Data Scientest - Data Science Project Report

Predicting Rain in Australia

A Machine Learning and Timeseries Application

Maximilian Holdt and Carolin Poullain

July 31, 2024

Table of Contents

1. Introduction 1

2. Project structure 1

3. Understanding and manipulation of data 2

   3.1 Framework 2

   3.2 Relevance 2

4. Limitation of the dataset 4

5. Pre-processing and feature engineering 6

   5.1 Manual replacement of certain missing values and KNN imputation of remaining missing values 6

   5.2 Encoding of categorical variables 8

   5.3 Building clusters 8

6. Visualizations and Statistics 11

7. Outlook 14

8. Modeling Introduction 16

9. Model Selection 17

   9.1 Data Preparation 17

   9.2 Results 18

   9.3 Feature importance evaluation 20

   9.4 Interpretation of the top 5 important features 21

10. SHAP Evaluation 22

11. SKTIME – Timeseries Prediction of 29

   11.1 30

   11.2 AutoARIMA with exogenous variables 29

11.3 Multivariate prediction with ExpandingWindowSplitter 31

12. Conclusion 33

Report Part 1: Preprocessing

1. Introduction

The aim of this project is to develop a machine learning model that can predict the occurrence of rain on the next day in Australia from weather measurements such as humidity, cloud cover or temperature. The provided dataset contains measurements from 49 different locations across Australia during the years 2007 through 2017.

The project is developed by Carolin Poullain and Maximilian Holdt. Caro has a background in agriculture and environmental economics with focus on applied econometrics, while Max has a background in Physics which may help in understanding measurement techniques and the validity of the measured data. Otherwise there is no explicit expertise with weather data.

2. Project structure

The code related to all steps of this project is published on GitHub (https://github.com/caropoullain/weather_australia).

It contains the following files:

- weatherAUS.csv *(raw data)*
- weatherAUS_agglomeration.zip *(preprocessed data)*
- data_audit.xls
- preprocessing_kmeans.ipynb
- preprocessing_agglomeration.ipynb
- xgb.ipynb
- sktime.ipynb
- weatherAUS_report.pdf

3. Understanding and manipulation of data

3.1 Framework

For the project the dataset “Rain in Australia” is used. It is freely available via kaggle and consists of an excerpt of data collected by the Bureau of Meteorology of Australia. The corresponding full set of data is freely available on their Website http://www.bom.gov.au/climate/data/ .
The dataset contains 145,460 observations of 23 variables, 9 of which are categorical and 14 of which are numerical. For a detailed description of every variable see the Data Audit available in the GitHub repository.

3.2 Relevance

To start with, we will take ‘RainTomorrow’ as the target variable and analyze our data with respect to this variable. For this it must first be noted that weather is a complex phenomenon in which all the given variables interact in numerous ways. A simple linear dependence of rain on one such variable is therefore highly unlikely. Still, some variables seem obviously very important, like cloud cover. Although the cloud cover is measured the day before the target rainfall is measured, this variable seems to be very important. Additionally, storms with a high amount of rain are often preceded by a sudden change in humidity and pressure. Thus, these variables are also of importance. Finally, rain is often a persistent phenomenon, i.e. (depending on location) a rainy day may often be followed by another rainy day. These considerations are somewhat validated by a look at the heatmap of correlations between the numerical variables (see Fig. 1 below). Here we have also introduced the variables *Rainfall_tomorrow* and *Rainfall_yesterday* to highlight the dependence of e. g. the humidity measurement of one day and the rainfall of the following day. The heatmap shows that, indeed, the highest correlation for the rainfall the next day is given by the variables Rainfall (the day before) and *Humidity3pm* and *Cloud3pm*.
The heatmap shows that, although no one singular variable is very highly correlated with the target variable, almost all variables are somewhat correlated. This shows the intricate connection of all the different effects which must be accounted for by the model.

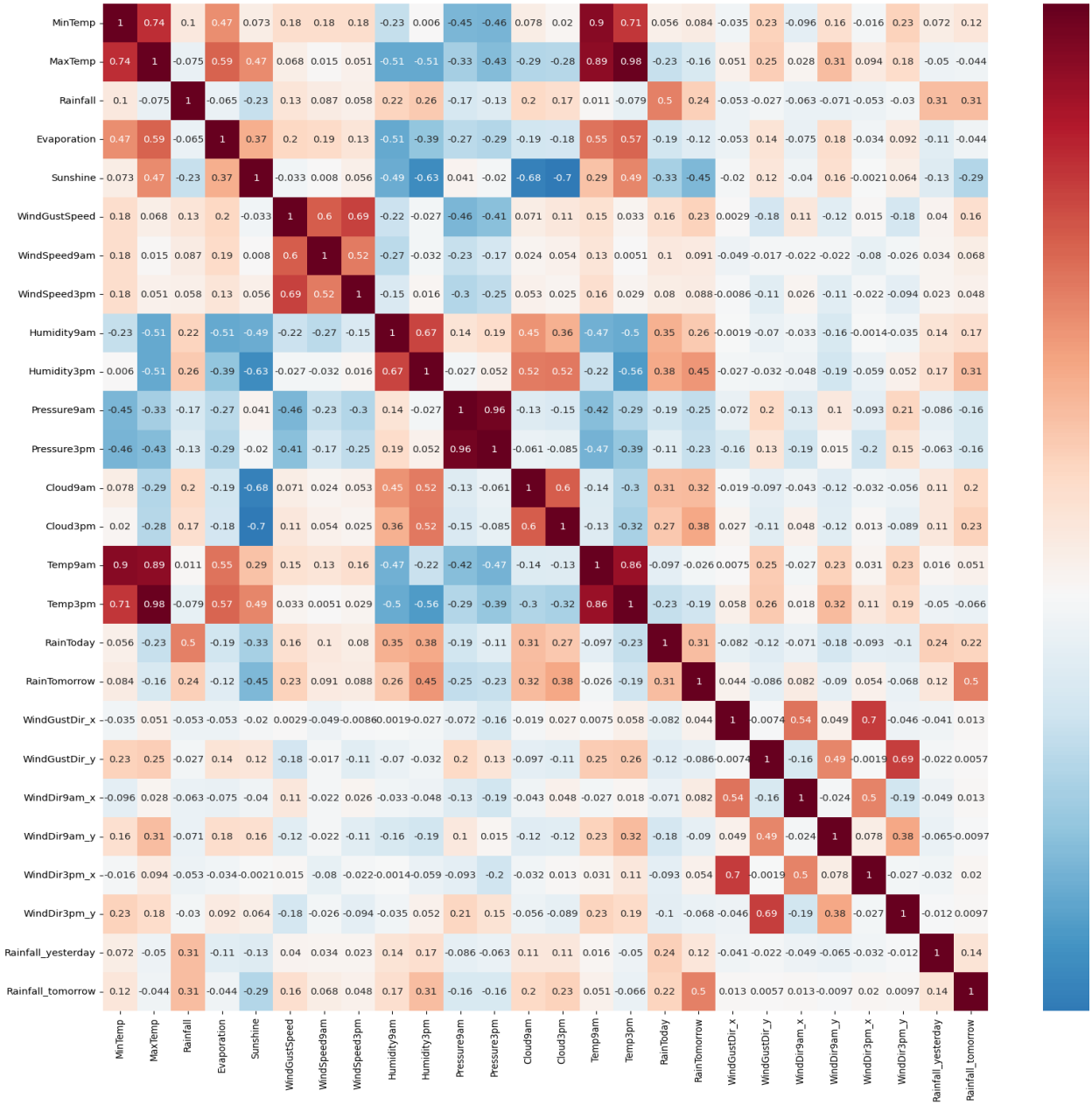


Figure 1: Correlation heatmap of all variables

Another interesting feature is the broad range of dates. The measurements are taken in very different regions with very different climate zones. This may lead to the necessity to consider

different regions on their own as some geographical features might lead to very different behavior in different places which the model might not recognize if not instructed accordingly. Finally, the data is measured over the course of ten years which might lead to impacts by other effects like global warming as well as simply the different seasons. Thus, the temporal aspect of the data must be accounted for.

4. Limitation of the dataset

A very important limitation is the number of missing values in the dataset. As seen in the table below some of the relevant variables have up to 48% missing values. Dropping all these rows would drastically alter the size of the data and is thus not recommended. Therefore, these values must be considered in detail. Here it must also be noted that the Website of the Bureau of Meteorology[1] states that "from time to time, observations will not be available, for a variety of reasons. Sometimes when the daily maximum and minimum temperatures, rainfall or evaporation are missing, the next value given has been accumulated over several days rather than the normal one day. It is very difficult for an automatic system to detect this reliably, so caution is advised." This phenomenon, hence, must be treated accordingly.

