Weather forecast in Australia

Geoffrey Foulon-Pinto, Pharm.D, Ph.D. & Lise Aujoulat, M.Sc.

January 17, 2022

DataScientest

Title: Weather forecast in Australia.

Abstract: Weather prediction is an integral part of our daily lives and the reliability of predictions is crucial. This project aimed at designing a supervised learning machine learning model to determine if precipitation is to be expected for the day following observations. The dataset used contained meteorological records from 49 stations spread across Australia, i.e. 145,460 dated records, over a 10-year period (2007 - 2017).
The raw database included 22 variables (temperature, atmospheric pressure, humidity, etc.) in addition to the variable to be predicted. Entries with a missing value on the predictor variable have been removed. The other missing values were replaced by the average, median or mode value over the current month, for each weather station. Data were adjusted by oversampling and undersampling.
Among the various machine learning models we developed, Random Forest was the model that had the best performance, with an accuracy of 83%, a recall of 63% and an f1 score of 62%. A deep learning model using 2 dense layers with 25 and 50 neurons (activation by rectified linear unit function) has also been developed, offering an accuracy of 86%. Finally, a time series study was carried out on the maximum temperature using another data set (with data from 1995 to 2021) in order to explore the diversity of what can be produced using meteorological data. The predictions that have come out for the year 2021 were satisfying
In conclusion, this data science project shows that machine learning and deep learning algorithms are promising tools in weather forecasting. Field knowledge is essential in this kind of study and it would indeed be interesting to take more global phenomena such as highs or other data such as satellite images into account to obtain better predictions.

Keywords: Data science, Machine learning, Deep learning, Time series, Weather forecasting.

Acknowledgement

We would like to thank the entire DataScientest team for the quality of the training provided and in particular the diversity of educational content made available.
We would particularly like to thank Maxime Michel for his invaluable advice, his critical outlook and his availability in supervising this project, and Gaspard Grimm, our promotion manager, for his kindness and his investment in the success of the promotion.
Finally, we would also like to thank all those who animated the set of masterclasses that we had the opportunity to attend.

Contents

1 Introduction

Meteorology is the science concerned with atmospheric phenomena (cloud formation, precipitation, etc.) and mankind has been interested in it since at least Antiquity. Due to the central place of agriculture in the development of different civilizations, it quickly becomes necessary to understand and anticipate weather phenomena. From the first sketches of anemometers made in the first century BC to 17th century barometers, weather forecasting today has a large number of powerful tools and techniques to study these weather phenomena. With an increasing number and variety of weather data available, designing machine learning and deep learning algorithms could be a very suitable approach for weather forecasting.

We carried out this project as part of a data science training course, provided from October 2021 to January 2022 by DataScientest. This project was an opportunity for us to put into practice the theoretical knowledge acquired during this training by working on real life data. The objective of this project was to produce a model for predicting whether or not precipitation will occur in Australia, using data from the Australian Government's Bureau of Meteorology. The database contained 145,460 entries and will be presented in more detail later in this report.

To meet our objective of producing a rainfall forecast model the day after the observations, we first studied all the data made available and dealt with the missing values. We carried out a pre-processing on the database to have a format suitable for modeling. We designed several machine learning models to determine the most suitable model and compared it to a deep learning model. Through these models, we focused on precision and recall metrics because we wanted to detect rainy days (rarer than dry days in Australia). Finally, the nature of the data allowing it, we carried out a time series analysis. All the programming part was done with the Python 3.10.1 language. The main libraries used were Pandas, NumPy, Matplotlib.pyplot, Seaborn, Scikit-learn and Tensorflow.

2 Data description & data visualization

The dataset used for this project come from `www.kaggle.com` and was provided by Joe Young & Adam Young. According to the authors, this dataset is based on data provided by the Australian government's Bureau of Meteorology.
The raw database contains 145,460 weather record entries. The target variable is *RainTomorrow*, which indicates whether or not it rains the day after the observations. There are 22 features, providing information on temperatures, amount of precipitation and evaporation, humidity, duration of sunshine, cloudiness, wind force and direction, atmospheric pressure and raining status on the recorded day. Several of these variables were recorded twice a day, at 9 a.m. and 3 p.m. These variables will be detailed below. To better understand the value of each variable for the response to the objective, we opted for a comparative description of the variables depending on whether it was raining or not.

2.1 Date and Location

2.1.1 Date

The date of observation. They are in yyyy-mm-dd format and the data come from statements dated 2007 to 2017. There are only 61 entries for the year 2007, 2270 for the year 2008 and 8623 for the year 2017. Regarding other years, there is an average of 16,813 entries.
Missing data: none.

2.1.2 Location

The common name of the location of the weather station. The median number of surveys per location is 3009. Data came from 49 different weather stations located across Australia. The locations with the most records are Canberra (3,436 records) and Sydney (3,344 records).
Missing data: none.

2.2 Temperatures

2.2.1 MinTemp and MaxTemp

The minimum and maximum temperature in degrees Celsius, respectively. These two variables vary in a similar way over the months: the highest temperatures are found in January and the lowest in July. In July, the average minimum and maximum temperatures are 6.9 ± 5.2 C and 16.7 ± 5.2 °C respectively, while in January they are 17.5 ± 4.8 C and 29.5 ± 6.0 C respectively.
Missing data: 1485 (1.0 %) for *MinTemp* and 1261 (0.9 %) for *MaxTemp*.

