

Rain Prediction Using Machine Learning

Aldimeola Alfarisy



Table of Contents

01



**Business Background
and Objectives**

02



Data Preparation

03



**Exploratory Data
Analysis (EDA)**

04



Feature Engineering

05



**Modelling and
Evaluation**

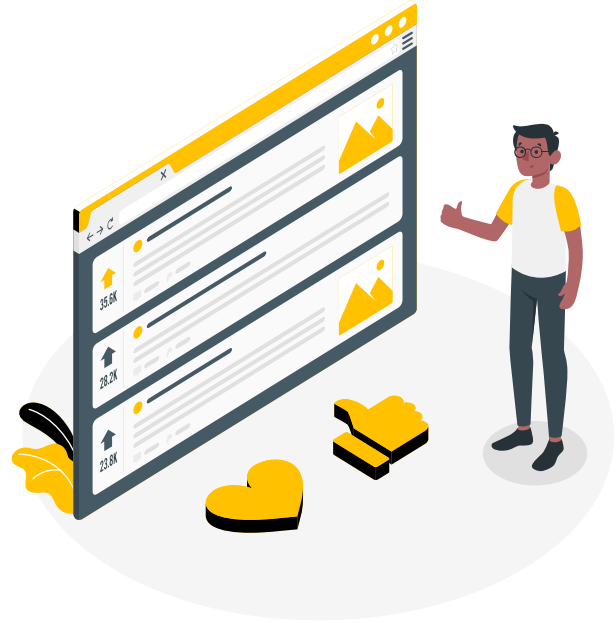
06



**Conclusion and
Recommendation**

Business Background and Objectives

01



Introduction and Problems



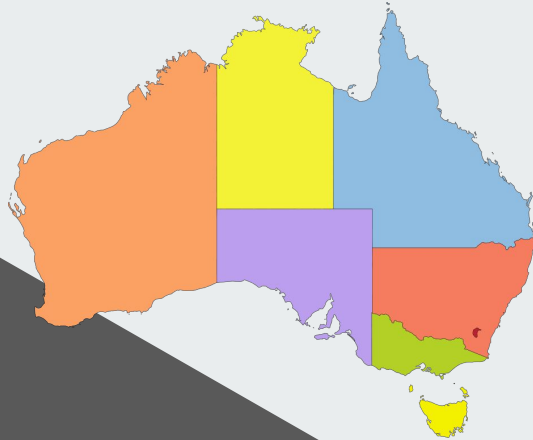
The weather has a significant impact on many life aspects, one of which is agricultural industry and because of that, being able to predict it helps farmers in their day-to-day decisions such as how to plan efficiently, minimize costs and maximize yields.

A major agricultural company needs to have an accurate **rain prediction algorithm** that will improve their decision-making on typical farming activities such as **planting** and **irrigating**.

Using historical rain information from Australia regions in 10 years as research data, it is necessary to **predict** weather(**rain**) in next day.

Objectives

- What factors and conditions in current day that will cause rain in the next day?
- What machine learning algorithms are suitable for predicting rain in the next day?
- Where is the location that has the highest frequency of rain?
- What is the impact of the predictive model for business problems operating in the agricultural sector?

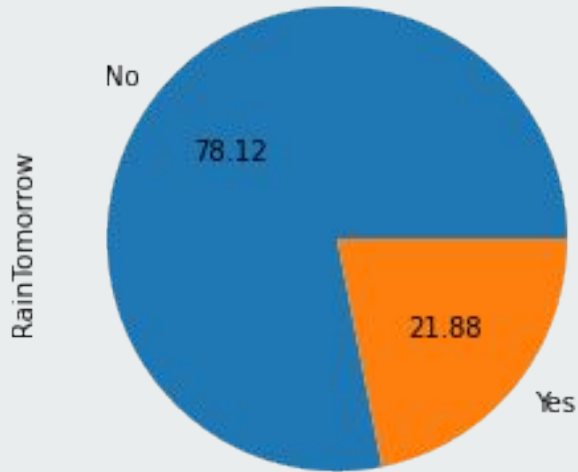


Data Preparation

02



Overview



Target distribution is **imbalance**,
I will use **AUC** as model metric
evaluation

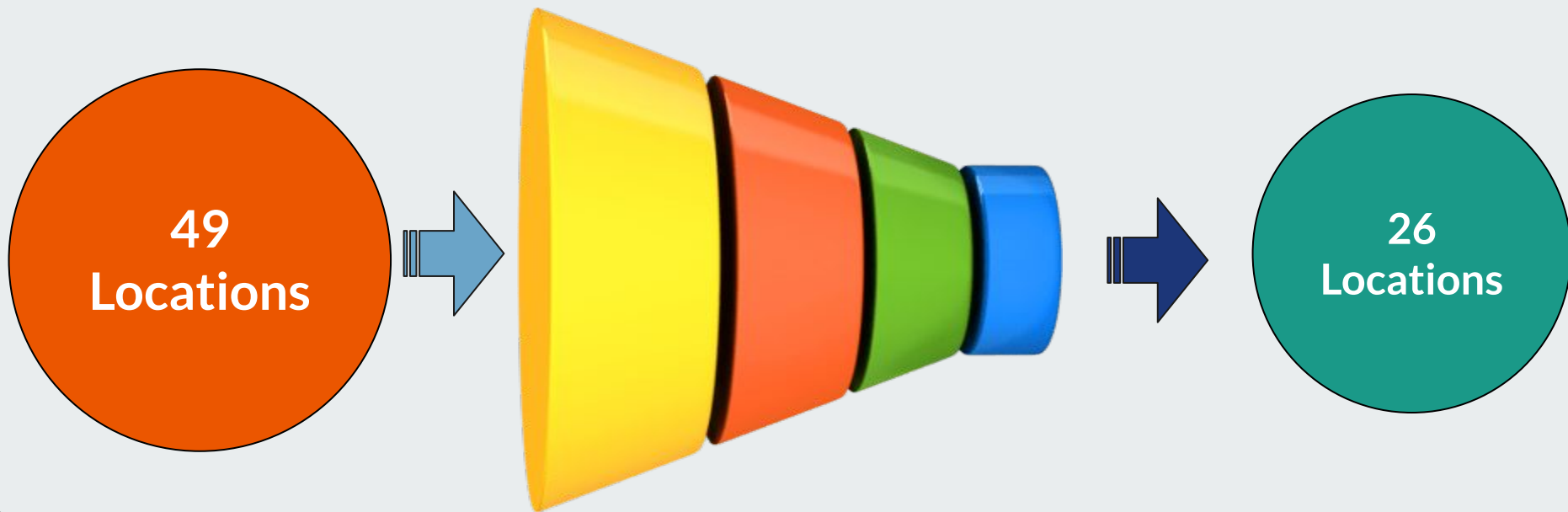
Data contains **145460 rows** and **22 features** and **1 binary column** with **RainTomorrow** as **target**.