Numerical Variables	% Missing values	Categorical Variables	% Missing values
Sunshine	48.009762	Cloud3pm	40.807095
Evaporation	43.166506	Cloud9am	38.421559
Pressure9am	10.356799	WindDir9am	7.263853
Pressure3pm	10.331363	WindGustDir	7.098859
WindGustSpeed	7.055548	WindDir3pm	2.906641
Humidity3pm	3.098446	RainTomorrow	2.245978
Temp3pm	2.481094	RainToday	2.241853

[1] http://www.bom.gov.au/climate/

Rainfall	2.241853	Date	0
WindSpeed3pm	2.105046	Location	0
Humidity9am	1.824557		
WindSpeed9am	1.214767		
Temp9am	1.214767		
MinTemp	1.020899		
MaxTemp	0.866905		

Table 1: Missing values of numerical and categorical variables.

Finally, rainy days are generally not as likely as sunny days. This is shown in the distribution of the target variable which is highly biased (see Fig. 2). This also must be accounted for as otherwise a model which predicts no rain on any day would already be correct on nearly 80% of the days.

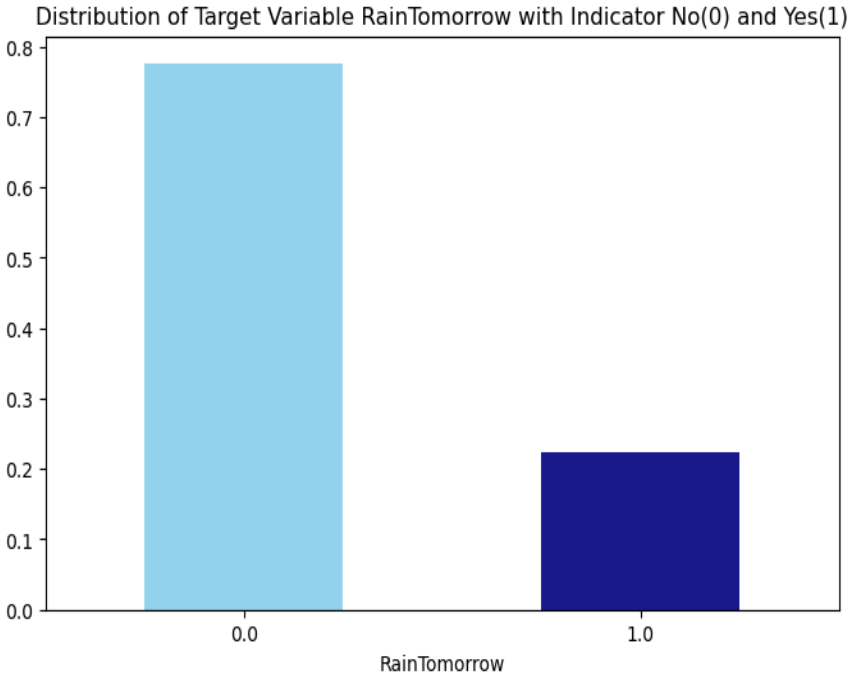


Figure 2: Distribution of the target variable.

5. Pre-processing and feature engineering

5.1 Manual replacement of certain missing values and KNN imputation of remaining missing values

Some of the variables may include outliers coming from the way the data is collected as described in the section above. To see such outliers, we looked at the corresponding boxplots (see Figs. 3, 4 & 5). The minimum and maximum temperatures seem to be reasonable and do not show obvious outliers. For *Rainfall* and *Evaporation* on the other hand there are very high values which might originate from the summation of the values of consecutive days. To confirm this, we looked at the rows containing the highest values and the measurements on the respective preceding days. Contrary to our suspicion, it turned out that for *Rainfall* these extreme values must be of natural origin as there were no missing values on the proceeding days and hence no summation has occurred.

For the variable *Evaporation*, the situation looks differently. Here the three highest values are indeed preceded by many missing values. For the values lower than the three highest, only very few values are missing before, if any at all. We thus concluded that the highest three values are summations of measurements from the days before. Accordingly, we replaced the highest values and the preceding missing values with the mean of the high values over these days. Rows containing missing values in the target variable were deleted as there is no way of filling them without impacting the validity of the model. The rest of the missing values were then replaced through a KNN imputation with n=5.

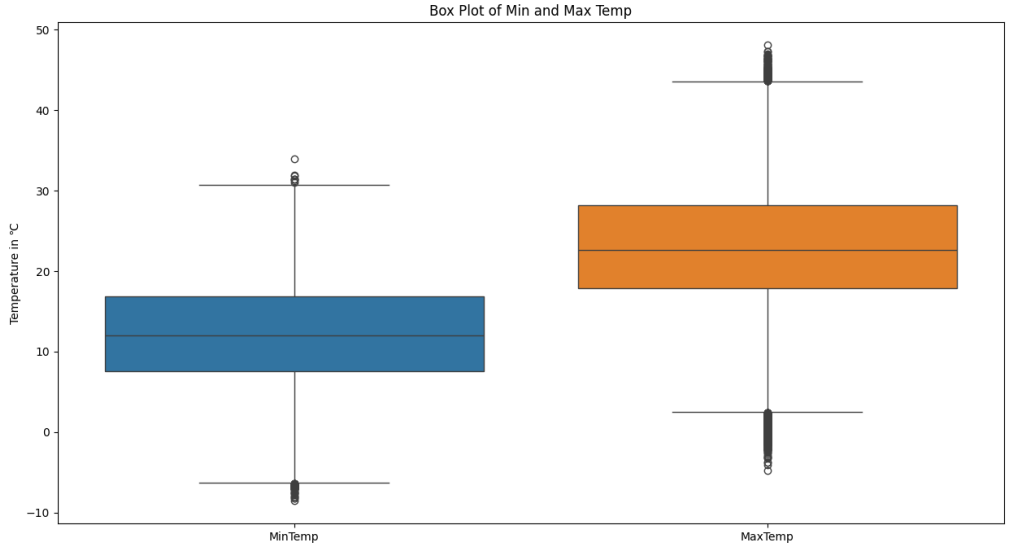


Figure 3: Boxplot of min. and max. temperatures to detect outliers

.

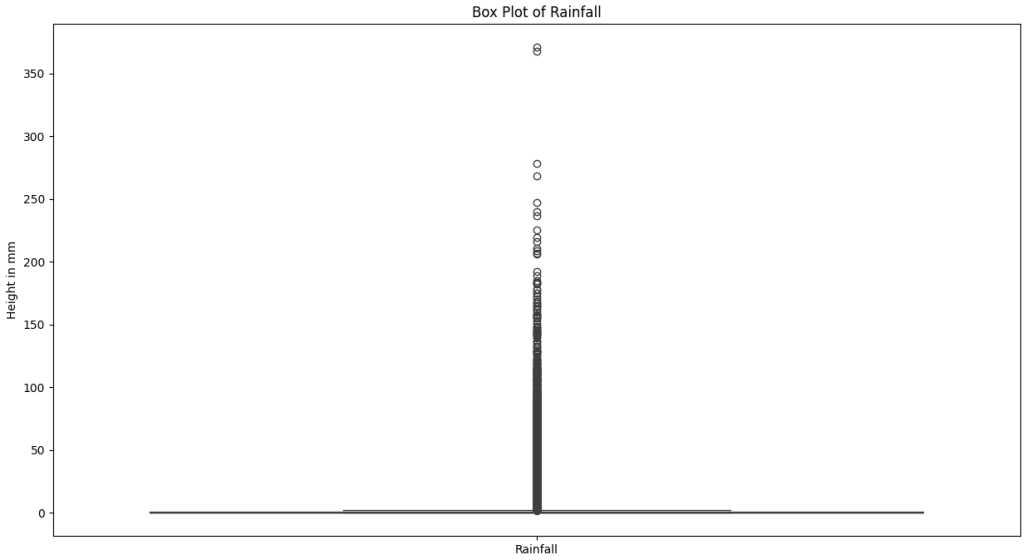


Figure 4: Boxplot of variables Rainfall for outlier detection.