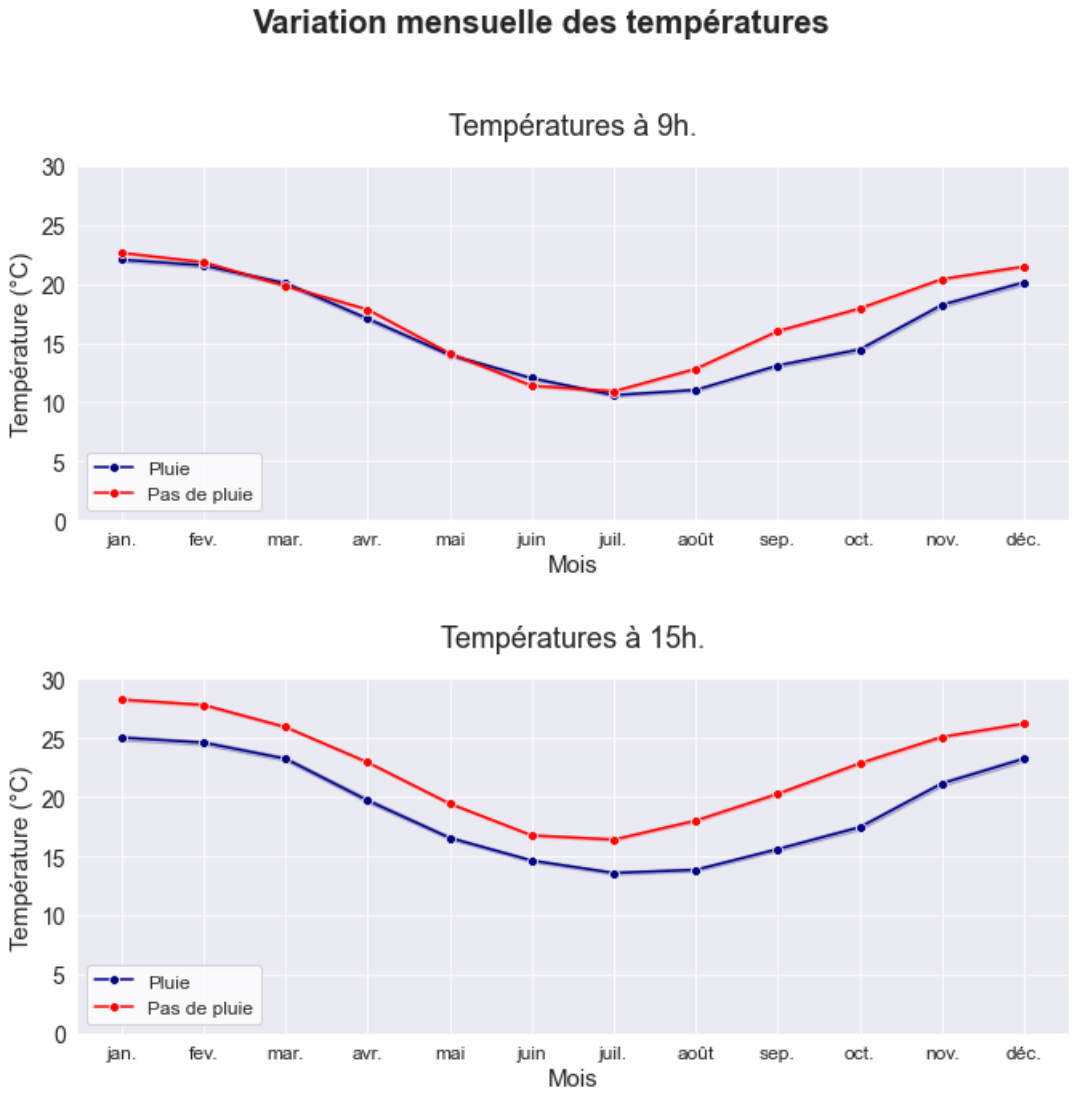
2.2.2 Temp9am

Temperature (degrees C) at 9am. The mean temperature at 9 am is lower on rainy days (15.8 ± 6.1 C) than on days without rain (17.3 ± 6.6 C). The variability is quite similar whether it rains or not and there are many outliers, mainly in the higher values. From one locality to another, the morning temperature seems to vary only slightly, but several northern cities stand out for their higher temperatures: Brisbane, Cairns, Gold Coast, Townsville, Alice Springs, Darwin, Katherine and Uluru.
Missing data: 1767 (1.2 %).

2.2.3 Temp3pm

Temperature (degrees C) at 3pm. The observations at 3 pm follow the same trends as at 9 am. The average temperature at 3 pm is lower on rainy days (18.6 ± 6.3 C) than on days without rain (22.6 ± 6.9 C).
Missing data: 3609 (2.5 %).

Data visualization



2.3 Rainfall, evaporation and humidity

2.3.1 Rainfall

The amount of rainfall recorded for the day in mm. In the database, 21.9 % of the records were for rainy days, with a large variability in the amount of rainfall. Indeed, on rainy days the average rainfall is 10.3 ± 15.5 cm, with a maximum value of 371.0 cm. This variable is directly related to *RainToday*: *RainToday* = 'no' when *Rainfall* $\leq$ 1.0 and *RainToday* = 'yes' when *Rainfall* $>$ 1.0 cm.
Missing data: 3261 (2.2 %).

2.3.2 Evaporation

The so-called Class A pan evaporation (mm) in the 24 hours to 9am. Evaporation has many outliers and seems to be correlated with *RainToday*. Evaporation is more important on days without rain, so this variable seems to behave inversely to *Rainfall* as a function of *RainToday*. The median evaporation height is 3.2 cm, IQR (1.8 - 5.4) on rainy days and 5.2 cm, IQR (3.0 - 7.8) on dry days.
Missing data: 62790 (43.2 %).

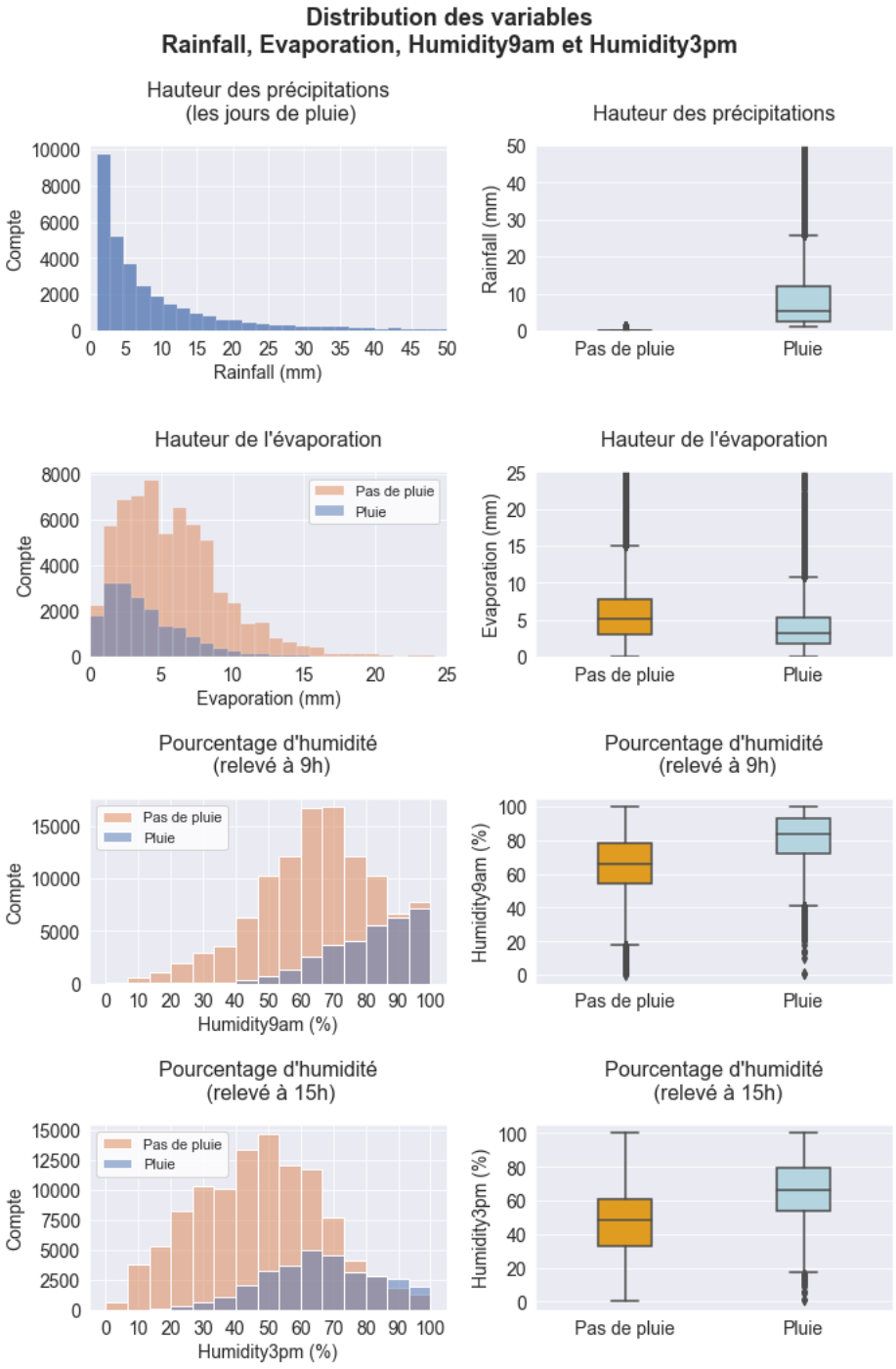
2.3.3 Humidity9am

Humidity (percent) at 9 am. The average percentage of humidity at 9 am is 81.4 ± 13.8 % on rainy days and 65.2 ± 18.8 % on dry days. The humidity at 9 am does not seem to be correlated with the humidity at 3 pm; on the other hand, a correlation with the height of the precipitation could exist (especially for high precipitation), but this is limited by the high variability of the humidity rate. The humidity level at 9 am varies relatively little from one locality to another, with the exception of certain areas located inland (Alice Springs, Woomera, Cobar, Uluru...) which present lower levels. Overall, there are many outliers for the lower humidity values.
Missing data: 2654 (1.8 %).

