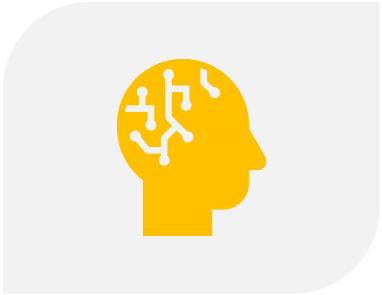
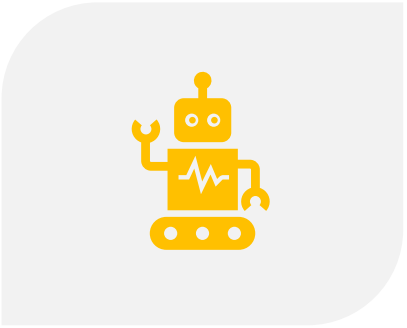
# Weather forecast in Australia

# Our Team

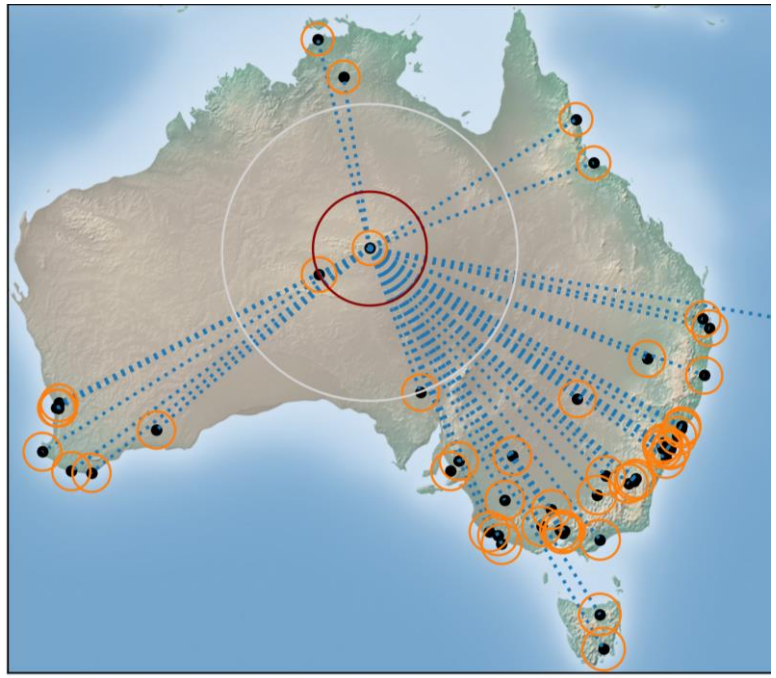


JAN PAUL (PHYSICIST, PH.D.)



KATARINA (ENGINEER, PH.D.)

KEVIN (MATH MODELING)



# Motivation

- Enhancing Preparedness and Resilience
- Optimizing Resource Allocation
- Supporting Economic Growth and Development
- Advancing Scientific Knowledge and Technological Innovation
- Empowering Decision-Makers and the General Public

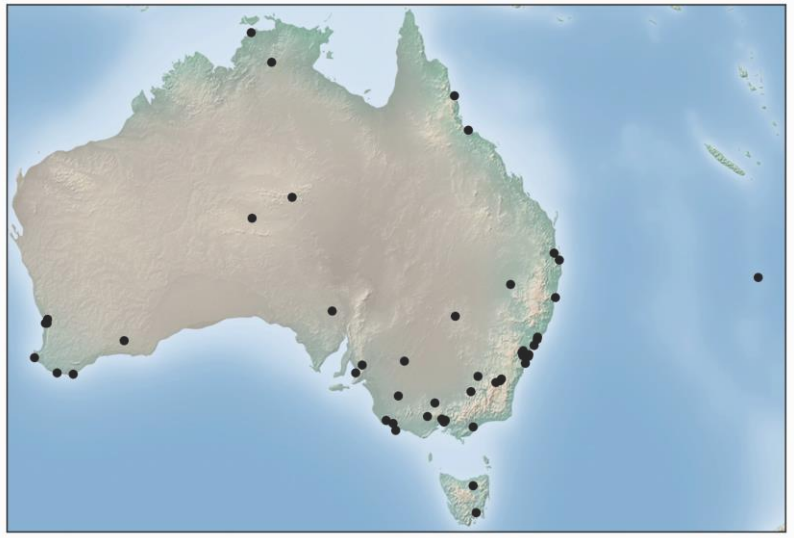
# Goals

- Leveraging AI and mathematical techniques
- Targeting agricultural businesses, insurance companies, energy distribution companies, and airlines
- Providing daily weather conditions
- Revolutionizing customizable, efficient, reliable, and highly accurate long-term predictions across Australia