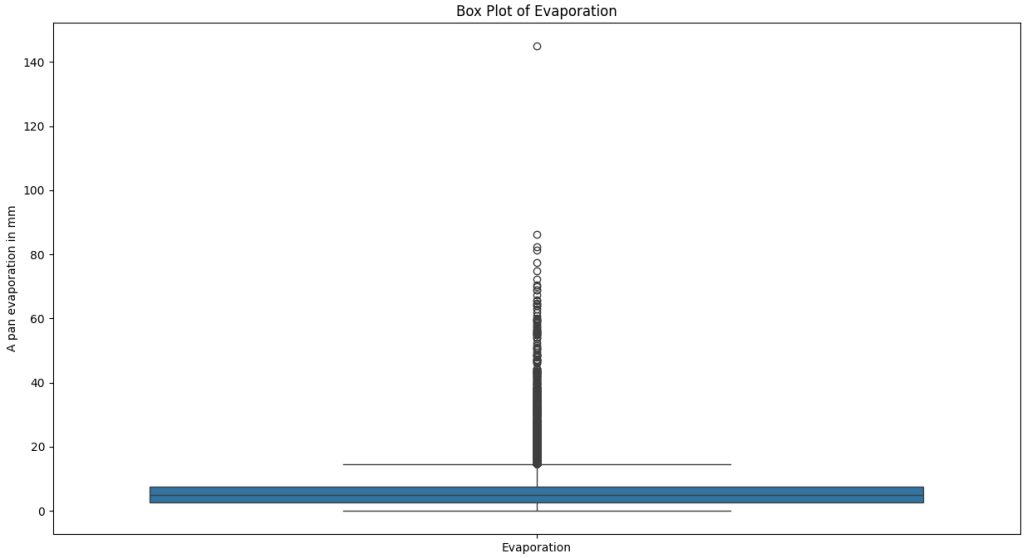


Figure 5: Boxplot of variables Evaporation for outlier detection.

5.2 Encoding of categorical variables

For our model to work we needed to replace the categorical variables (apart from *Date* and *Location*) with numerical values. For this the values 'Yes' and 'No" in the columns *RainToday* and *RainTomorrow* were replaced with 0 and 1. The wind directions were encoded in corresponding sine and cosine values: Each column with a direction was encoded in two columns representing the x- and y-value of their circular representation.

5.3 Building clusters

After imputing the missing values, we clustered the data by location to be able to improve the model in e. g. different climate zones, if needed later in the process.

An analysis via the elbow method and comparison to the climate zones of Australia for different numbers of clusters led us to choose five different clusters, see figure 6 below.

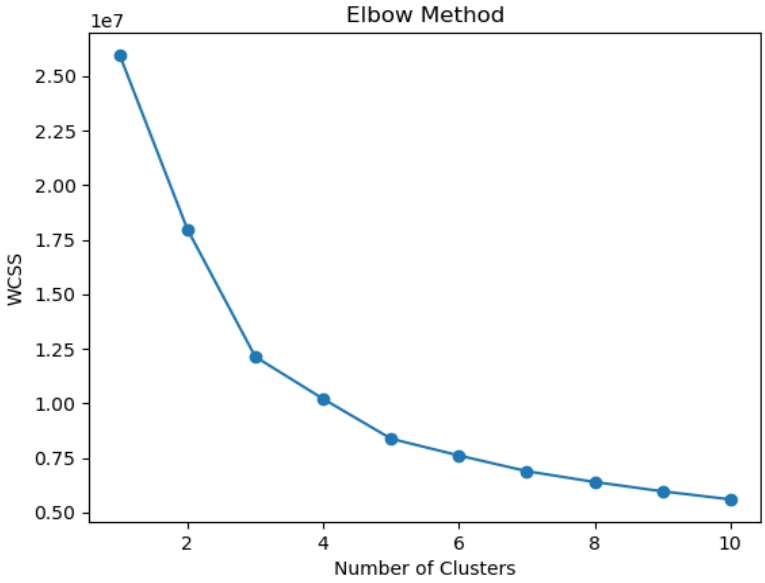


Figure 6: Elbow method for determining the optimal number of clusters.

We then tested two different ways of clustering the data, firstly, using a KMeans Clustering Algorithm and, secondly, using an Agglomerative Clustering algorithm. The resulting clusters differ (compare figures 7 and 8), and both ways do not lead to clusters which match the climate zones perfectly, although the Agglomerative clusters seem to be a slightly better fit. Still, a perfect match here should not be expected since other local aspects like mountains or valleys may have a significantly higher impact on the weather locally.