2.3.4 Humidity3pm

Humidity (percent) at 3 pm. The average percentage of humidity at 3 pm is 66.2 ± 17.5 % on rainy days and 47.3 ± 19.7 % on non-rainy days. There is a notable variability in the humidity at 3 pm from one locality to another. Unlike the humidity at 9 am, there are also outliers for high values; the other observations at 9 am are transposable to 3 pm. Finally, the humidity level at 3 pm is generally lower than that recorded at 9 am.
Missing data: 4507 (3.1 %).

Data visualization



Data for Rainfall is shown up to 50 mm, the outliers are up to 371.0 mm. Evaporation data is shown up to 25 mm, outliers up to 145.0 mm.

2.4 Sunshine and clouds

2.4.1 Sunshine

The number of hours of bright sunshine in the day. The median sunshine is 5.4 hours on rainy days and 9.3 hours on non-rainy days. Regardless of the weather, sunshine duration is highest between December and February (10.3 hours) and lowest between May and July (6.5 hours).
Missing data: 69835 (48.0 %).

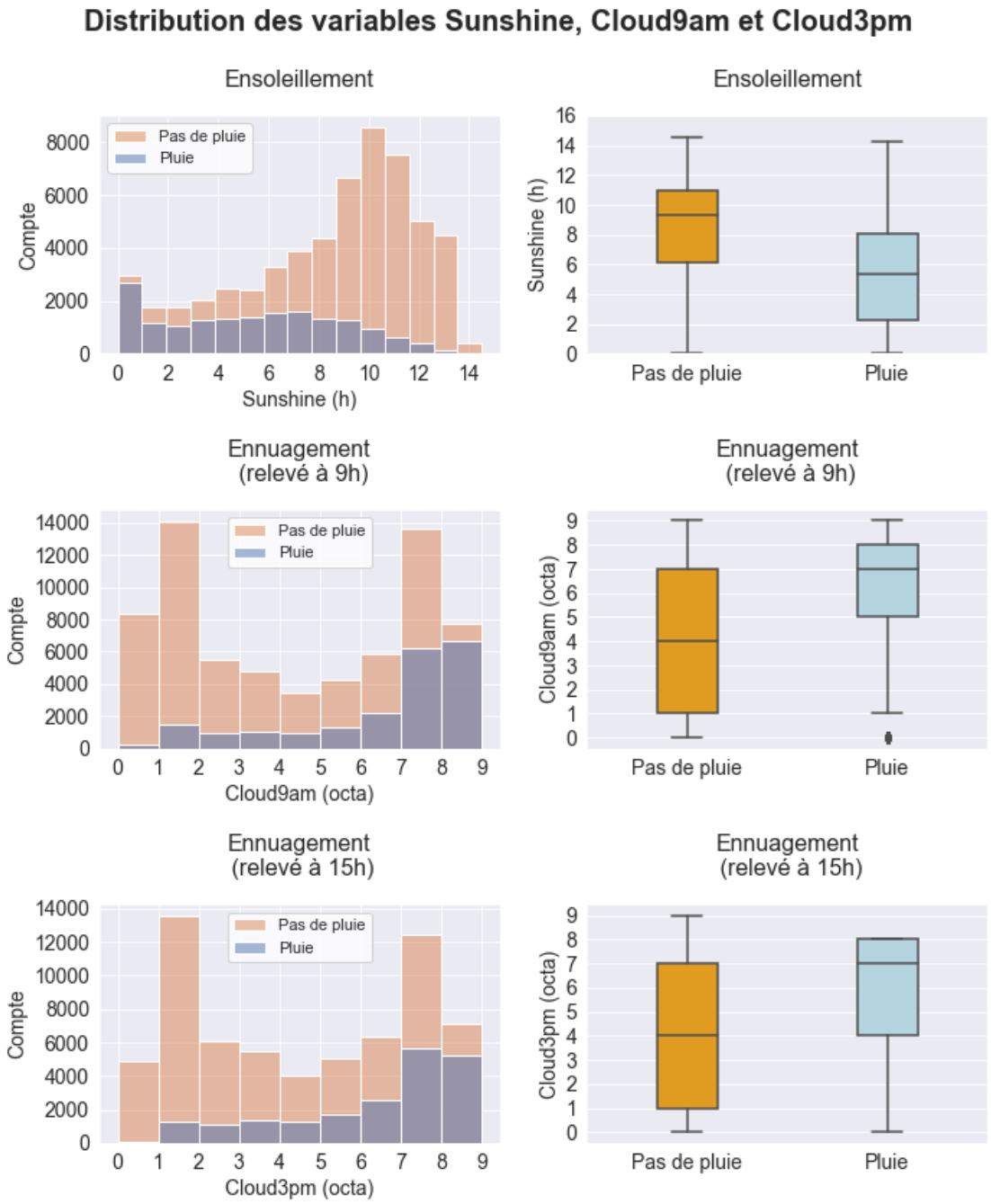
2.4.2 Cloud9am

Fraction of sky obscured by cloud at 9 am, measured in "oktas", which are a unit of eights. The median cloudiness at 9 am is 7.0 oktas on rainy days and 4.0 oktas on non-rainy days. For clarification, concerning the measurement in oktas: a perfectly clear sky is indicated by the value of 0 oktas, while a completely overcast sky is estimated at 8 oktas. The special value of 9 oktas is used when the sky is not observable due to an obstruction to visibility (e.g. fog). The variability is much higher on days without rain. The cloudiness at 9 am is very inconsistent from one location to another, but with few outliers.
Missing data: 55888 (38.4 %).

2.4.3 Cloud3pm

Fraction of sky obscured by cloud at 3 pm, measured in "oktas". The median cloudiness at 3 pm is 7.0 on rainy days and 4.0 on dry days. Despite the lack of correlation with morning cloudiness, *Cloud3pm* behaves similarly to *Cloud9am*.
Missing data: 59358 (40.8 %).

Data visualization



Data for Rainfall is shown up to 50 mm, the outliers are up to 371.0 mm. Evaporation data is shown up to 25 mm, outliers up to 145.0 mm.

2.5 Wind direction

2.5.1 WindGustDir

The direction of the strongest wind gust in the 24 hours to midnight. This variable has 16 modalities (North, North-North-East, North-East, East-North-East, East...). Within the same state, the prevailing winds are never largely predominant: there is a significant variability of directions whether it is raining or not.
Missing data: 10326 (7.1 %).

2.5.2 WindDir9am and WindDir3pm

Direction of the wind at 9 am and at 3 pm, respectively. These two variables have the same 16 modalities as *WindGustDir*. While it is very common for the wind direction to be different between these two times, it is quite rare for the wind to change direction completely. For example, for a wind coming from the South at 9 am, the main directions at 3 pm are South, South-South-East, South-East, South-South-West and South-West.
Missing data: 10566 (7.3 %) for *WindDir9am* and 4228 (2.9 %) for *WindDir3pm*.

2.6 Wind speed

2.6.1 WindGustSpeed

The speed (km/h) of the strongest wind gust in the 24 hours to midnight. The readings range from 6.0 to 135.0 km/h, with an average speed of 40.0 km/h. There are many outliers in the high values, but their impact is limited (median speed: 39.0 km/h). Interestingly, the data suggest that the wind blows stronger on rainy days: 44.0 km/h on average against 38.9 km/h on non-rainy days. Finally, the variability seems to be higher on rainy days.
Missing data: 10263 (7.1 %).