All features are the weather and climate elements that occur on that day in certain location that are used to predict rain in the next day.

Features consist **6 categorical** features, and **16 numerical** features (Date will be transform to datetime format).

| Numerical Features | | Categorical Features |
|---|---|---|
| <ul style="list-style-type: none">• MinTemp• MaxTemp• Rainfall• Evaporation• Sunshine• WindGustSpeed• WindSpeed9am• WindSpeed3pm | <ul style="list-style-type: none">• Humidity9am• Humidity3pm• Pressure9am• Pressure3pm• Cloud9am• Cloud3pm• Temp9am• Temp3pm | <ul style="list-style-type: none">• Date• Location• WindGustDir• WindDir9am• WindDir3pm• RainToday |

Location Feature Filtering

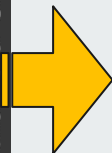


There are some **features** in certain locations that values is **completely missing** in 10 years observation. I decide to drop those locations.

Missing Values Handling

Missing Values Ratio

| | |
|---------------|-----------|
| Location | 0.000000 |
| Month | 0.000000 |
| Year | 0.000000 |
| Day | 0.000000 |
| MinTemp | 0.705897 |
| MaxTemp | 0.688406 |
| Rainfall | 2.153923 |
| Evaporation | 11.068216 |
| Sunshine | 17.833583 |
| WindGustDir | 3.445771 |
| WindGustSpeed | 3.413293 |
| WindDir9am | 3.516992 |
| WindDir3pm | 1.136932 |
| WindSpeed9am | 0.864568 |
| WindSpeed3pm | 0.845827 |
| Humidity9am | 1.101949 |
| Humidity3pm | 1.014493 |
| Pressure9am | 1.444278 |
| Pressure3pm | 1.434283 |
| Cloud9am | 10.805847 |
| Cloud3pm | 12.435032 |
| Temp9am | 0.770865 |
| Temp3pm | 0.715892 |
| RainToday | 2.153923 |
| RainTomorrow | 2.160170 |



Group by

Median value
Numerical Data

Mode value
Categorical Data

| | | MinTemp | MaxTemp | Rainfall | Evaporation | WindGustDir | WindDir9am | WindDir3pm |
|--------------|---|---------|---------|----------|-------------|-------------|------------|------------|
| AliceSprings | 1 | 21.90 | 36.55 | 0.0 | 12.9 | ESE | E | ESE |
| | 2 | 20.55 | 36.10 | 0.0 | 12.4 | SE | E | SE |
| | 3 | 18.40 | 34.10 | 0.0 | 10.4 | ESE | ESE | ESE |
| | 4 | 12.65 | 29.30 | 0.0 | 7.8 | ESE | ESE | SE |
| | 5 | 7.70 | 23.20 | 0.0 | 4.8 | ESE | ESE | SE |
| | 6 | 3.70 | 19.55 | 0.0 | 3.8 | ESE | E | ESE |

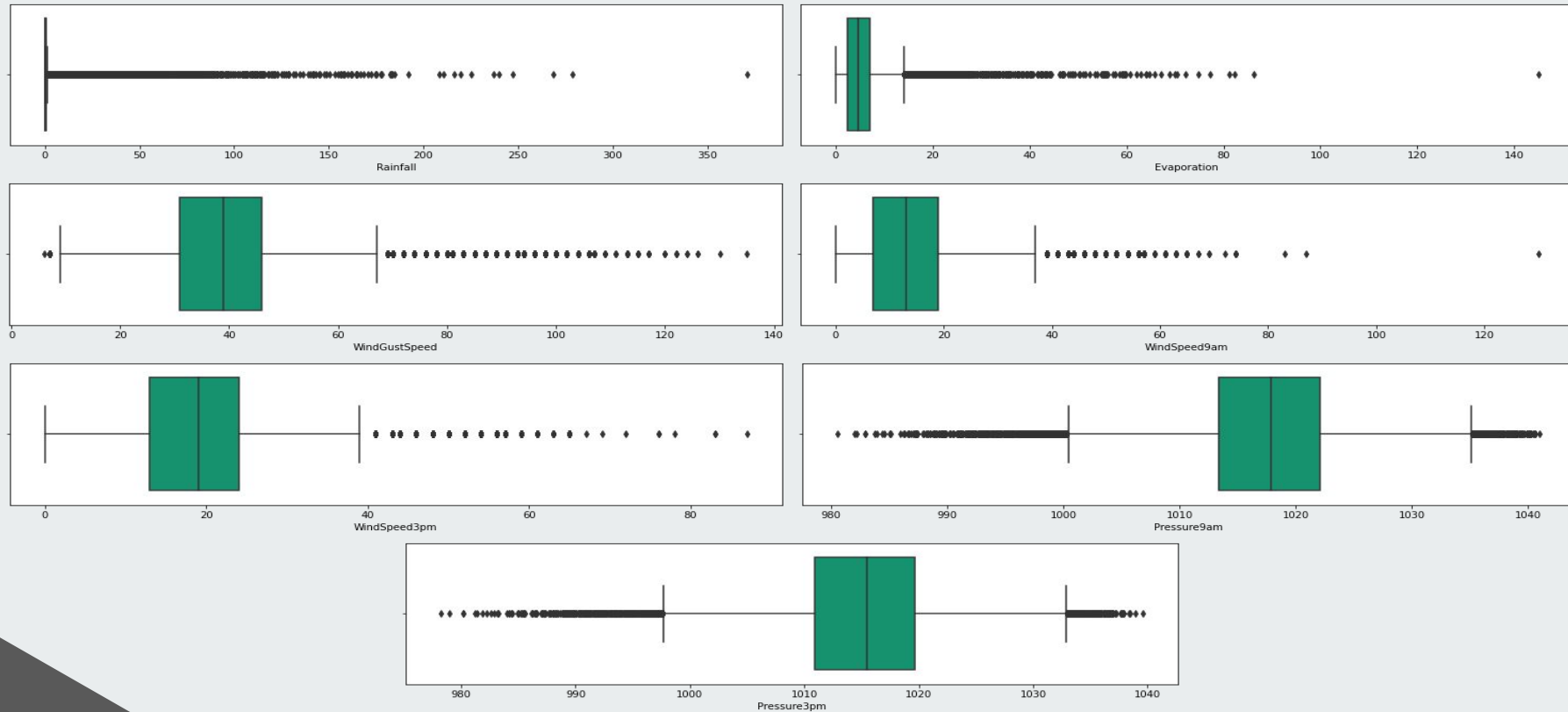
Fill missing value by **median** value for **numerical** data and **mode** value for **categorical** data based on each **Location** and each **Month**.