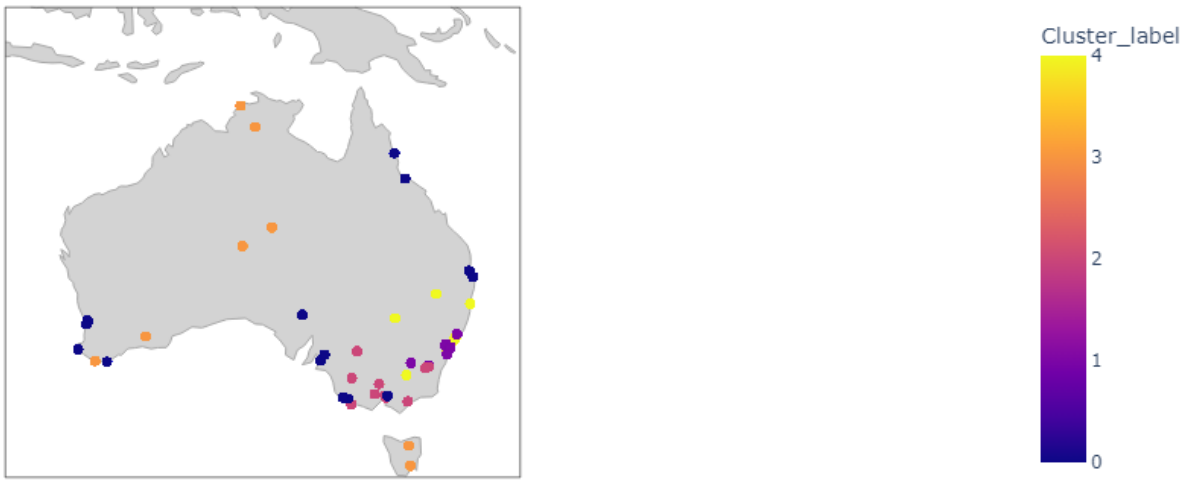


Figure 7: Resulting clusters using the Agglomerative method

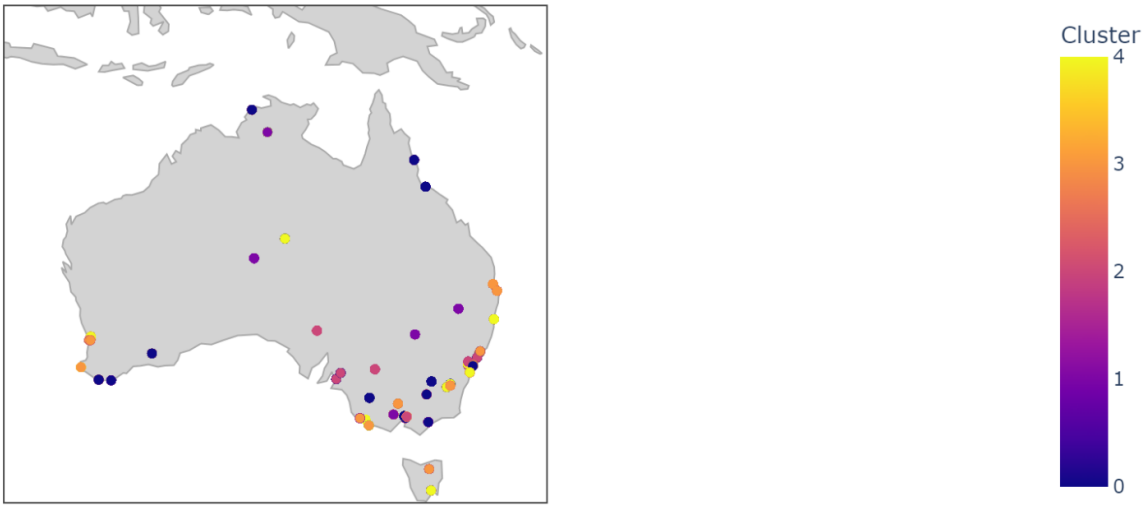


Figure 8: Resulting clusters using a simple KMeans approach

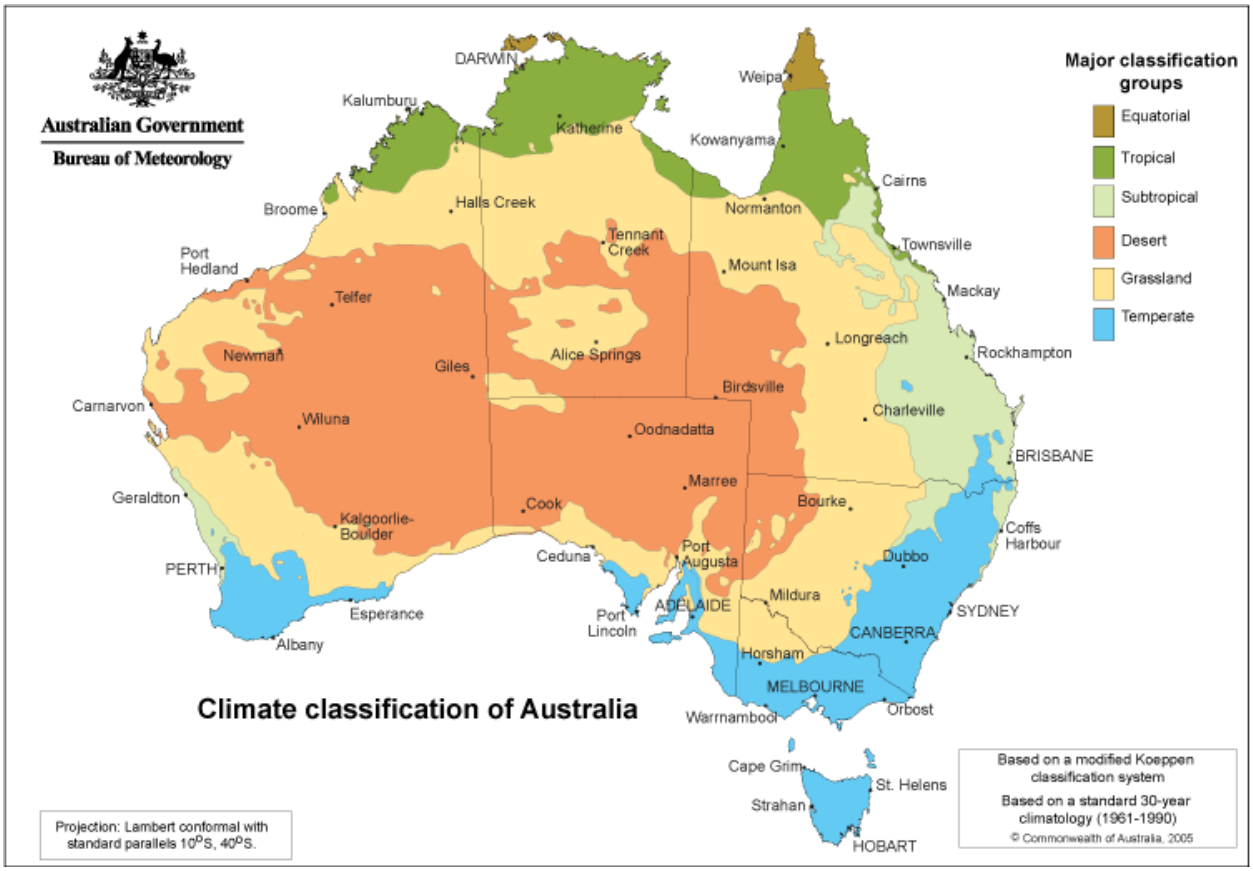


Figure 9: As reference, the climatic zones of Australia. Source: http://www.bom.gov.au/climate/

As a last step for the classification to work, it is necessary that all variables are of similar dimensions. This is not the case at first as they are represented in different physical units. To counteract this, we carried out a standardization.

Dimensional reduction was not necessary in this case and thus no such technique was considered.

6. Visualizations and Statistics

The correlation heatmap (see Fig. 1 above) already gives us a first overview of the relationship between the variables. To have a better visualization of the relationships we made a pairplot of the most important variables, see figure 10 below.

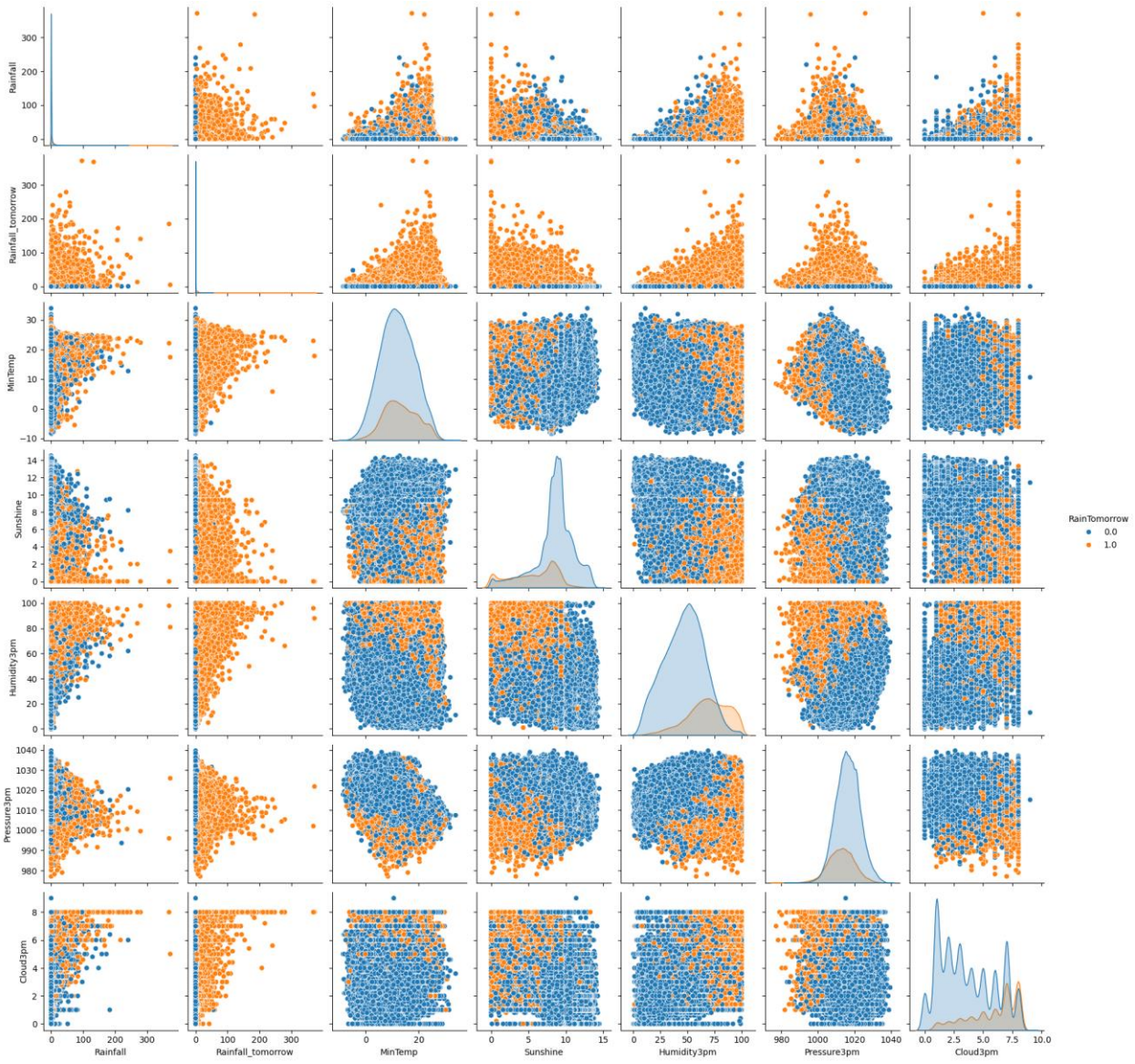


Figure 10: Pairplot of relevant variables

We conclude from the pairplot that no strong linear correlation exists between these variables, although some dependencies do exist. For example, we see that larger amounts of rainfall the next day correlate with smaller number of hours of sun and greater humidity at 3 pm. Of course, many other relations exist (e. g. between sunshine and temperatures or sunshine and cloud cover), but they are not directly important for the study of rainfall. The diagonal in the heat map is also quite interesting as it seems to show that for sunshine and humidity the functions are wider for days with rain the next day and narrower otherwise. Also, the cloud cover at 3pm has quite a different form for both possibilities. This interpretation must be carried out carefully,

though, as there are significantly more observations with no rain the next day (RainTomorrow = 0) then with rain (RainTomorrow = 1). The target variable is highly unbalanced, but likely due to natural factors. However, this matches the idea that rain follows or is accompanied by changes in weather aspects like temperature, pressure or humidity.

To see this in more detail we take a closer look at the distribution of *Humidity3pm* for the two different values of our target variable. From figure 11 below, we see that not only are the mean and median higher if it rains the next day. We also see that the probability distributions have different shapes and it is less regular if it rains the next day.