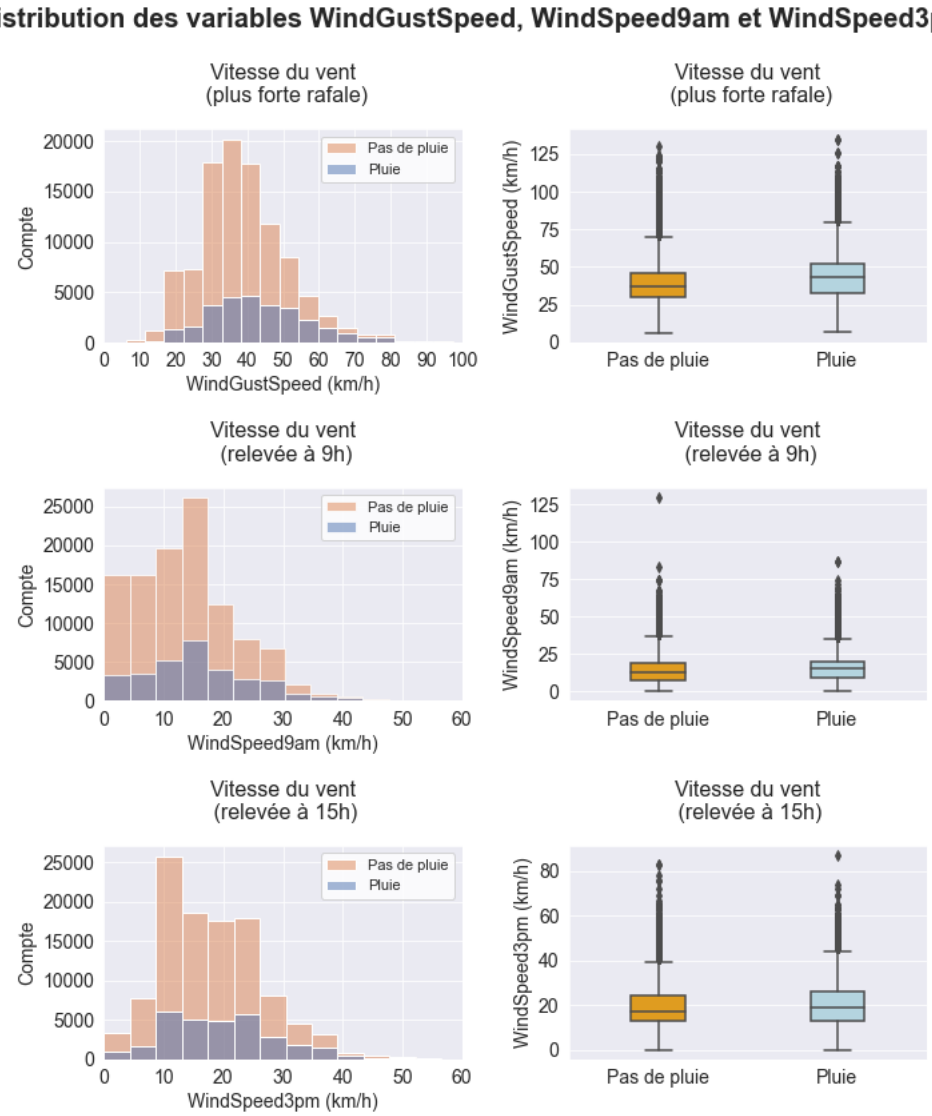
2.6.2 WindSpeed9am

Wind speed (km/hr) averaged over 10 minutes prior to 9 am. The mean wind speed at 9 am is 15.7 ± 9.5 km/h on rainy days, and 13.5 ± 8.7 km/h on dry days. The data at 9 am seem to be roughly correlated with those measured at 3 pm, as well as with the speed record of the strongest gust, but with a significant variability.
Missing data: 1767 (1.2 %).

2.6.3 WindSpeed3pm

Wind speed (km/hr) averaged over 10 minutes prior to 3 pm. The average wind speed at 15h is 20.0 ± 9.4 km/h on rainy days, and 18.3 ± 8.6 km/h on dry days. The data at 15:00 correlate roughly with the wind speed of the strongest gust, but with significant variability.
Missing data: 3062 (2.1 %).

Data visualization



Data for WindGustSpeed are shown up to 100 km/h, the outliers are up to 135.0 km/h. The data for WindSpeed9am and WindSpeed3pm are shown up to 60 km/h, the outliers are up to 130.0 km/h (9am) and up to 87.0 km/h (3pm)..

2.7 Atmospheric pressure

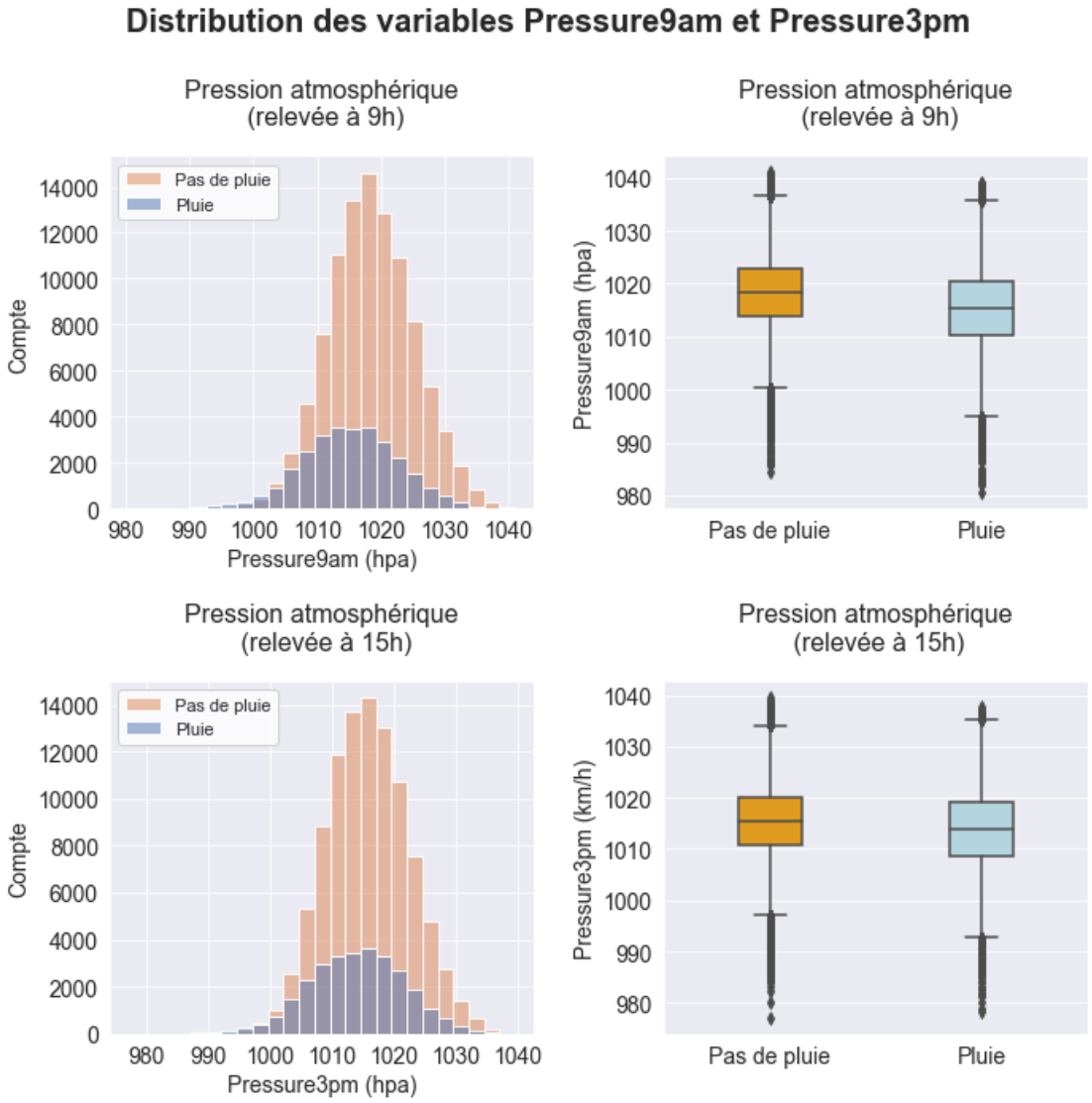
2.7.1 Pressure9am

Atmospheric pressure (hpa) reduced to mean sea level at 9 am. The mean atmospheric pressure at 9 am is higher on rainy days (1015.1 ± 7.6 hpa) than on non-rainy days (1018.4 ± 6.8 hpa). There are many outliers, for both high and low values. This variable is remarkably constant from one location to another, with very little variation in amplitude. There are four notable exceptions: Darwin, Katherine, Cairns and Townsville which have lower pressures (all located in northern Australia).
Missing data: 15065 (10.4 %).