Exploratory Data Analysis (EDA)

03

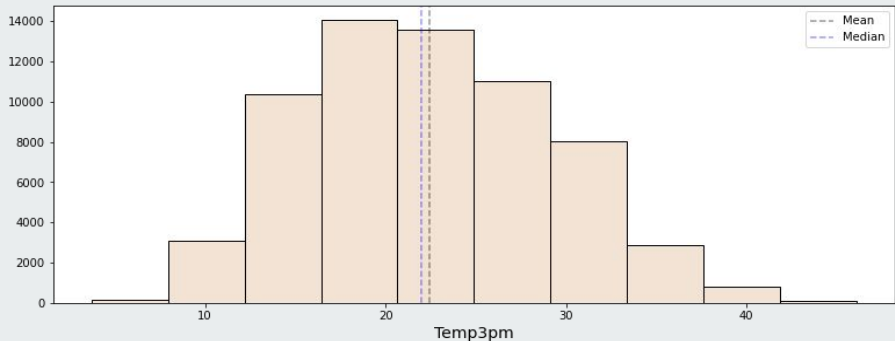
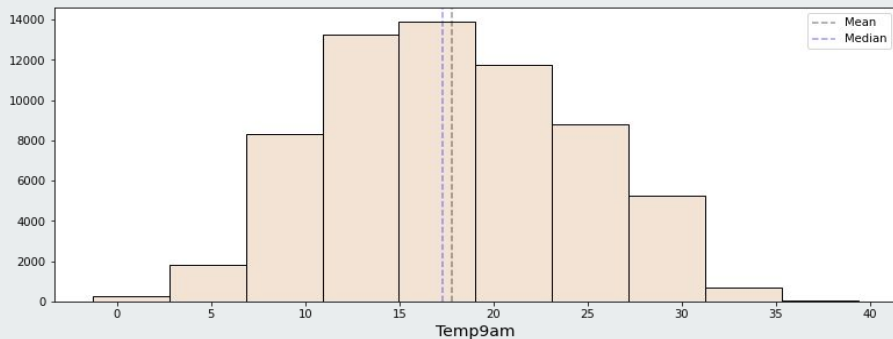
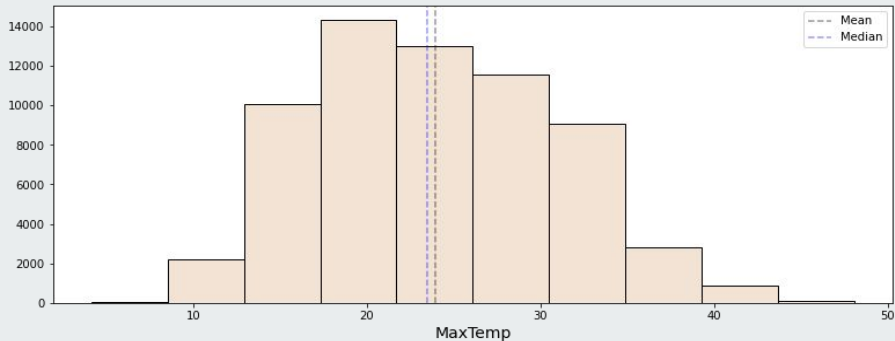
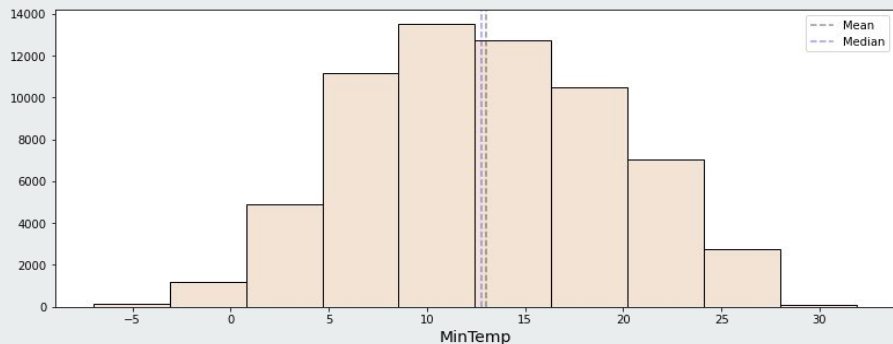


Outliers Detection



Seven features have extreme **outliers** and need to be removed based **IQR** (Interquartile Range) **upper** and **lower** limit.