# Federal Bureau of Meteorology BOM of Australia

| Date | Location | Min Temp | Max Temp | Rain fall | Evaporation | Sun shine | Wind Gust Dir | Wind Gust Speed | Wind Dir 9am | Wind Dir 3pm | Wind Speed 9am | Wind Speed 3pm | Humidity 9am | Humidity 3pm | Pressure 9am | Pressure 3pm | Cloud 9am | Cloud 3pm | Temp 9am | Temp 3pm | Rain Today | Rain Tomorrow |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2010-04-02 | Mount Gambier | 10.4 | 20.1 | 0.4 | 1.6 | 10.1 | S | 39 | SE | SSE | 20 | 24 | 84 | 47 | 1024.9 | 1023.4 | 2 | NaN | 15 | 18.9 | No | No |
| 2010-04-03 | Mount Gambier | 9 | 19.4 | 0 | 4.8 | 2.6 | SSE | 43 | SSE | SSE | 22 | 26 | 62 | 56 | 1024.8 | 1022.6 | 7 | 7 | 14.5 | 17.9 | No | No |
| 2010-04-04 | Mount Gambier | 9.6 | 25.3 | 0 | 3 | 7.8 | SSW | 30 | ENE | NNE | 15 | 7 | 81 | 39 | 1020 | 1015.5 | 7 | 7 | 15.6 | 24.2 | No | No |
| 2010-04-05 | Mount Gambier | 9 | 27.6 | 0 | 3.4 | 9.7 | SSE | 30 | NE | SE | 13 | 11 | 55 | 28 | 1015.6 | 1011.9 | 6 | 4 | 18.3 | 26.6 | No | Yes |
| 2010-04-06 | Mount Gambier | 15.7 | 17.6 | 2 | 5.4 | 0.2 | ENE | 57 | NNE | ENE | 15 | 43 | 91 | 88 | 1010.5 | 1006 | 8 | 7 | 16.7 | 15.9 | Yes | Yes |

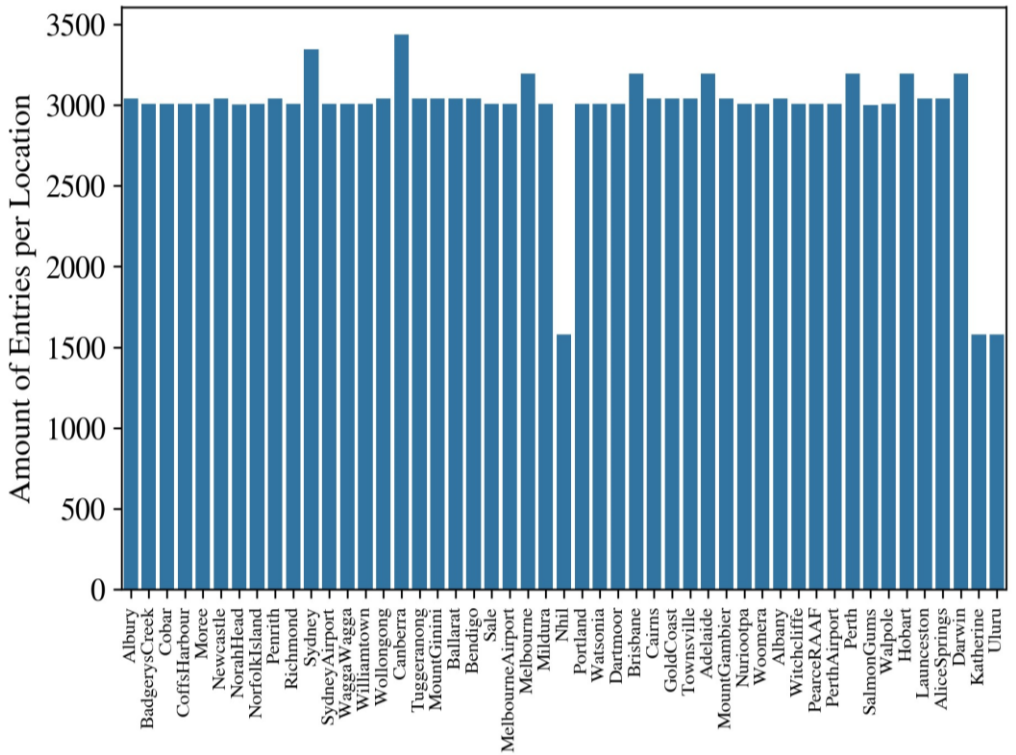
https://www.kaggle.com/jsphyg/weather-dataset-rattle-package // http://www.bom.gov.au/climate/data/