2.7.2 Pressure3pm

Atmospheric pressure (hpa) reduced to mean sea level at 3 pm. The average pressure at 3 pm is lower on rainy days (1013.9 ± 7.6 hpa) than on non-rainy days (1015.7 ± 6.8 hpa). The air pressure distribution is very similar at 3 pm and at 9 am. However, at 3 pm, the difference in pressure between the days with and without rain is less marked than at 9 am. There are also many outliers at 3 pm, for both high and low values.
Missing data: 15028 (10.3 %).

Data visualization



2.8 Rain status

2.8.1 RainToday

Boolean: 1 if precipitation (mm) in the 24 hours to 9 am exceeds 1 mm, otherwise 0. Of the total data, there are 21.9 % rainy days and 75.8 % non-rainy days.
Missing data: 3261 (2.2 %).

2.8.2 RainTomorrow

Boolean: 1 if precipitation (mm) in the following day exceeds 1 mm, otherwise 0 . This is the variable to be predicted.
Missing data: 3267 (2.2 %).

3 Machine learning modelling

The objective of this part of the work was to test different machine learning models to define which one would be the most adapted to our problem, and then to apply this model to our data.

3.1 Model selection

3.1.1 Primary data cleaning

In order to evaluate which model fitted best, we opted for a simple data cleaning which focused on keeping all variables. Rows with missing data for *RainToday* and *RainTomorrow* were removed. For the variables *MinTemp* and *MaxTemp*, the missing values were replaced by the averages per month, year and station. Then all rows with missing values were deleted. The values of the variables *RainToday* and *RainTomorrow* were discretised (Yes/No: 1/0) and all other qualitative values were encoded. This quick data cleaning leaves 56,441 rows in the database, however data from only 26 of the 49 stations were retained.

3.1.2 Model testing: unbalanced data

The following models were then tested: K-Nearest Neighbours (KNN), Decision Tree, Random Forest and Bagging. the results of which are summarised in table 1 below. Recall values shows that class 1 (rainy day) is hardly detected. The KNN model, in addition to the unsatisfactory results despite the search for the best parameters, was not further developed due to the processing time during fitting (more than 15 minutes).
Two other models were tested but are not listed in 1: a logistic regression after data normalisation, with C = 1.0 as argument, but which did not detect class 1 (rainy day). Knowing that other models gave promising results, and that logistic regression did not seem to be an appropriate model for our problem (binary class predictions), we did not go further. A Support Vector Machine (SVM) was tested, but the model had an excessively long running

time. We therefore discarded this model.

Table 1: Scores of machine learning models tested on unbalanced data.

Model	Accuracy	Precision	Recall	f1 score
KNN	0.84	0.76	0.43	0.55
Decision Tree	0.80	0.55	0.55	0.55
Random Forest	0.86	0.77	0.54	0.63
Bagging	0.85	0.74	0.52	0.61

Displayed values for recall, precision and f1 score are those of class 1 (= rainy day). Values for class 0 (= non-rainy day) are not shown but were all very satisfying with very little variability between models.

3.1.3 Model testing: balanced data

In order to improve the detection of rainy days, we decided to test the models that performed best with a rebalancing of the data beforehand, using undersampling or oversampling.
According to the results shown in 2, the rebalancing of the data leads to better performances for most of the models, especially for the Random Forest which obtains a recall of 0.82 for class 1. We therefore selected this model.
For the remainder of this work, we seek to improve the performance of the Random Forest model, firstly through a more thorough pre-processing in order to keep as much data as possible.

3.2 Data preprocessing for machine learning

3.2.1 Data cleaning

After analysing the database as a whole, we decided to remove some variables, namely *Rainfall*, *WindGustSpeed*, *Cloud9am*, *Cloud3pm*, *Pressure9am*, *Pressure3pm*, *Evaporation* and *Sunshine*. The reason for this was that these

Table 2: Scores of machine learning models tested on unbalanced data.

Model	Accuracy	Precision	Recall	f1 score
Bagging	0.81	0.54	0.77	0.64
Decision Tree	0.74	0.43	0.73	0.54
Random Forest	0.81	0.54	0.82	0.65

Displayed values for recall, precision and f1 score are those of class 1 (= rainy day). Values for class 0 (= non-rainy day) are not shown but were all very satisfying with very little variability between models.

variables had too many missing values, without the possibility of relevant replacement values. Furthermore, the descriptive analysis of the variables suggested that some of these variables might be correlated with others.

Missing values were removed for the variables *RainToday* and *RainTomorrow*, as the latter was our target variable and dependent on the former, we could not replace missing values for these variables. Missing values were also removed for the categorical variables (*WindGustDir*, *WindDir9am*, *WindDir3pm*) as we considered the proportion of null values negligible so we did not apply replacement.

For the remaining quantitative variables, concerning temperature and humidity, the missing values have been replaced by the average values per station, per month and per year. For example, if the Albury station has missing values for the variable *MaxTemp* for the month of June 2015, these will have been replaced by the average value of *MaxTemp* for that month (June) and year (2015). Following this correction, we found that there were remaining missing values for the variables *Humidity9am*, *Humidity3pm* and *Temp3pm*. We concluded that for these variables, some stations must have had zero values for a whole month, making the correction impossible (average over the current month). These last missing values were removed.

Once all these corrections have been applied, 122,649 entries remain, so 84.3 % of the data has been retained for future use. Two stations (Newcastle and Albany) were also removed indirectly via the removal of missing values, however these stations are located in areas that include many other weather

stations.

3.2.2 Data encoding

The values *'Yes'* (rainy day) and *'No'* (non-rainy day) of the variables *RainToday* and *RainTomorrow* are replaced by the integers 1 and 0 respectively. The variables concerning the wind direction are encoded via the *LabelEncoder* method for the sake of practicality for the analyses that will follow. Indeed, 16 different directions are listed and the application of the *get_dummies* method would lead to the creation of 48 new columns. However, this method was used and compared to ensure that the *LabelEncoder* had no impact on the final results compared to dummification.

3.2.3 Standardization

We have chosen not to apply any normalisation or standardisation. Indeed, the table described below, which follows the previous corrections and deletions of variables, shows a relatively similar order of magnitude between the remaining variables. Normalization/standardization steps were nevertheless applied in other versions in order to check the impact of this step on the results, which turned out to be zero. Furthermore, the Random Forest model does not require any prior normalisation of the data. We therefore considered this preprocessing step unnecessary in our case.

3.2.4 Train and test datasets

The target variable, *RainTomorrow*, was isolated in a dataframe called *target* and the explanatory variables were stored in a dataframe called *data*. The data was then separated into a training set and a test set, the latter representing 20 % of the data.

3.2.5 Data balancing

As experienced in the preliminary models, we noticed that a rebalancing of the data was necessary to be able to detect class 1, "rainy day". Thus, the *UnderSampling* and *OverSampling* methods were used, combined and compared: an *OverSampling* (with the argument sampling_strategy = 0.6 to apply the method on 60 % of the data) was applied on the training set before applying an *UnderSampling*. We also tested *Undersampling* alone and *Oversampling* alone to compare results.

3.3 Use of random forest for rainy day detection

3.3.1 Model performances

The Random Forest model applied on data re-equilibrated by *OverSampling* then *Undersampling* shows an acceptable f1 score (0.62), the class "rainy day" is correctly detected by the model. The accuracy is 83 %, our model seems reliable for the correct detection of the classes (true negatives and true positives). This model applied on data rebalanced by *UnderSampling* shows a lower accuracy and precision, but a higher recall (0.76). This means that the class "rainy day" is well detected, but also includes observations of other classes. Finally, this model applied on data re-equilibrated by *OverSampling* shows a higher accuracy (84 %) and precision than the two others The class is not well detected. Results are displayed in table 3.