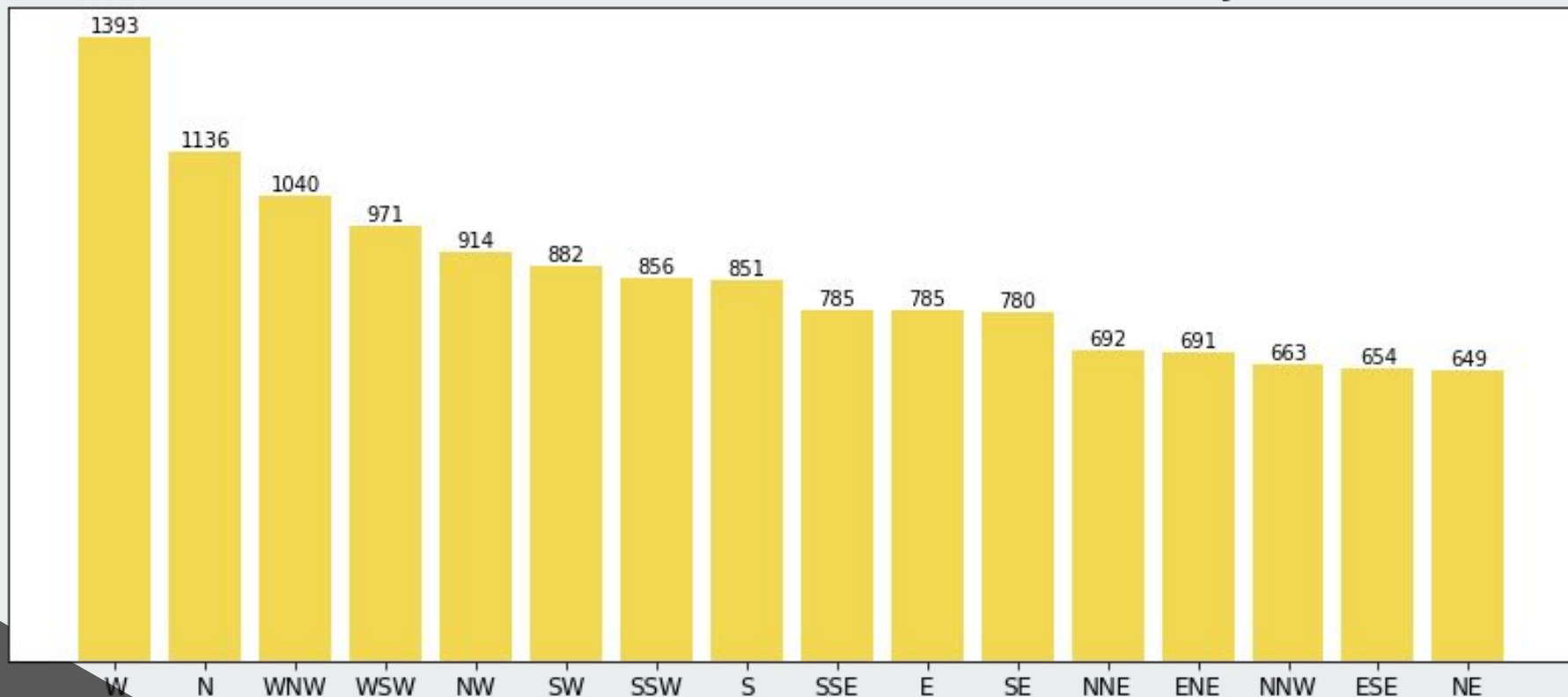
Temperature Aspects



Temperatures in Australia regions have range **around** from **10** until **30 degrees Celsius**. The data **distribution** quietly **normal**. We can conclude temperature in Australia regions relatively **stable**.

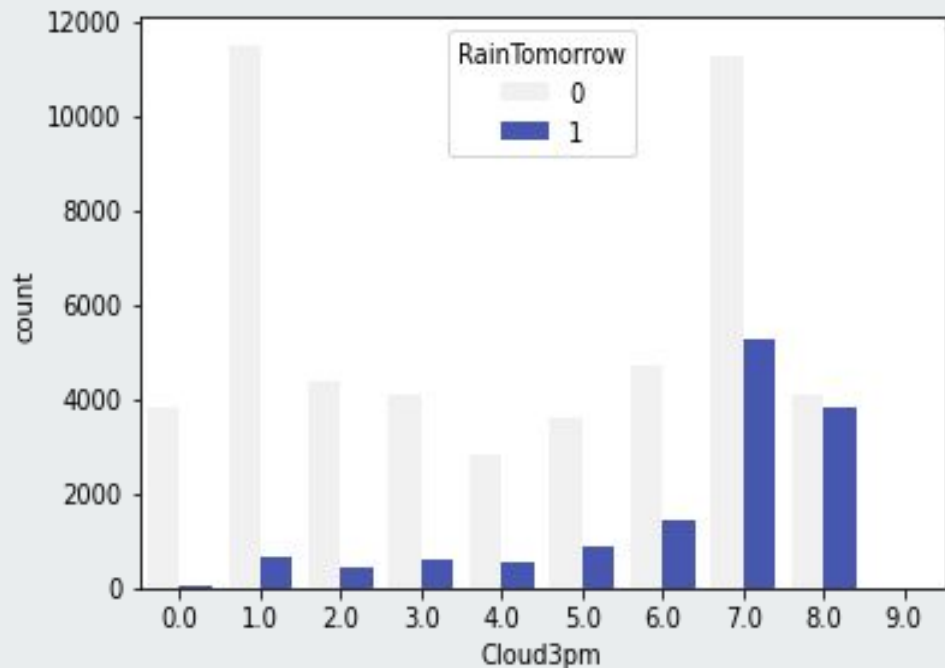
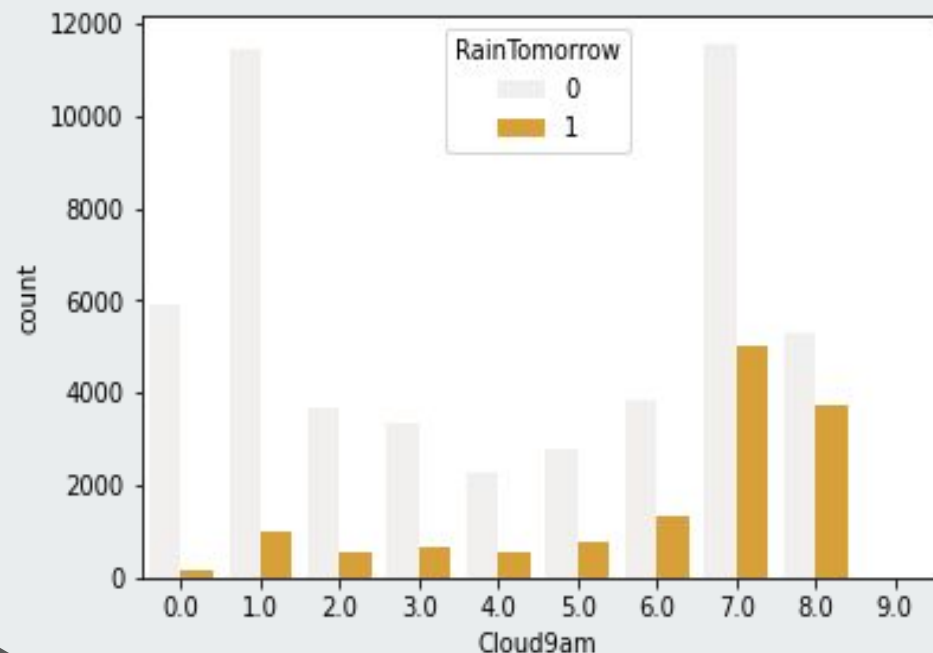
Wind Gust Direction

Wind Gust Direction that Cause Rain in the Next Day



Rain is tend to happens in the next day majority when direction of wind gust in between **West** to **North**.