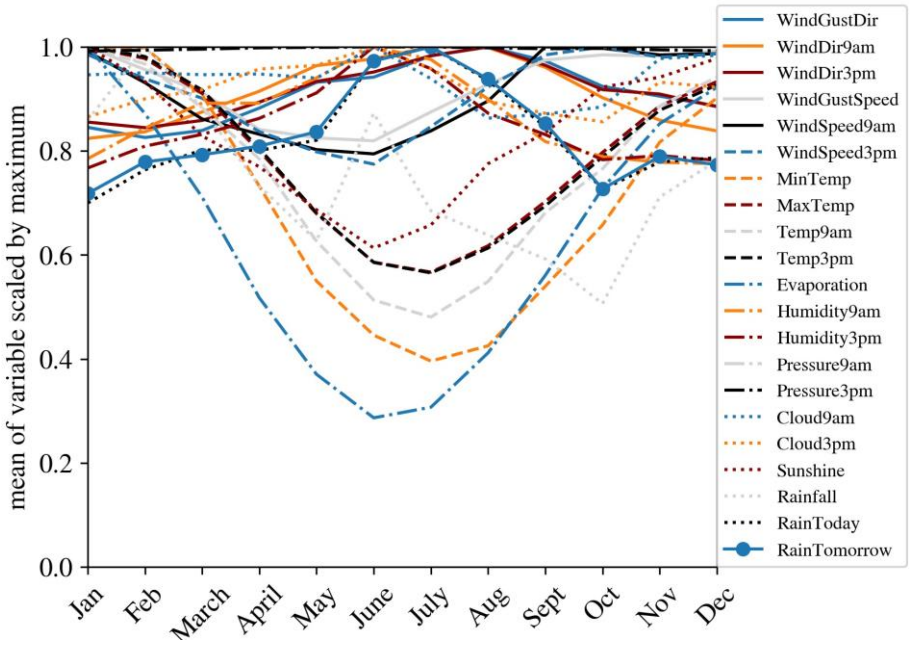


# I Development step´s: preprocessing

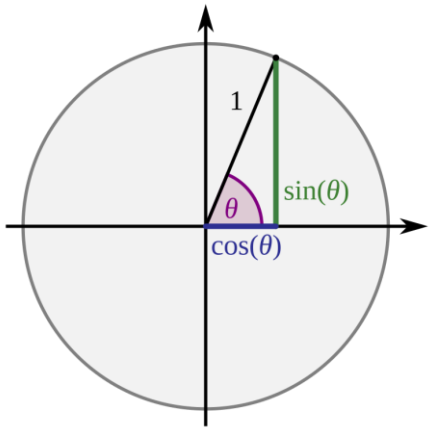
# Data cleaning

- Categorical variables RainToday/RainTomorrow:
  0 for no rain and 1 for rain
- Remove locations with minimal data:
  Exclude Uluru, Katherine, and Nhil
- Change variable type to datetime
- Perform feature engineering:
  Split the date into three new variables
- Model seasons in Australia: Account for summer/winter
- Handle missing values: Drop missing values, replace with mean for numerical variables, and mode for categorical variables.