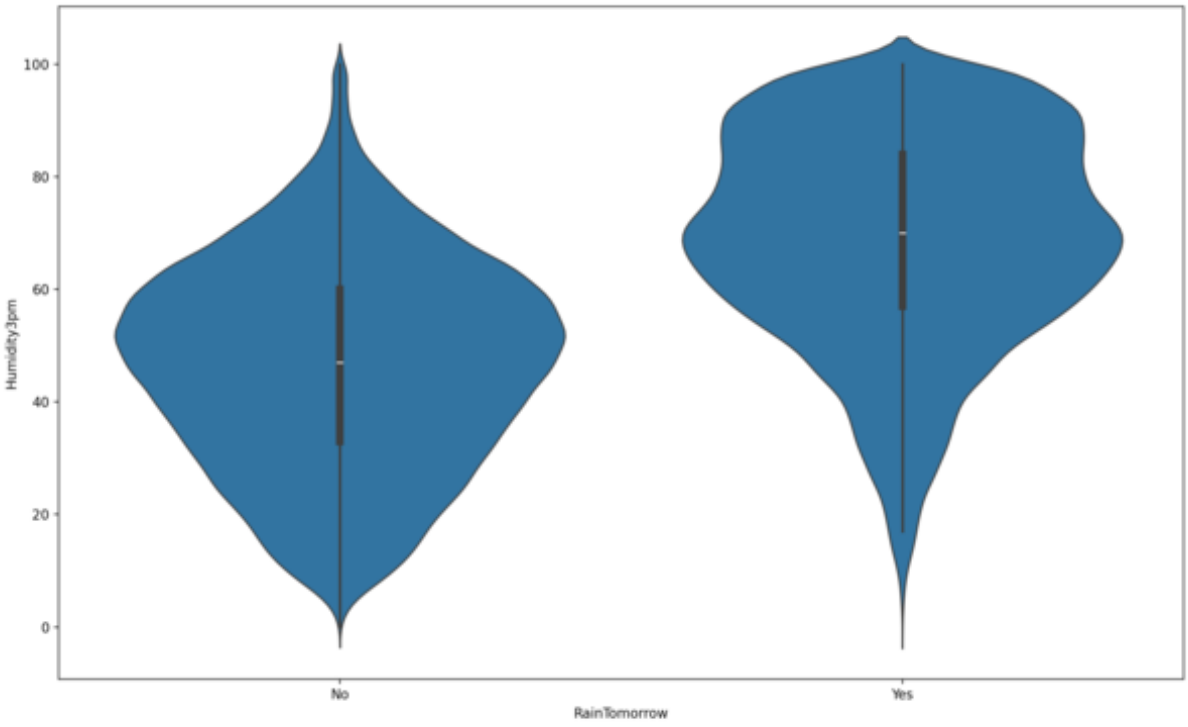


Figure 11: Violin plot of the humidity for the different values of the target variable 'RainTomorrow'

The correlation heatmap shows no strong linear correlation between the target and one explanatory variable, but some variables like humidity and rainfall are correlated. On the other hand, there are variables which do not seem to be correlated to the target at all, like the wind

direction or evaporation. But a closer study of the correlated variables shows that rain might be correlated not to the values of the explanatory variables but their changes, like sudden drops in pressure or humidity as it is often the case e.g. in storms. Drops in pressure and corresponding storms on the other hand are naturally associated with changes in wind speed and direction. Thus, these variables might still be very important as their changes might have value for the model. As all the variables are directly related to weather aspects in this way we conclude that we keep all the variables for the modeling part.

7. Outlook

As an outlook for further interesting aspects to study we also considered the evolution of the accumulated amount of rainfall over the years (see Fig. 12 below). We would expect that the global effect of climate change is somewhat present in this data, but a more detailed study would be necessary to confirm, like sub-setting by the clusters or climate zones. Still, the graph shows a persistent dependence of the amount of rain on the month with more rain in the months from December to April then in the other half of the year. This can be of importance for a better tuning of the model later, when we introduce the time aspect of the data.

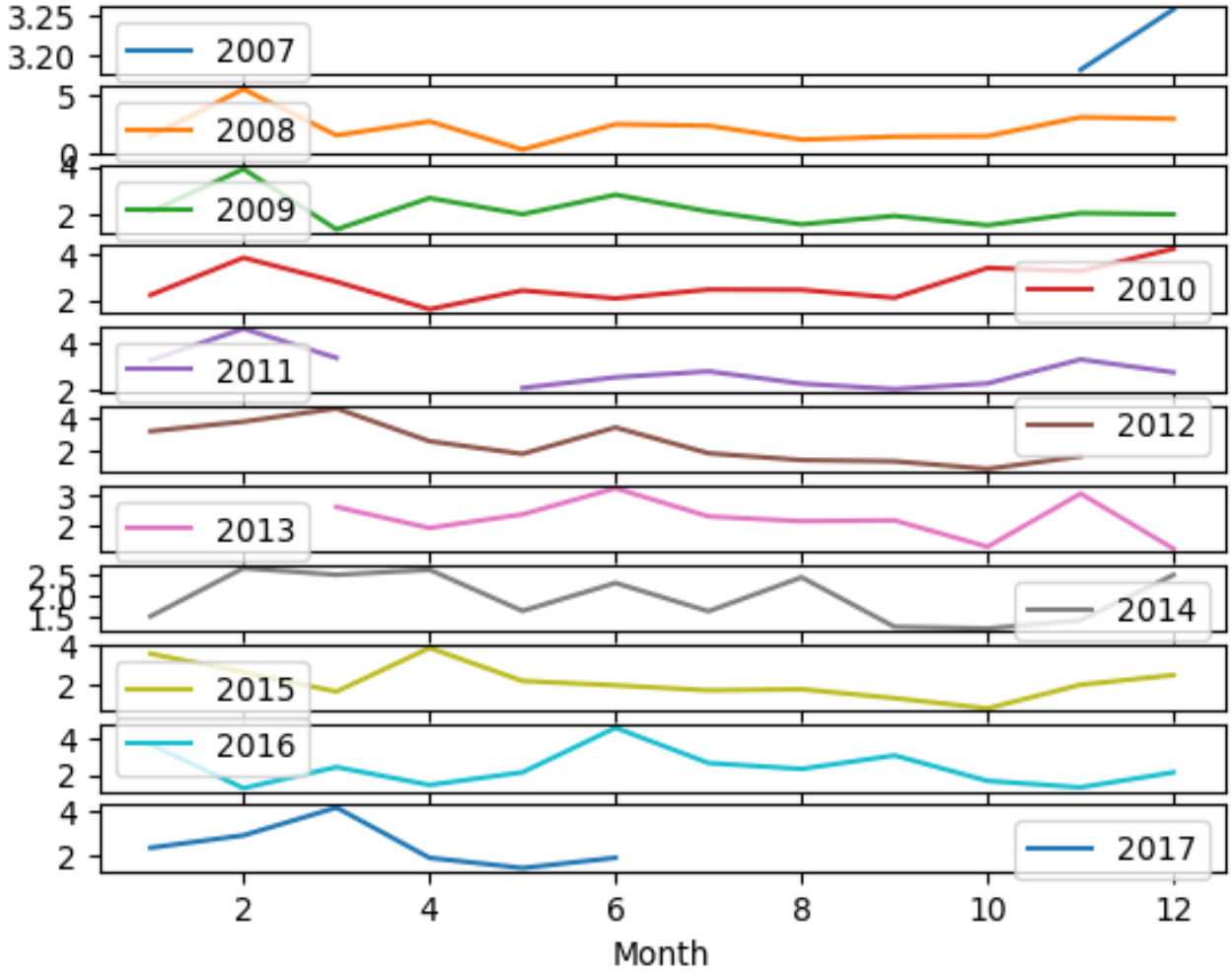


Figure 12: Monthly rainfall by year, accumulated

Report Part 2: Modeling

8. Modeling Introduction

Australia, known for its diverse and often extreme climate, experiences significant variability in rainfall patterns. Predicting rainfall is crucial for various sectors, including agriculture, water resource management, and urban planning. In this project, we delve into the complexities of predicting rainfall in Australia using a comprehensive dataset. Our primary objective is to predict whether it will rain tomorrow (the variable *RainTomorrow*) through the application of various classification machine learning models. Additionally, we aim to forecast the amount of rainfall using multiple time series forecasting techniques.

Classification Problem: Predicting *RainTomorrow*

Our first task is a classification problem focused on predicting the binary outcome of whether it will rain tomorrow or not. We developed and evaluated several classification models, including:

- Logistic Regression
- Decision Trees
- Random Forest
- SVM
- KNN
- XGBoost

To enhance the interpretability of our models, we applied a SHAP (SHapley Additive exPlanations) analysis. SHAP values provide insights into the contribution of each feature to the model's predictions, enabling us to understand the driving factors behind rainfall prediction in different regions of Australia.