Fraction of Sky Obscured by Cloud (in Oktas)



Sky in Australia mostly **cloudy** (Fraction of sky obscured by cloud is **7 oktas**).
Rain in the next day tend to happen when **cloud** in **9am** and **3pm** is **7** or **8 oktas**.

(Reference: [Worldweather](#))

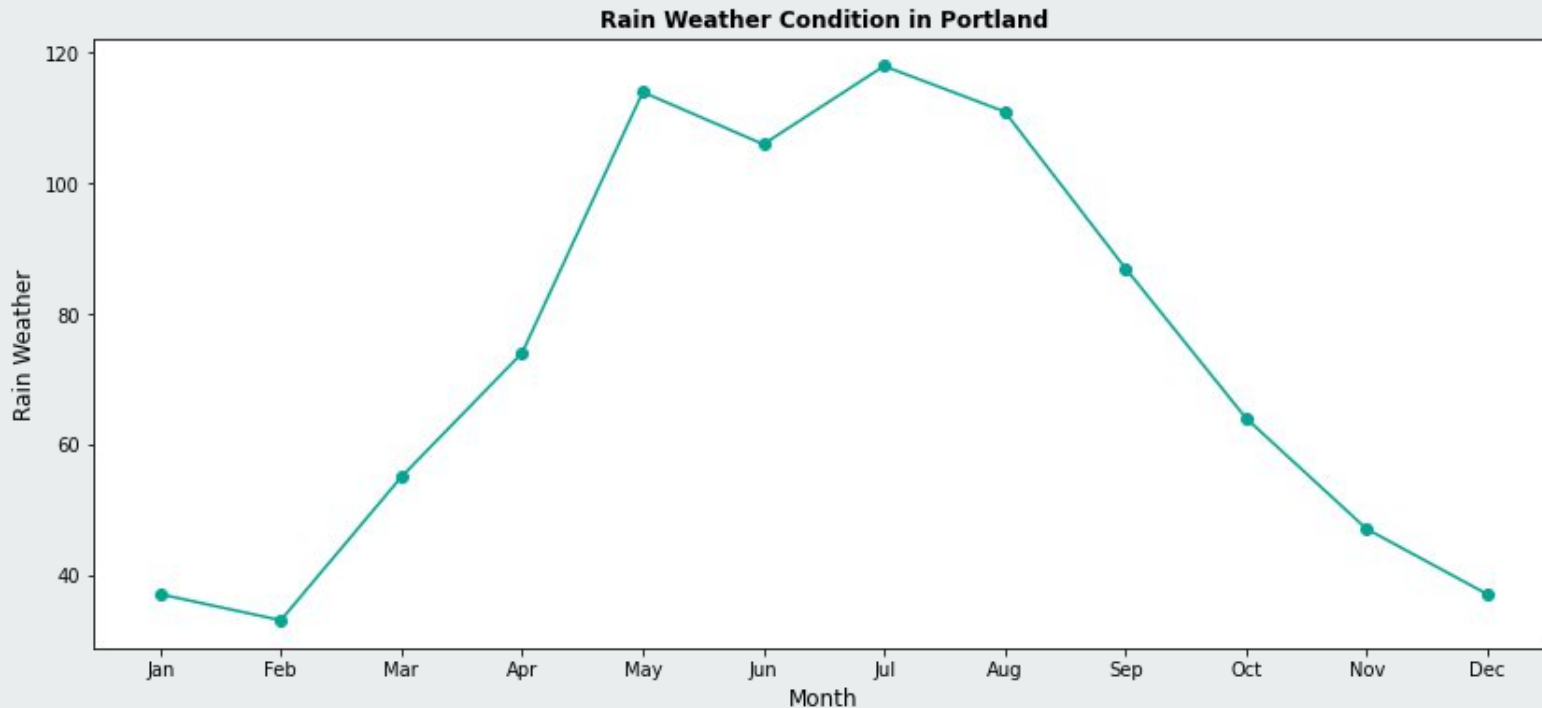
Rain Condition Based on Locations

Top 10 Location With the most experience rain



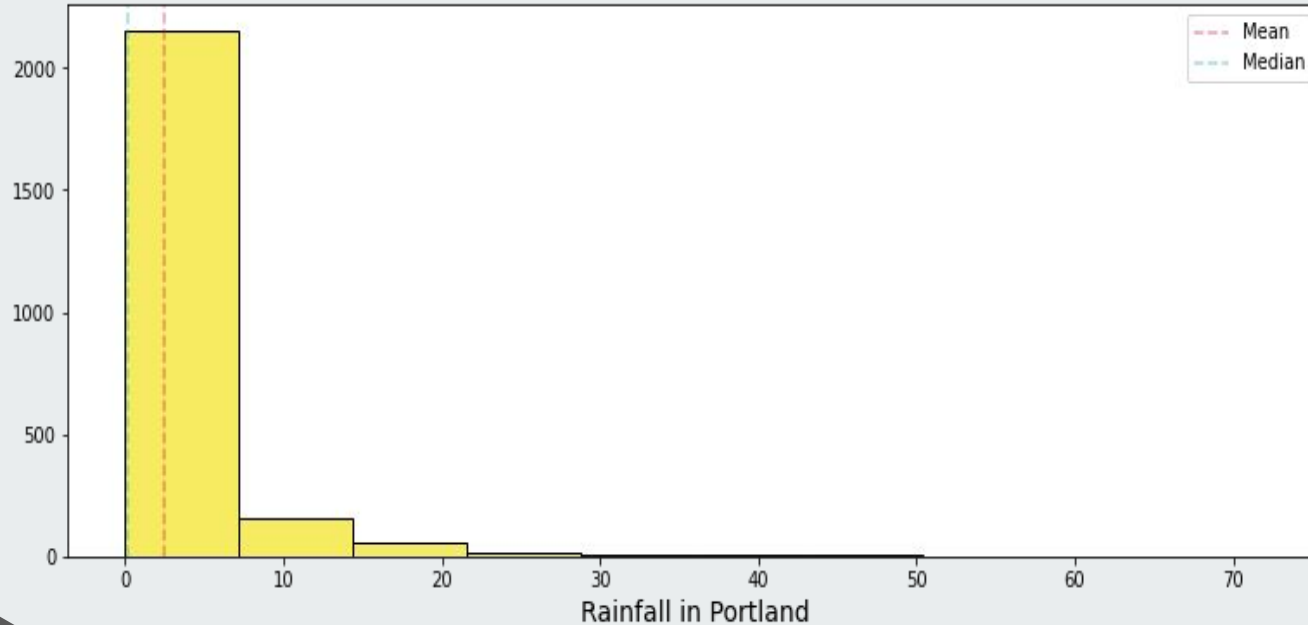
Portland is the location in Australia that experiences the most rain.

Rain Condition in Portland



Rain in Portland is most common to happen from May to August. We can conclude that these period is the rainy season in Portland.

Amount Rainfall in Portland



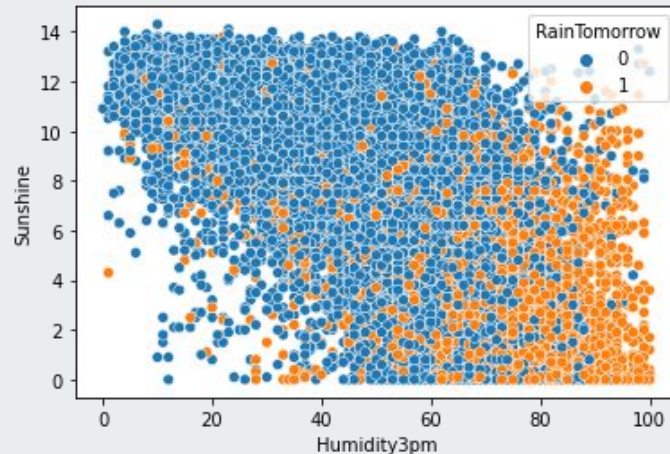
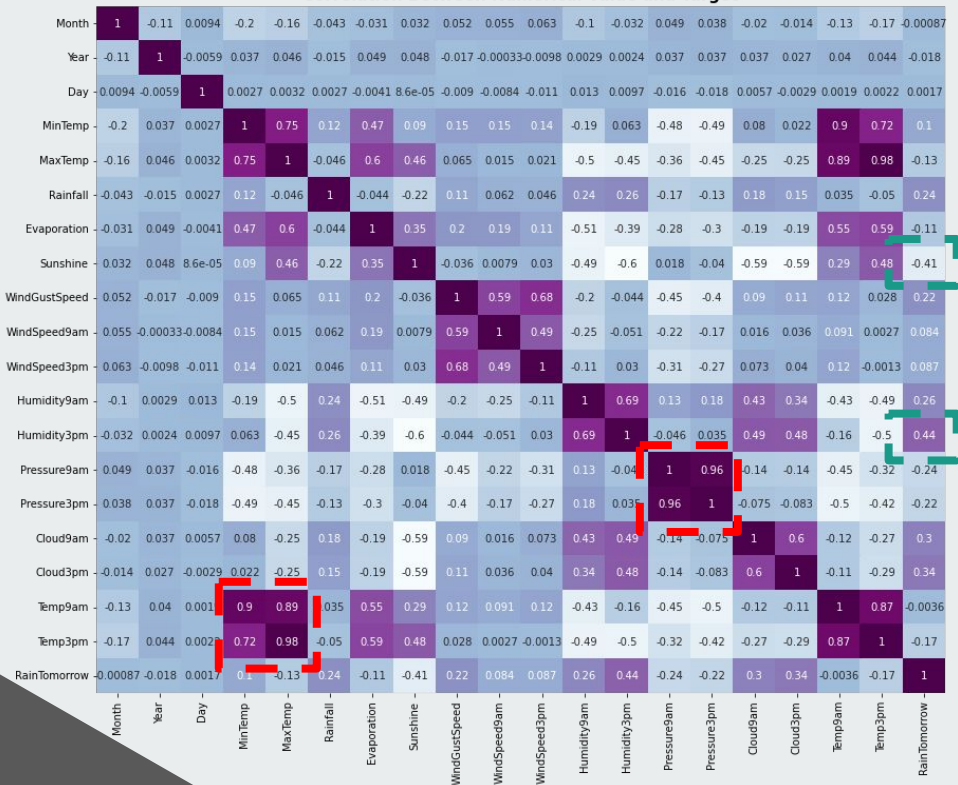
Mean : 2.49 mm

Median : 0.2 mm

Amount of **Rainfall** in Portland majority in range around **0 - 8 mm** with average of rainfall amount is **2.49 mm**

Multivariate Analysis

Correlation Between Numerical Value and Target



- **Humidity3pm** (positive correlation), and **Sunshine** (negative correlation) are the features that have the **highest correlation** with **target**.
- There are **multicollinearity** that correlated with **Temperature** and **Pressure** aspects.
- **Pressure9am** and **Temp3pm** have **higher correlation** with **target**. These two features will be kept for modelling.
- When **Humidity3pm** ratio is high and number of hours of bright **Sunshine** is low, **rain** in the next day tend to **happen**.

