adaboost/.

space/#:~:text=The%20videos%20are%20stored%20in.

Zach (2021) “How to Interpret a ROC Curve (with Examples).” www.statology.org/interpret-roc-curve/.