Time Series Forecasting: Predicting *Rainfall*

Beyond predicting the occurrence of rain, forecasting the precise amount of rainfall is essential for effective planning and decision-making. We approached this task using several time series forecasting techniques of the library *SKTime*:

- Univariate Models (Naïve Forecaster): These models predict future values based solely on the historical values of the rainfall series.
- Multivariate Models (ExpandingWindowSplitter): These models incorporate multiple exogenous variables that may influence rainfall, such as temperature, humidity, and wind speed.
- AutoARIMA (Auto-Regressive Integrated Moving Average): Using the `sktime` library and considering all exogenous variables, we employed the AutoARIMA model, which automatically selects the best parameters for ARIMA models based on the data. By considering all exogenous variables, we aimed to improve the accuracy of our rainfall forecasts, capturing the intricate dependencies between weather variables.

This project provides a comprehensive analysis of rainfall prediction in Australia, addressing both the binary classification of rain occurrence and the quantitative forecasting of rainfall amounts. By leveraging advanced machine learning and time series forecasting techniques, we aim to deliver robust and interpretable models that can assist in better managing the impacts of rainfall variability across the continent.

9. Model Selection

9.1 Data Preparation

To fit a suitable model to our dataset, we first had to prepare the clustered data further. Since we applied different clustering techniques to our raw data during the preprocessing process, we were still working with two different datasets. Further along the modeling evaluation, it turned out that the clustered version based on the agglomerative clustering method (after taking the mean values

for every variable for each location) performed better. Hence, only the results of this dataset are presented here.

After splitting the dataset into test (20%) and training data (80%), we standardized X and y applying a StandardScaler. Fitting a first Random Forest Classifier didn't achieve the desired accuracy and a respective classification report suggested that our model performed well regarding the prediction of RainTomorrow = 0 but underperformed otherwise. This was somewhat expected as we have a naturally unbalanced dataset. Hence, we decided to apply oversampling (i.e. RandomOverSampler and SMOTE) as well as undersampling techniques (i.e. RandomUnderSampler and ClusterCentroids) to level our data.

9.2 Results

As previously mentioned, we applied different machine learning algorithms to optimize our classification model using both under- and oversampled training and test data. Since XGBoost achieved the best results by far, we will only present those in this report.

To apply an XGBoost classifier to our data, we first split off a validation set and transformed the data into matrix form. We then applied a GridSearch to obtain the ideal parameters for our model. The best results could be achieved with the following parameters:

- booster: 'gbtree'
- learning_rate: 0.1
- max_depth: 8
- objective: 'binary: logistic'
- num_boost_rounds: 700

We also set the threshold for predicting rain to 0.6 which evened out the false positive and false negative predictions. These settings then produced the following results on the test data set (which was also used for validation during the training of the model):

	precision	recall	f1-score	support

0 (no rain)	0.94	0.92	0.93	19754
1 (rain)	0.92	0.94	0.93	19924
accuracy			0.93	39678
macro avg.	0.93	0.93	0.93	39678
weigthed avg.	0.93	0.93	0.93	39678

Table 2: Classification report for the XGBoost model on the test data set

These results are very satisfactory which is also easily seen in the corresponding crosstab:

Observations\Predictions	0	1
0	18147	1151
1	1607	18773

Table 3: Confusion Matrix for the test data set of the XGBoost model

On the validation data set (which the model never saw before, and which wasn't oversampled) we obtained the following results:

	precision	recall	f1-score	support
0 (no rain)	0.91	0.92	0.91	11122
1 (rain)	0.69	0.67	0.68	3098
accuracy			0.86	14220
macro avg.	0.80	0.79	0.80	14220
weighted avg.	0.86	0.86	0.86	14220

Table 4: Classification report for The XGBoost model on the validation data set

Obviously, the scores on the validation set are not as good as on the test set. But we still obtained a high accuracy overall, and the model is very good at predicting no rain while also having high

scores for the rain prediction even though the data is very imbalanced. The crosstab for this data set shows in some more detail that the false predictions are roughly the same amount for false positive and false negative predictions.

Observations\Predictions	0	1
0	10211	1030
1	911	2068

Table 5: Confusion Matrix for the test data set of the XGBoost model

9.3 Feature importance evaluation

In our analysis of the classification models predicting *`RainTomorrow`*, we examined feature importance using the XGBoost model, focusing on the importance type "gain." This metric measures the improvement in accuracy (or reduction in error) brought by each feature to the splits it makes in the trees. By identifying and ranking these key features based on their gain values, we can better understand which variables contribute the most to enhancing the model's performance. This analysis allows us to gain deeper insights into the factors that significantly influence rainfall predictions, ensuring more informed and reliable forecasting.

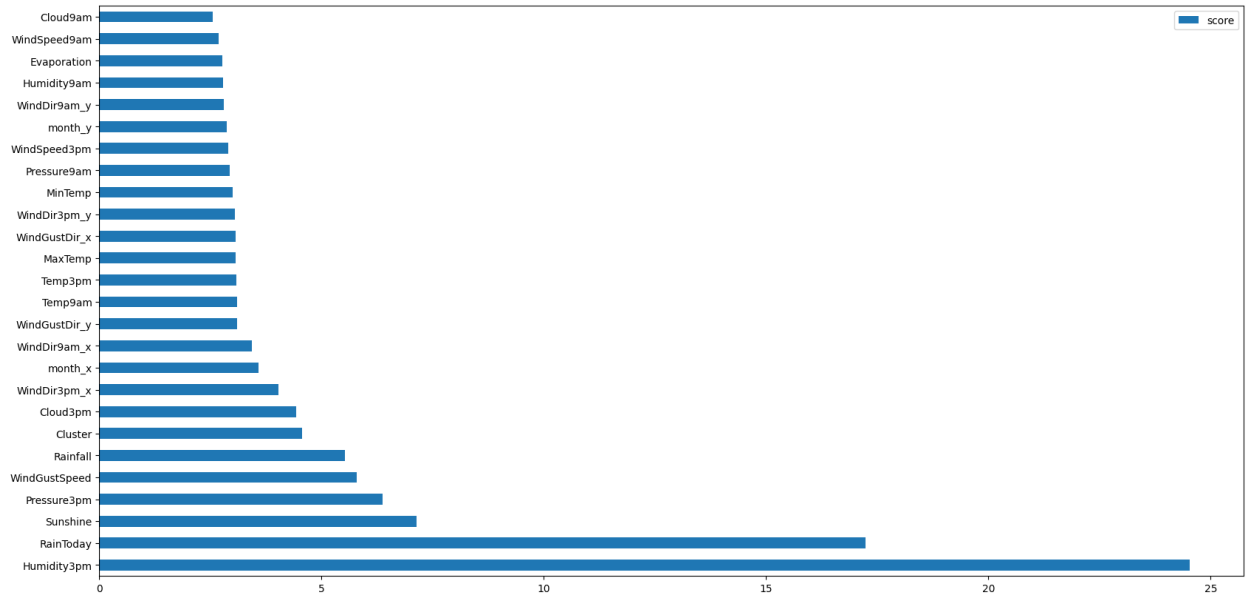


Figure 13: Feature importance of XGBoost model

9.4 Interpretation of the top 5 important features

Humidity3pm emerged as the most influential variable in our model. High humidity in the afternoon typically indicates a higher likelihood of rain the following day. This feature's substantial gain suggests that it significantly improves the model's predictive accuracy, likely due to its strong correlation with precipitation patterns. The presence of *RainToday* is a strong predictor of whether it will rain tomorrow. This variable's high importance is intuitive, as ongoing weather conditions often persist into the next day. Its contribution to the model indicates that recent rainfall is a critical factor in forecasting future rain. *Sunshine* is another key variable, representing the amount of sunlight received. More sunshine generally correlates with lower chances of rain, and vice versa. Atmospheric pressure at 3 pm (*Pressure3pm*) is a crucial predictor, as pressure systems are closely linked to weather patterns. Lower pressure often indicates stormy weather, while higher pressure suggests fair weather. This feature's importance underscores how pressure variations can affect rainfall predictions. Finally, *WindGustSpeed* represents the speed of the strongest wind gusts. High wind gusts can be associated with stormy conditions and frontal systems, which are often accompanied by rain. The inclusion of this variable in the top five highlights its relevance in capturing dynamic weather conditions that influence rain likelihood.

10. SHAP Evaluation

To enhance the interpretability of our classification models predicting *RainTomorrow* further, we conducted a SHAP (SHapley Additive exPlanations) analysis on the randomly oversampled XGBoost model. SHAP values help us understand the contribution of each feature to the model's predictions. By decomposing the output of our models into the sum of individual feature effects, SHAP provides clear insights into which variables are most influential in determining whether it will rain tomorrow. This analysis is crucial for validating our models and gaining deeper insights into the factors driving rainfall predictions across different regions in Australia.