Feature Engineering

04



Feature Engineering

| Aspects | Action |
|----------------------|---|
| Categorical features | Categorical feature with 2 distinct values -> Binary Encoding Categorical with more than 2 distinct values -> One Hot Encoding |
| Drop Features | Multicollinearity -> Pressure3pm, MinTemp, MaxTemp, and Temp9am Don't contribute for modelling -> Year |
| Imbalance Data | Resampling use Undersampling |
| Scaling | Min-Max Scaler |

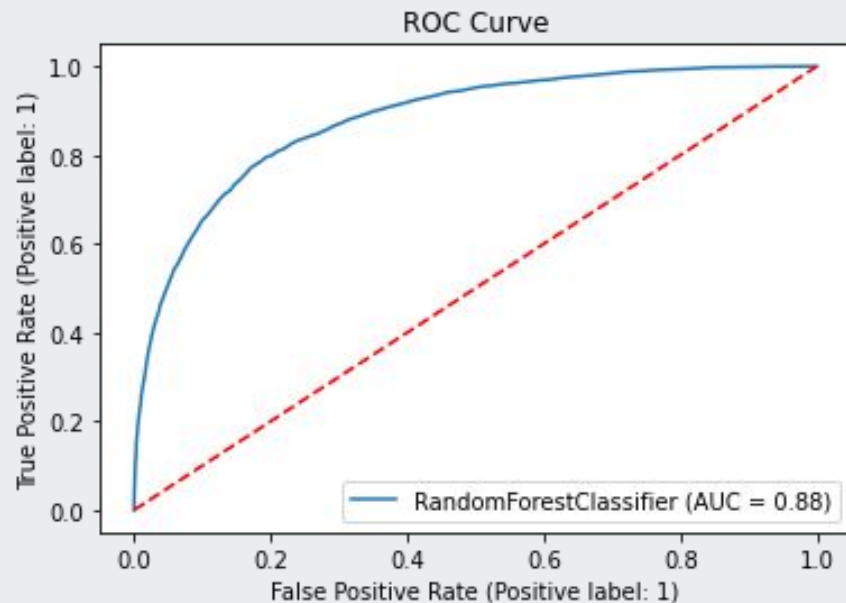
Modelling and Evaluation

05



Modelling

| Model | AUC |
|---------------------|-----|
| K-Nearest Neighbors | 78% |
| Logistic Regression | 87% |
| Decision Tree | 71% |
| Random Forest | 88% |
| XG-Boost | 87% |



I split dataset become **train** and **test** data with **80:20** proportion. Since the data is **imbalance**, I use **AUC** (Area Under Cover) instead.

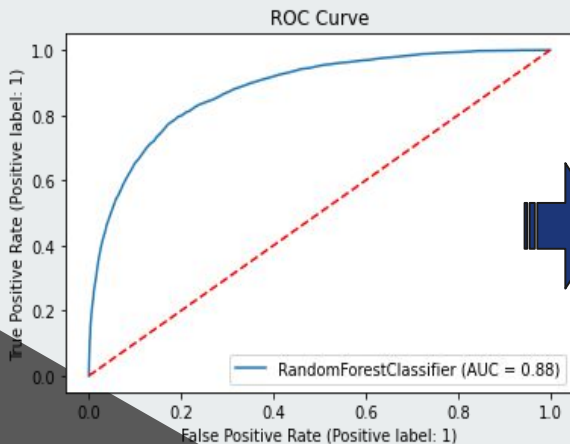
Random Forest have the best **AUC** score compared than other models.

Hyper Parameter Tuning

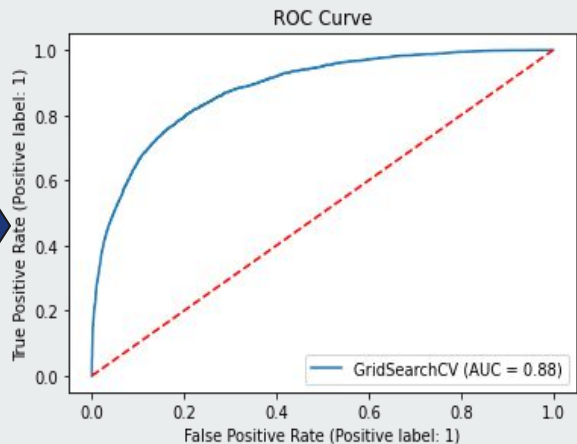
| Random Forest Algorithm | AUC |
|-------------------------|-----|
| Before Tuned | 88% |
| After Tuned | 88% |

Overall, the **AUC** score before and after **tuned** is relatively same.