Since our first objective was to correctly and reliably detect rainy days (class 1), which are rarer than non-rainy days in Australia, we prefer the combined rebalancing (*OverSampling* then *UnderSampling*) which gives satisfactory results on the detection of the rainy day class (recall) and the accuracy of this detection. This is the model we use for the following analyses.

Note that we tried to combine a *RandomizedSearchCV* (finding the best grid of parameters) and a *GridSearchCV* (testing the best parameters) in order to improve the model but the results were not better. We therefore chose to focus on improving the preprocessing afterwards.

Table 3: Random forest performances.

	Accuracy	Recall	Precision	f1 score
OverSampling + UnderSampling	0.83	0.63	0.61	0.62
UnderSampling	0.78	0.76	0.51	0.61
OverSampling	0.84	0.50	0.70	0.59

Displayed values for recall, precision and f1 score are those of class 1 (= rainy day). Values for class 0 (= non-rainy day) are not shown but were all very satisfying.

3.3.2 Features importance

Feature importance of our model was analysed to see if there were any areas for improvement in preprocessing (e.g. by creating or removing variables). According to the results, the variables that seem to have the most influence on the prediction are the variables related to humidity and temperature, although no feature clearly stood out:

- *Humidity3pm*: 0.249;
- *Humidity9am*: 0.092;
- *MinTemp*: 0.088;
- *Temp3pm*: 0.086;
- *Temp9am*: 0.078.

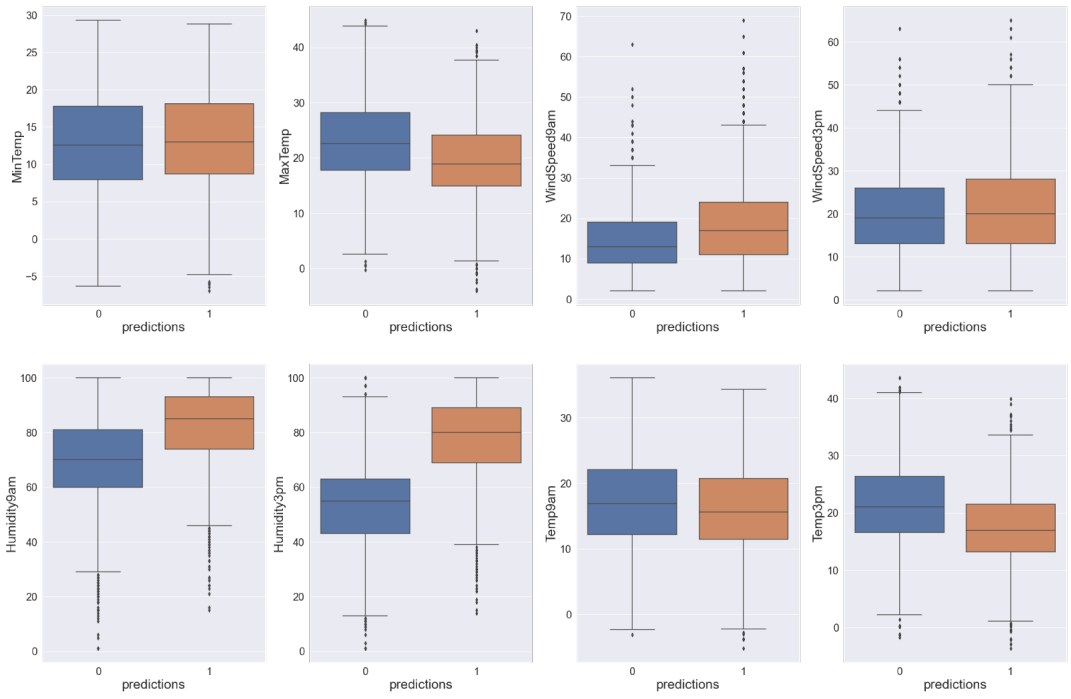
3.3.3 Analysis of correct and incorrect predictions

We analysed the distribution of quantitative variables across predictions to see if we could identify any particular patterns that might have influenced the predictions. To do this, we created four data frames corresponding to the four classes of predictions:

- True Negative (TN): prediction of days without correct rainfall;
- True Positive (TP): prediction of correct rainy days;
- False Positive (FP): prediction of a rainy day when it did not rain the next day;
- False Negative (FN): prediction of a day without rain when it rained the next day.

In particular, we have displayed graphs comparing FN and TP, in an attempt to see which patterns might have favoured the model in obtaining correct rainy day predictions, or on the contrary biased it and led to wrong predictions.

In the box plots shown below, all real rainy days are considered. The blue boxes correspond to false predictions (FN) while the orange ones are good predictions (TP) with respect to these real rainy days.



We can notice that, compared to the bad predictions, the good predictions come from lower values for the high temperature variables (*MaxTemp* and *Temp3pm*) and higher values for the humidity and wind direction variables in the morning (*WindSpeed9am*). We can also notice that these variables have many extreme values.

Concerning the wind directions, only the wind directions at 9 am could show a difference between the TP and the FN: it can be noticed that for the TP, the North-East wind stands out much more than for the FN. Perhaps further statistical analysis (comparison of averages for example) could have revealed more differences.

3.4 Trying to improve the performance of the model

3.4.1 Revision of pre-processing

Analysing the metrics obtained with the machine learning algorithms trained on the cleaned database as presented above, motivated us to rethink our pre-processing. We hypothesised that removing 8 variables could limit the performance of the model, considering that these variables could significantly influence the prediction. Given the size of the Australian territory and the diversity of climates present, we initially chose to replace the missing values station by station. However, the absence of values for entire months prevented such a replacement, so we opted secondarily for a geographical approximation. Of the 49 stations, 6 were not eligible for such a reconciliation with at least one other station. Of the 43 eligible stations, we grouped them into 15 distinct groups, following the geographical proximity.

Thus, there were initially 343,248 missing values in the database (10.3 % of the total data), spread over 145,460 entries. This second pre-processing reduced the number of missing values to 57,622: 83.2 % of the initial missing values were replaced. After deleting 4 columns with too many missing values (*Sunshine*, *WindGustDir*, *WindDir9am*, *WindDir3pm*), and deleting the remaining missing values, 129127 records remain in the database. In total, 88.8 % of the weather records were retained.

The code used for the random forest algorithm presented above was reused identically with the database from the geographical pre-processing. The algorithm was tested under the same three conditions: *OverSampling + UnderSampling*, *UnderSampling* alone, *OverSampling* alone. The results obtained with this pre-processing are quite satisfactory, nevertheless the improvement brought compared to the initial pre-processing does not seem to have a significant influence on the model performances.

Table 4: Random forest performances improvement.

	Accuracy	Precision	Recall	f1 score
OverSampling + UnderSampling	0.84	0.62	0.69	0.66
UnderSampling	0.84	0.70	0.51	0.59
OverSampling	0.86	0.72	0.57	0.64

Displayed values for recall, precision and f1 score are those of class 1 (= rainy day). Values for class 0 (= non-rainy day) are not shown but were all very satisfying.

4 Deep learning modelling

4.1 Base model

4.1.1 Model description

This first model had 2 dense layers:

- A first layer with 25 neurons and a *ReLu* activation function;
- A second layer with 50 neurons and a *ReLu* activation function;
- An output layer with 2 neurons and a *softmax* activation function;
- Learning on 6 epochs per batch of 32.

4.1.2 Model performances

This deep learning model predicted rainy days with an accuracy of 73.34 % (recall = 52.13 %) and non-rainy days with an accuracy of 87.79 % (recall = 94.78 %). The overall accuracy of the model was 85.57 %. The parameters of this first model were set arbitrarily. The performance of this model is globally satisfactory, even very satisfactory for the detection of days without rain. However, it would be interesting to improve the prediction of rainy days.