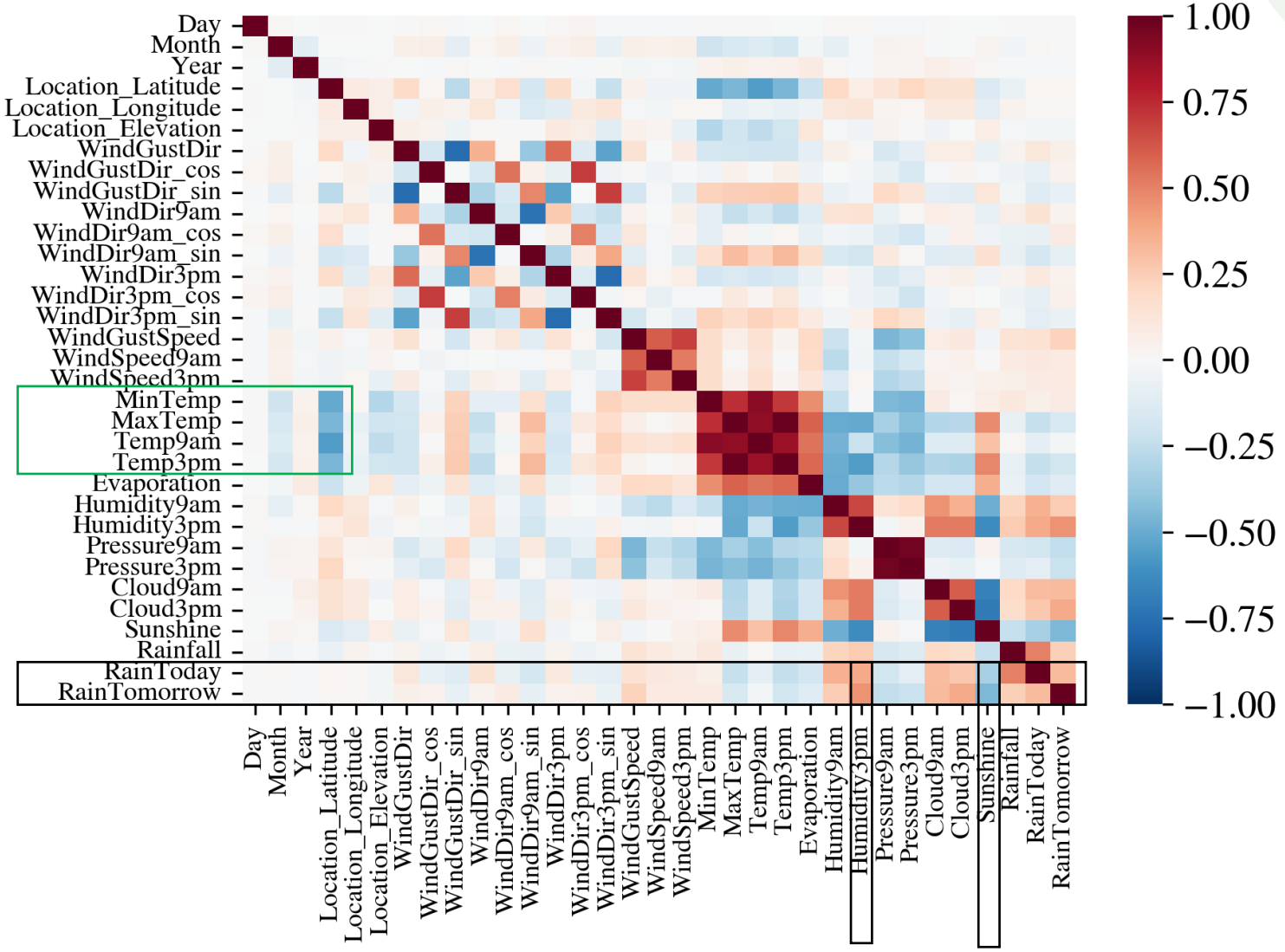



# Create Features

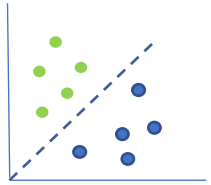
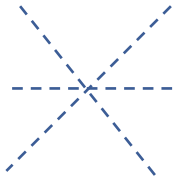