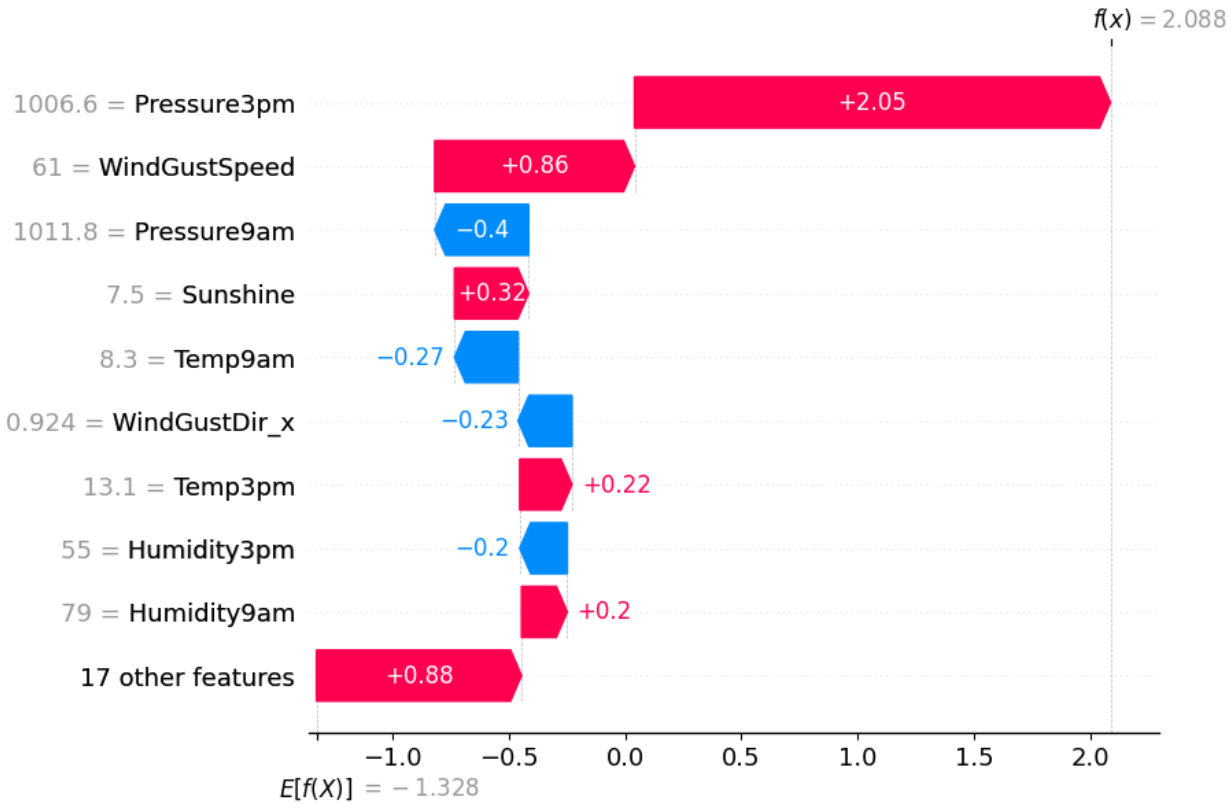


Figure 14: Waterfall plot of SHAP values for a single prediction

The top five features in the waterfall plot provide valuable insights into the factors influencing the prediction for rain for a single prediction. High pressure in the afternoon and wind gust speed have the most significant positive impact, while lower pressure in the morning and temperature

contribute negatively. *Sunshine* also plays a role in reducing the probability of rain, although its influence is relatively moderate compared to pressure and wind gust speed.

Figure 15: SHAP force plot for a single prediction

The SHAP force plot in figure 15 provides a visual explanation of how each feature contributes to a specific prediction by showing the direction and magnitude of each feature's impact. It is particularly useful for understanding individual predictions in detail. The base value (0.4999) is the average prediction of the model across all instances. Each feature's contribution is represented as a horizontal bar pushing the prediction to the right (positive contribution) or to the left (negative contribution). Features contributing to an increase in the prediction value (positive SHAP values) are shown in red, while features contributing to a decrease (negative SHAP values) are shown in blue. The prediction value equals 1, meaning that it will rain tomorrow. Positive contributions push the prediction value to the right, increasing it, while negative contributions push it to the left, decreasing it. The longer the bar, the more significant the feature's impact on the prediction.
As the feature importance analysis of the XGBoost model already suggests, *Humidity3pm* has a large positive impact on the prediction, as well as if it has rained the previous day (*RainToday*) and the quantity of rainfall on that day (*Rainfall*).

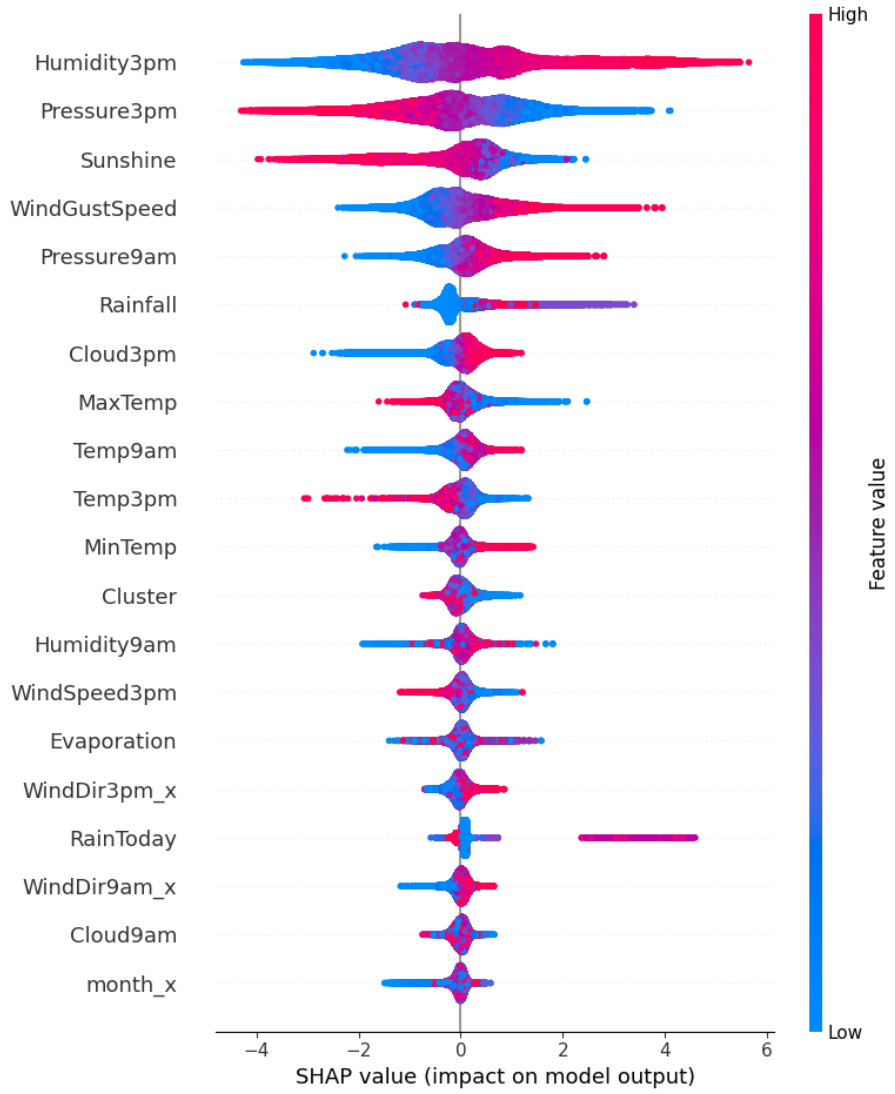


Figure 16: SHAP summary plot of the entire dataset

Interpreting the summary plot in figure 16 involves understanding how each feature contributes to individual predictions and the overall model behavior. As features higher up on the plot are more important in predicting the target variable (i.e. features with high SHAP values), *Humidity3pm, Pressure3pm* and *Sunshine* can be evaluated as the three main influential features when predicting if it rains tomorrow. Positive SHAP values indicate features that push the prediction higher, while negative SHAP values indicate features that push the prediction lower. The longer the bar or the larger the dot, the greater the impact of the feature on the prediction. This underlines one's intuition – higher values of *Humidity3pm* have a positive impact on *RainTomorrow*, as well as lower values of *Pressure3pm* and *Sunshine* of the previous day. Higher

values of *WindGustSpeed* and *Pressure9am* also have a positive impact on *RainTomorrow*, which also aligns with general expectations. High pressure in the morning and low pressure in the afternoon correspond to a fast decline in pressure, resulting in higher wind speed and an overall worsening of the weather.