4.2 Exploring the parameters of a deep learning model

In order to better understand the influence of the different parameters of a deep learning model, we modified the model several times by modulating the parameters one by one. The modifications made in isolation were as follows:

- number of neurons: 250 in layer 1 and 500 in layer 2;

- number of epochs: 18;
- batch size: 320;
- activation function: *hyperbolic tangent*;
- addition of a layer: third layer of neurons (50 neurons, *relu* activation).

For each of these models, the performance was very similar to that of the first model, with an overall accuracy of around 86 %. The main notable difference was observed with the increase in batch size (320 instead of 32), considerably reducing the training speed of the model. Detailed results are shown in table 5.

Table 5: Deep learning models performances.

	Accuracy	Precision	Recall	f1 score
Primary model	0.86	0.73	0.52	0.61
neurons	0.86	0.76	0.50	0.60
epochs	0.86	0.73	0.53	0.62
batch	0.86	0.72	0.53	0.61
activation	0.85	0.69	0.57	0.62
layers	0.86	0.73	0.53	0.61

Displayed values for recall, precision and f1 score are those of class 1 (= rainy day). Values for class 0 (= non-rainy day) are not shown but were all very satisfying.

In the case of our dataset, parameter modulation did not improve the performance of the model. The initial model was already performing very well despite a disappointing recall for the prediction of rainy days, and our tabular data were clearly not subject to improvement.

5 Time series analysis

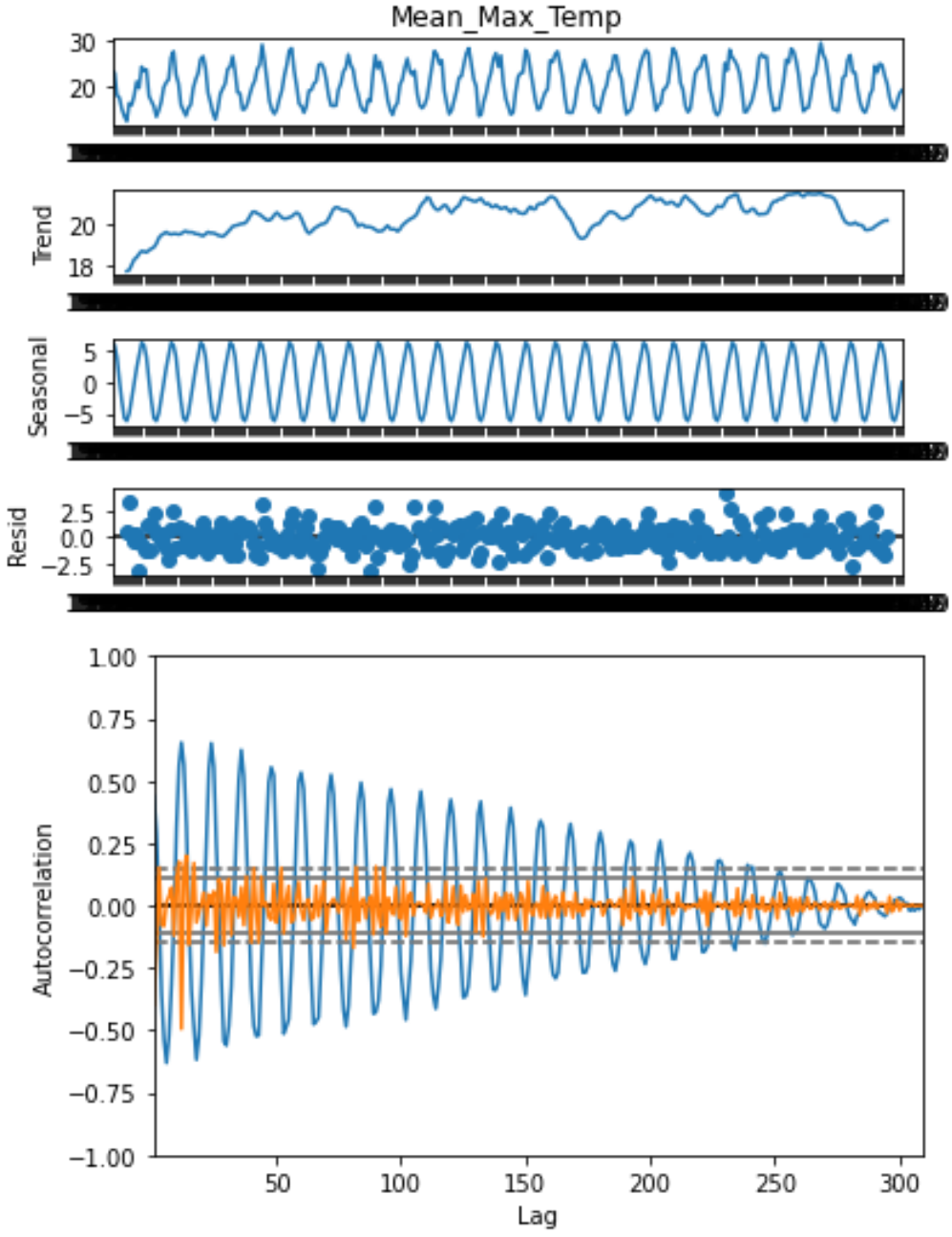
In order to explore the possibilities of studies on this kind of data, we carried out a time series study on another variable: the maximum temperature (average per month). We chose this variable because rainfall data (*Rainfall* in particular) was very often missing in the original site dataset. Temperature is also a seasonally varying variable and is therefore suitable for the application of time series.
To make the most accurate predictions possible, we have chosen to manipulate a larger time period data set found at the same original site as our first data set: this is the maximum temperatures (monthly averages) between 1995 and 2021, at a station near Melbourne.

5.1 Data preprocessing

A short data cleaning was performed to remove variables that were not relevant to our problem, such as station identification numbers or other codes. As the dataset contained some null values, these were replaced by values for the same date from other stations near Melbourne. Then, the dataset format was adapted for the Time Series application: transformation into series with the date in yyyy-mm format as index.
After a quick visualisation we confirmed the apparent seasonal behaviour of the maximum temperatures, with maximum peaks at the beginning of each year (austral summer) and minimum peaks in the middle of the year (austral winter).

5.2 Seasonal decomposition and differentiation

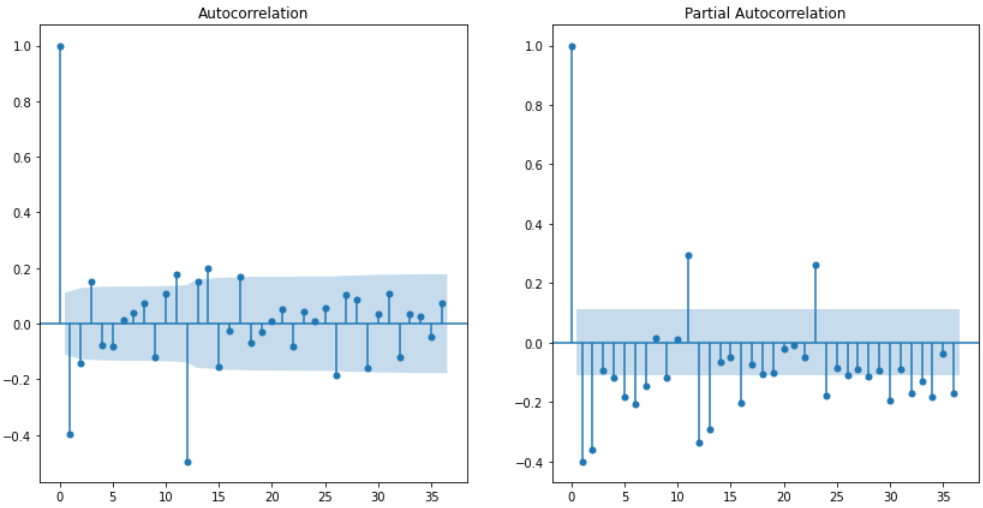
Using the *seasonal_decompose* module of the *tsa.seasonal* sub-library, we were able to see that the data follow an additive pattern (no apparent increase in amplitude over time, see figure below). To stationaryise the function, we then applied two successive differentiations, the second being a seasonal differentiation (d, D = 1, k = 12). According to the p-value obtained, well below 0.05, these two differentiations were sufficient to stationaryise the function.