- WindDir: Create sin and cos features to represent compass directions as periodic values between 0 and 2pi
- Longitude/Latitude: Retrieve coordinates and elevation for each location from the BOM website, considering climate zones and their impact on temperature and seasons
- Train/Test Split Preparation: Handle categorical variables by dropping or replacing them, sort data by date, perform a time-series split without shuffling, and apply standard scaling to ensure comparability of data from different stations
- Imbalance in Target Variable: Address the imbalance (75% no rain, 25% rain) through different approaches, such as oversampling, undersampling, or no specific action

# Heatmap


Day
Month
Year
Location_Latitude
Location_Longitude
Location_Elevation
WindGustDir
WindGustDir_cos
WindGustDir_sin
WindDir9am
WindDir9am_cos
WindDir9am_sin
WindDir3pm
WindDir3pm_cos
WindDir3pm_sin
WindGustSpeed
WindSpeed9am
WindSpeed3pm
MinTemp
MaxTemp
Temp9am
Temp3pm
Evaporation
Humidity9am
Humidity3pm
Pressure9am
Pressure3pm
Cloud9am
Cloud3pm
Sunshine
Rainfall
RainToday
RainTomorrow
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

# Modeling and solving the problem using Machine Learning techniques

| | | Adventages | Disadventages |
|---|---|---|---|
| ▪ Logistic Regression |  | • Simplicity and interpretability of results | • Limited ability to capture complex non-linear relationships |
| ▪ Random Forest Classifier (Bagging, imblearn BalancedRF) |  | • Ability to handle large datasets with high dimensionality<br>• Improves the performance of the model on the minority class by oversampling | • Can be computationally expensive for large datasets and complex models |
| ▪ Support Vector Machine (SVM) |  | • Effective in high-dimensional spaces and cases with a clear margin of separation | • Less efficient with large datasets |
| ▪ Neuronal Networks Keras |  | • Provides a wide range of pre-built layers and model | • May require substantial amounts of data to avoid overfitting |

# I modelling results

| | Logistic Regression | | Support Vector Machine (SVM) | | Random Forest Classifier | | Bagging for Random Forest Classifier | | Gaussian Naive Bayes | | Random Forest Classifier (NaN replaced, no Sampling) | | Random Forest Classifier (NaN dropped, over sampling) | | Deep Learning - Keras | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target class: 0/1: no/ rain | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Accuracy | 0,78 | | 0,80 | | 0,85 | | 0,85 | | 0,79 | | 0,85 | | 0,86 | | 0,85 | |
| Precision | 0,92 | 0,51 | 0,93 | 0,53 | 0,87 | 0,76 | 0,87 | 0,76 | 0,88 | 0,54 | 0,87 | 0,76 | 0,88 | 0,73 | 0,87 | 0,76 |
| Recall | 0,79 | 0,76 | 0,80 | 0,79 | 0,95 | 0,51 | 0,96 | 0,50 | 0,86 | 0,58 | 0,95 | 0,51 | 0,94 | 0,55 | 0,95 | 0,51 |
| F1-Score | 0,85 | 0,61 | 0,86 | 0,64 | 0,91 | 0,61 | 0,91 | 0,60 | 0,87 | 0,56 | 0,91 | 0,61 | 0,91 | 0,63 | 0,91 | 0,61 |