Before Tuned

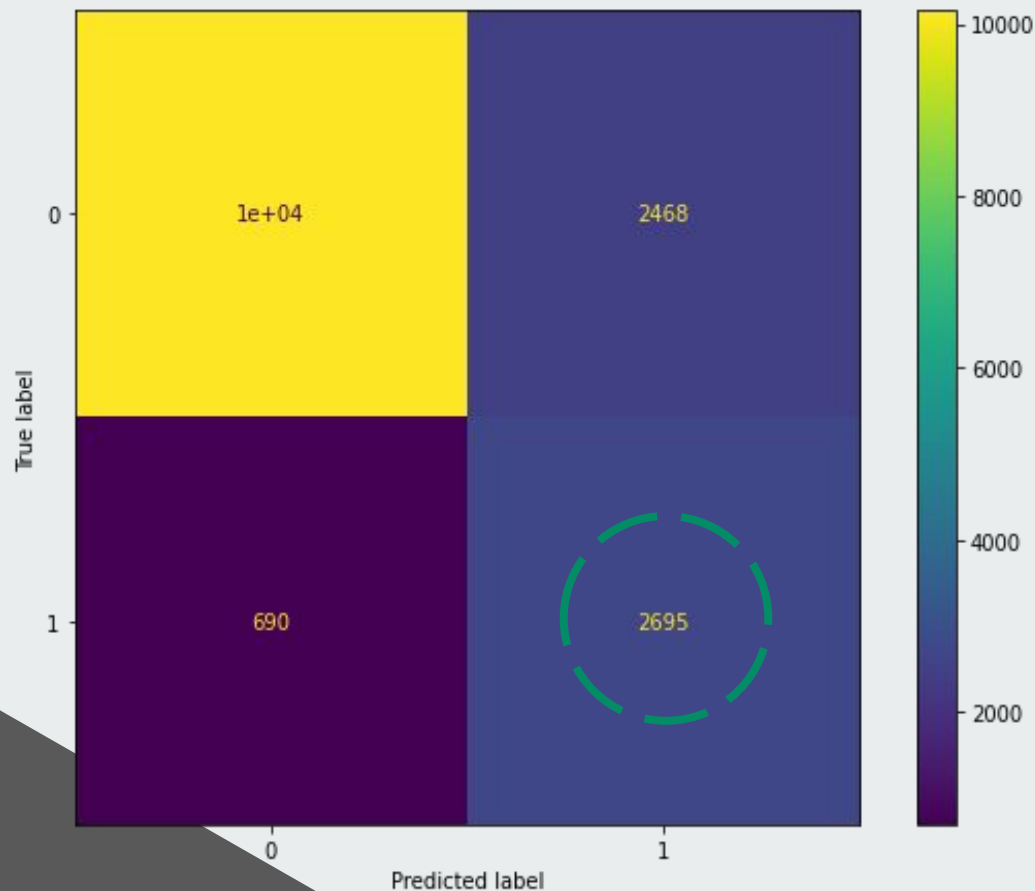


After Tuned



ROC curve from before and after tuned also didn't really have differences. It means before tuned, the model already have optimum performance. But still, **hyperparameter tuning** is essentials part for controlling model behavior. It's also very important to **avoid** model **overfitting**.

Confusion Matrix



From confusion matrix we can analyze model have **True Positive Rate** by **80%**. It means, from **3385 days** when our model predict will be rain, **2695 days** it really rained

Potential Impact in Business

We can do simulation our predictive model for business in the agricultural sector in Australia. For example, in irrigation activities.



Before

Water needs = 155.5 kl/month
Price of water = 0.26 USD/m³
Cost for water = 40.43 Millions AUD

After

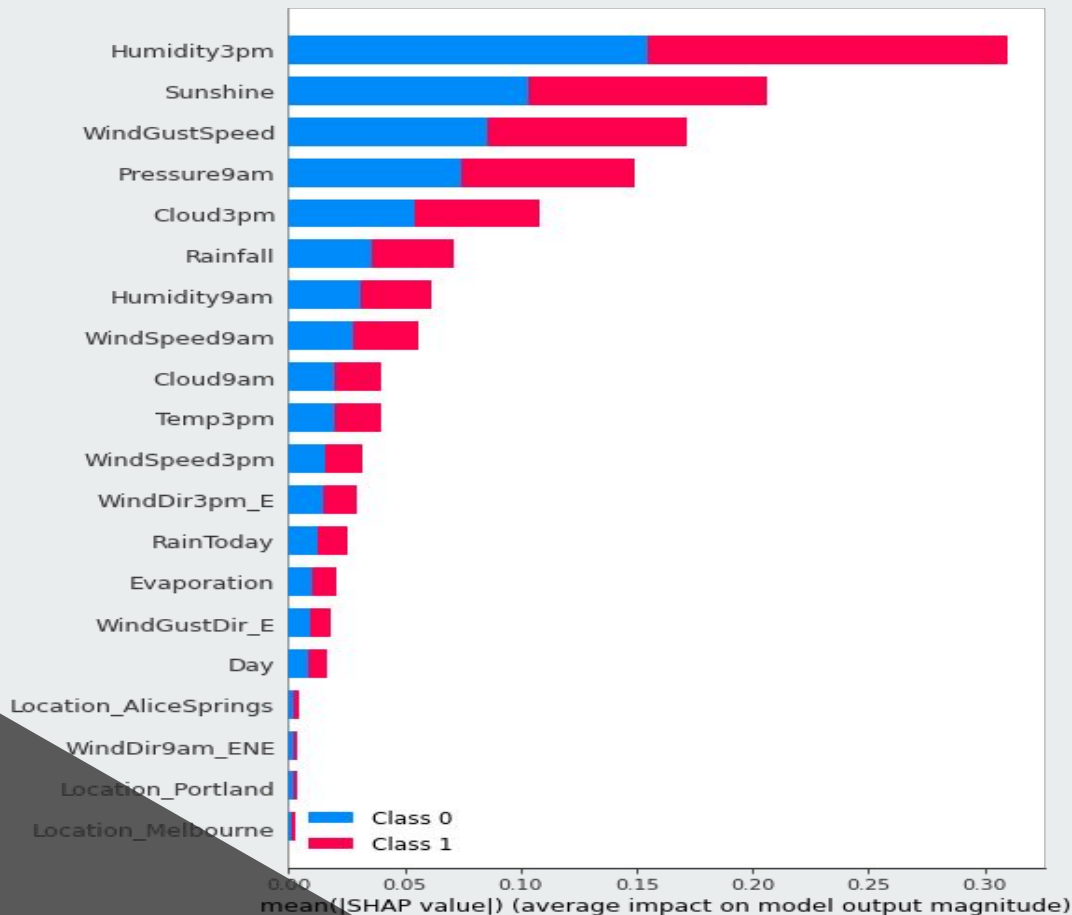
Water needs = 95.7 kl/month
Price of water = 0.26 USD/m³
Cost for water = 24.89 Millions AUD

SAVING

\$ 15,540,000 (38%)

on water supply for irrigation per month

Feature Importances



Based on SHAP value, Humidity3pm and Sunshine feature have the highest effect on the prediction. It's same with heatmap correlation.

Conclusion and Recommendation

06



Conclusion

- Based on Heatmap and SHAP values for feature importances, Humidity at 3pm and Sunshine has the big impact to cause rain in the next day.
- Random Forest classifier is the best model algorithm for predicting rain the next day because have the highest AUC score than other classifier algorithm.
- The location with the most rain frequency is Portland with rain season tend to happen in May until August.
- Based on simulation, model performance can help saving company cost for water supply by 38%.

Recommendation

If we want to start running the company that operating in agricultural sector, I think Australia regions is become one of the good choice since the temperature is relatively stable. Many agricultural products that can grown well based on temperatures range. Also, I think Portland is the best location due to high rain frequency in a year, because it's very helpful for doing farming activities, such as planting and irrigation.

The best time for harvesting and make as much water stock is May until August. So, when the dry season comes, we won't too worried about lack of water and still can do activities like irrigation and farm will not easily to drought.

As a model performance test, it would be better if the model was applied in Portland as the location with the highest rainfall frequency.



Thank You

Script: <https://colab.research.google.com/drive/1fejSHbTbXUMvdqn-Qor6jRWV223WthXq?usp=sharing>