Seasonal-decompose (additive model) and Stationnarisation by differentiation (p-value: $1.785e^{-13}$*)*

5.3 SARIMAX model

We decided to apply a SARIMAX model which seems to answer our problem in a complete and adapted way. We tried to estimate the orders p and q of our model using the ACF and PACF functions:

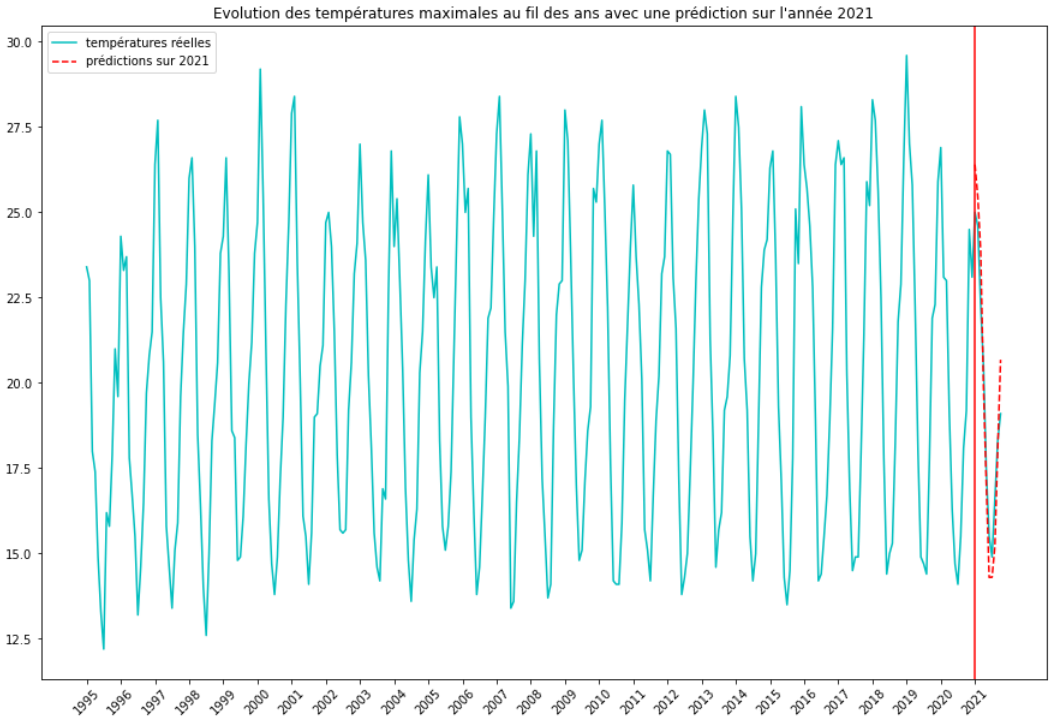


ACF and PACF curves over a period of 36 months. The ACF seems to tend towards 0 and may correspond to an AR process. The PACF does not seem to show any particular behaviour.

The orders p and q were therefore estimated more or less arbitrarily as follows: p, d, D, q, Q = 1, P = 0, k = 12.
These parameters appeared to be sufficient according to the different results of the statistical tests applied to the model: all parameters appear to be significant (p-value ¡ 0.05) so none were subsequently removed. The Ljung-Box test shows a p-value of 0.61, so we do not reject the hypothesis that the residual is white noise, which seems to follow a normal distribution according to the result of the Jarque-Bera test.

5.4 Time series prediction

The prediction was therefore applied for the year 2021 (January to October). Results are displayed in table 6.



Predictions made by the SARIMAX model are shown in red.

These first results are encouraging, the model seems to predict temperatures rather correctly, despite a difference of around 1 degree compared to the real values. With a search for the best parameters we believe that it would be possible to obtain good predictions in further term.

Table 6: Time series prediction for mean maximal temperature in 2021.

Month	Measured temperature	Predicted temperature
January	25.0	26.4
February	24.6	25.5
March	22.3	23.9
April	20.5	19.9
May	17.4	17.0
June	15.5	14.3
July	14.9	14.3
August	16.6	15.2
September	18.3	18.1
October	19.1	20.7

Actual values of the mean maximum temperature in Celsius degree for each month from January 2021 to October 2021, compared to the predicted values.

6 Discussion

The objective of this data science work was to develop machine learning models for forecasting the day's rainfall for the following day. Among the different models tested, satisfactory predictions were obtained with a Random Forest algorithm as well as with a Deep Learning model. The latter was very appropriate given the large amount of data in the database.

A first pre-processing of the data consisted in replacing the missing values by the averages over the current month for each station considered. This approach showed satisfactory results: the Random Forest algorithm was able to predict the absence or presence of rain the following day with an accuracy of 83 %. However, the detection of the presence of rain was only 63 % reliable. We therefore considered a reprocessing of the raw data: this second pre-processing applied a geographical rational: the replacement by the mean values was done by stations grouped according to their geographical location. It is indeed well known that weather conditions vary according to altitude, environment, human activity etc. This grouping was based on data from the Australian Government Bureau of Meteorology. This method of replacing missing values allowed many of the inputs and almost all of the variables to be retained. Nevertheless, the application of our Random Forest algorithm to this model did not provide a significant improvement in performance for detecting rainy days. This could be explained by the fact that the first pre-processing already provided a suitable database with a sufficient number of entries. The deleted variables did not have a major influence on the prediction.

In parallel, the development of a deep learning model showed similar results to the Random Forest: an accuracy of 86 % but a lower recall (52 %) and a comparable f1-score (61 %). On the other hand, this model was very good at detecting non-rainy days (recall: 95 %). The modulation of the model parameters (number of neurons, number of dense layers, training batch size, etc.) did not lead to improvements in the detection of rainy days. This could be explained by the fact that we were dealing with tabular data, with an already large number of inputs. On the other hand, in the descriptive study of the data, the comparative analysis according to *RainToday* revealed that the variability of many explanatory variables was often higher on rainy

days: this could explain the greater difficulty of the models to detect class 1 (rainy day) compared to class 0 (no rain). We also lacked the meteorological knowledge that would have helped us interpret our results and possibly find ways to improve the performance of our models. For example, we assumed that atmospheric pressure variations could be interesting to study in the prediction of rainfall, but we did not observe this phenomenon in our models. It would also have been very interesting to study the impact of more global phenomena on rainfall prediction. Indeed, high pressure systems such as *El Niño* are known to have a significant impact on the climate (such as rainfall), especially in Australia. In addition, weather predictions based on satellite imagery appear to be much more reliable in predicting weather and such data could have been modelled in deep learning.

The Time Series model with the prediction of maximum temperatures over the year 2021 also complemented our study. A better understanding of the parameters and their improvement could lead to better predictions in the medium term. Practical knowledge was again important for this model: one of the difficulties in assembling the data was to take into account that, even if some stations are close (around Melbourne), temperatures vary enormously depending on whether the station is in the middle of a forest or in the city.

In general, we became aware of the difficulty of managing data in a real context and taking into account business knowledge, especially with rainfall data that was regularly missing from some stations.

Therefore, in this study, we chose to focus much more on data manipulation and exploration of different models, rather than on improving the parameters of any particular model (our wish was also to manipulate different techniques for practical learning purposes).

7 Conclusion

Beyond the usefulness of these predictions, such algorithms could be used in various meteorological applications: longer-term predictions to help travellers plan the best dates for a trip, help in understanding climate change and predict temperatures over several decades, identify areas of housing at risk in order to suggest measures for adapting to changes. These ideas could also be adapted for use in the development of applications for the general public (drought risks in the coming years; tourism application, etc).