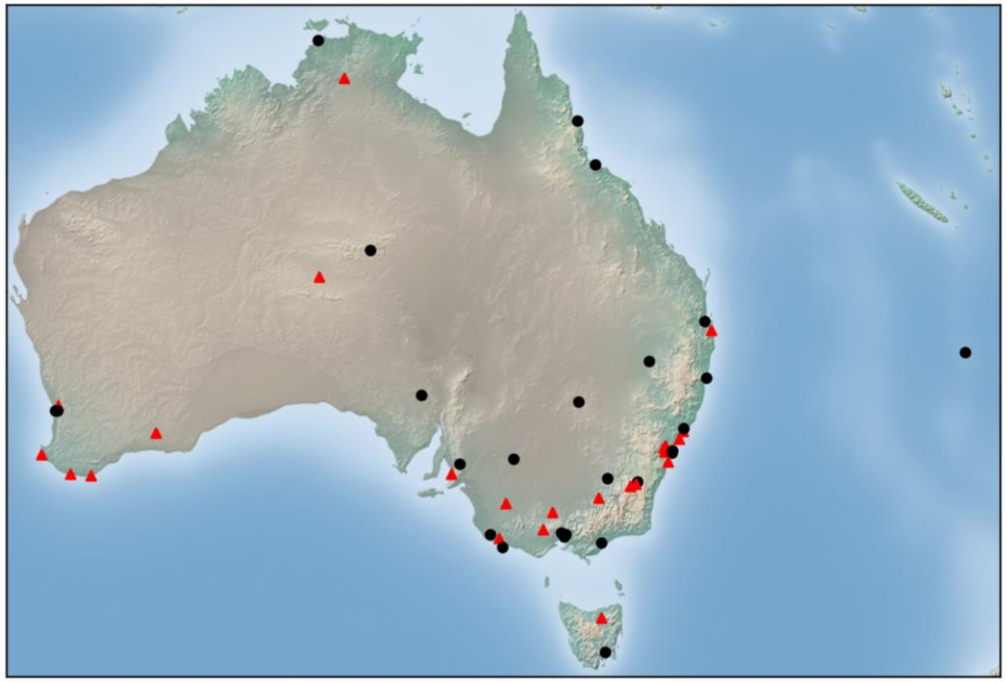
| feature | importance |
|---|---|
| Rainfall | 0.04 |
| Humidity9am | 0.04 |
| WindGustSpeed | 0.05 |
| Pressure9am | 0.05 |
| Pressure3pm | 0.05 |
| Cloud3pm | 0.06 |
| Humidity3pm | 0.14 |
| Sunshine | 0.14 |

# II Development steps: preprocessing

# Preprocessing II – Remove Irrelevant Data…

- … in space:
  - Drop locations lacking at least one feature: 26 stations remain
- … in time:
  - Drop data from before 2009-01-01

Red triangles: dropped locations

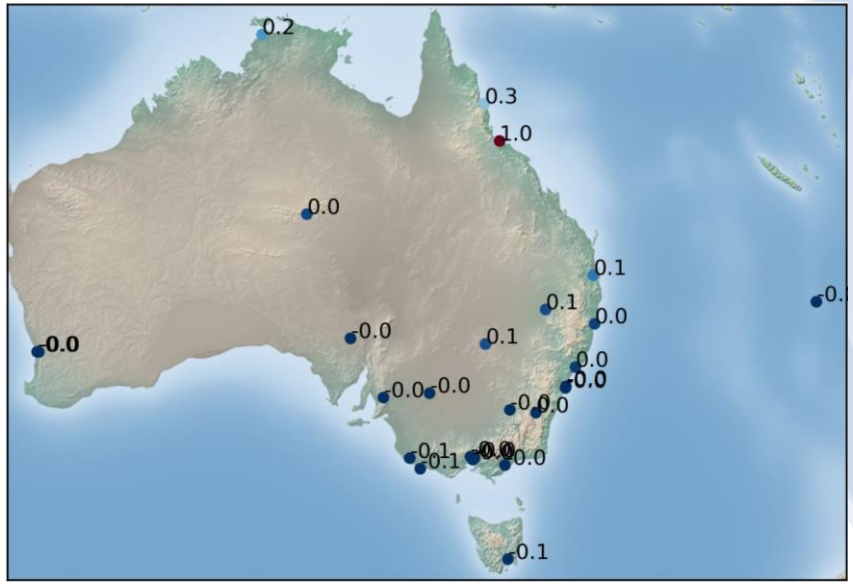


# Preprocessing II – Enhanced Feature Engineering

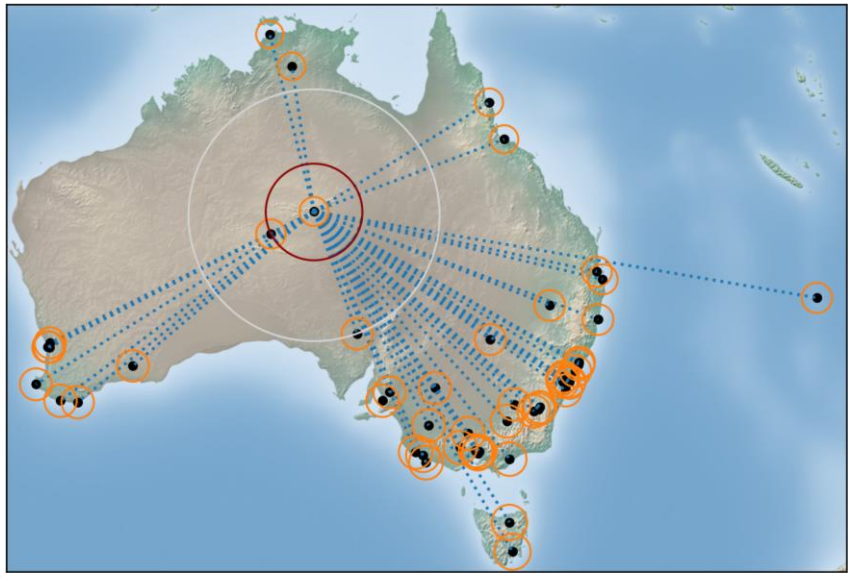
Add information from nearby stations assuming these stations share similar information / weather conditions

Metrics to classify stations as “nearby”:

- Correlation based on *Rainfall* feature




- Geometric distance: radius around each station



- For each feature: create a new feature *“avg_feature”* based on weighted-average over nearby stations
- Use *avg_features* as additional features
- Use *avg_features* to replace missing values in corresponding features

# II Modelling Results

- Best metric for *avg_features*: 100km distance
- Handling imbalance: Oversampling
- Scaling: Standard Scaler

- new preprocessing does not improve the performance
- More complex deep learning algorithms are required and might benefit from enhanced feature engineering (*avg_features)*

- *avg_features* are relevant, although they do not improve the performance as anticipated

| | **Random Forest Classifier (preprocessing II)** | | **Random Forest Classifier (preprocessing I)** | |
|---|---|---|---|---|
| | 0 | 1 | 0 | 1 |
| Accuracy | 0.86 | | 0.86 | |
| Precision | 0.89 | 0.72 | 0.88 | 0.73 |
| Recall | 0.94 | 0.55 | 0.94 | 0.55 |
| F1-Score | 0.91 | 0.63 | 0.91 | 0.63 |

| **feature** | **importance** |
|---|---|
| avg_Pressure3pm | 0.03 |
| Cloud3pm | 0.03 |
| avg_WindGustSpeed | 0.03 |
| avg_Cloud3pm | 0.04 |
| Sunshine | 0.06 |
| avg_Sunshine | 0.07 |
| avg_Humidity3pm | 0.08 |
| Humidity3pm | 0.08 |

| **feature** | **importance** |
|---|---|
| Rainfall | 0.04 |
| Humidity9am | 0.04 |
| WindGustSpeed | 0.05 |
| Pressure9am | 0.05 |
| Pressure3pm | 0.05 |
| Cloud3pm | 0.06 |
| Humidity3pm | 0.14 |
| Sunshine | 0.14 |

# Conclusions and Outlook

- We achieve the **best possible performance** given the **limitations of the available data** (low spatial / temporal resolution) by:
  - Exploring **extensive feature engineering** routes (e.g. features based on weighted-average over nearby stations)
  - Implementing **Machine- and simple Deep-Learning models**
- With a **Random Forest Classifier**, we are able to **predict** if it will **rain on the next day** with an **accuracy of 0.86**, a macro-averaged F1-score of 0.77, and a weighted-averaged F1-score of 0.85 **on the test data set**

- Based on the results, the project can be improved and extended in the **future**:
  - Use meteorological **data of higher resolution in time and space** (e.g. gridded satellite data with an hourly and 100km resolution, respectively.
  - Implement a **sophisticated recurrent LSTM-based neural network or a physics-informed neural network** (including meteorological and climatological equations)
  - Predict other meteorological variables (e.g. temperature) and do **longterm predictions**

# Demo: Streamlit Application