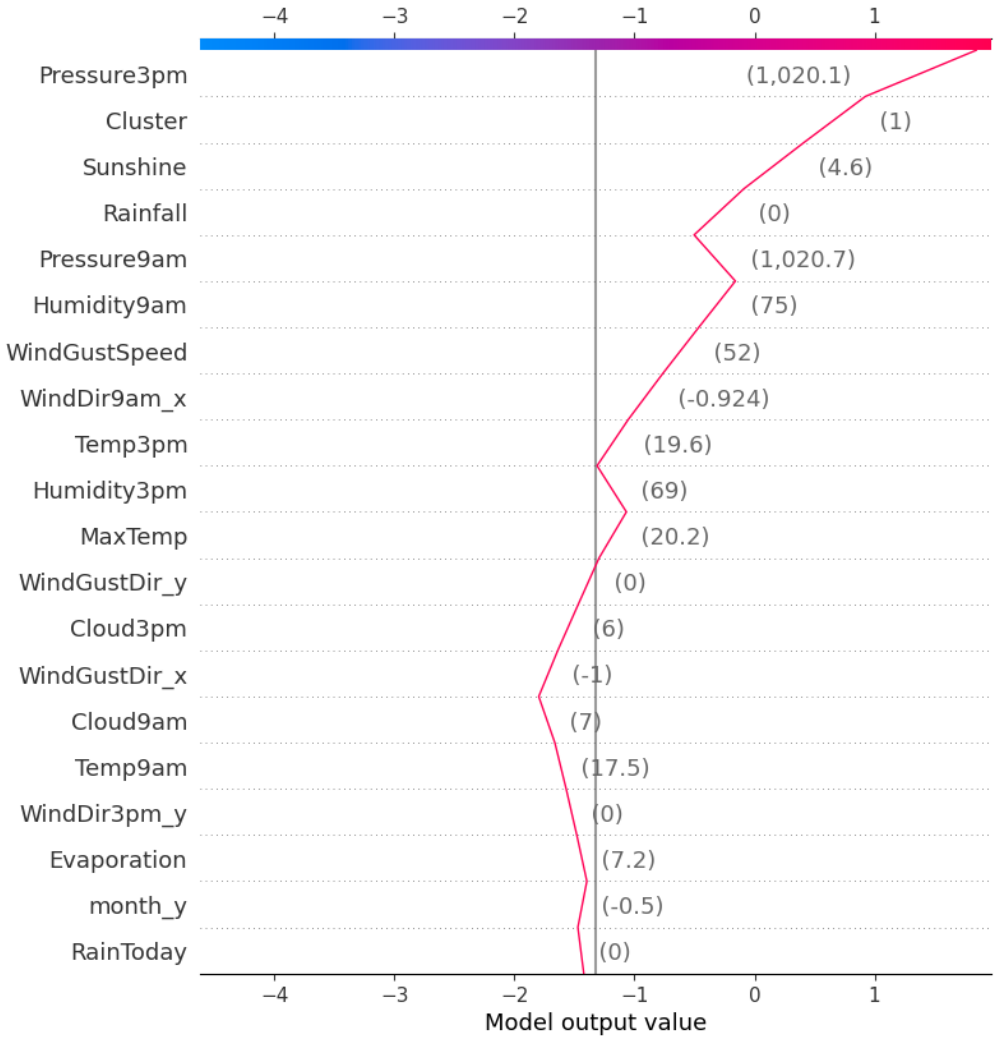


Figure 17: SHAP decision plot for a single prediction

The decision plot in figure 17 shows the cumulative effect of features along the model's decision path. The y-axis represents the decision path labeled with the features in the order they were applied by the model. The x-axis represents the model output or prediction value. It starts at the base value (the average model output if no features are considered) and ends at the final prediction for each instance (in figure 17 only the 20 most important features are shown which is

why the path doesn't start at 0). Each point along the decision path shows the impact of a particular feature on the model's prediction. The movement left or right between points indicates the increase or decrease in the prediction due to that feature and the line in the plot represents the decision path for a single instance. Thus, we can see how relevant the different features are in a single prediction.

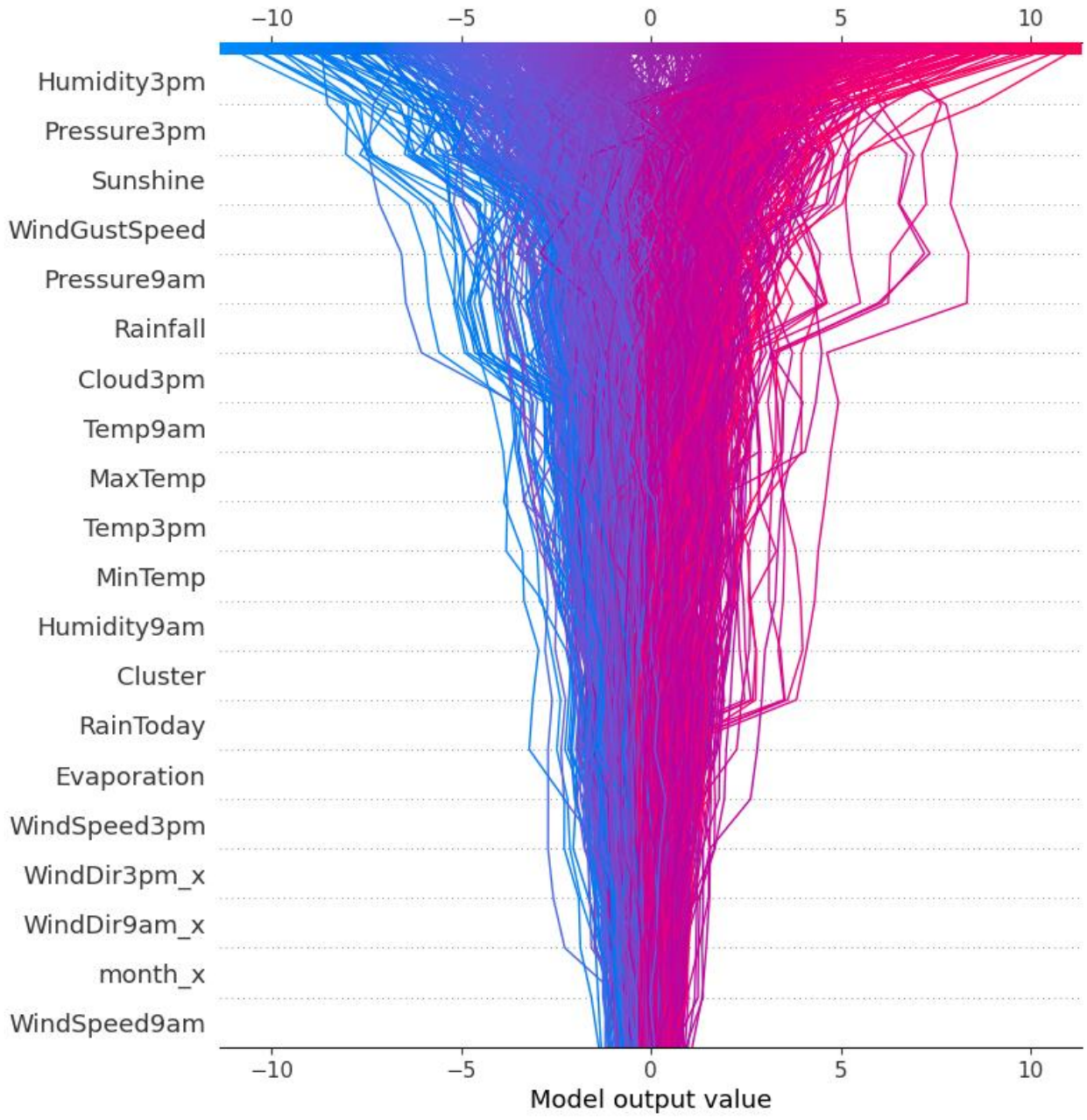


Figure 18: SHAP decision plot for 1000 predictions

It is also helpful to look at the plot for a couple of samples together as in figure 18 where the decision plot for 1000 predictions is shown. We see that the paths are mostly very similar which tells us that the relevance of the different predictions is rather homogeneous. But we also see some predictions which stray from the usual path and show a much more significant movement in features which are generally less relevant, e. g. there are paths which are highly impacted by *RainToday* and *Rainfall*. Such inconsistency may shine some light on the shortcomings of the model as they show that some features might have a different effect depending on the values of other features. Such highly non-linear effects can be difficult to adapt for by the XGBoost model.

11. SKTIME – Timeseries Prediction of *Rainfall*

In this chapter, we delve into the application of time series forecasting techniques to predict rainfall using the sktime library. Our analysis encompasses three distinct modeling approaches:

Univariate Prediction with NaiveForecaster

We started with a straightforward univariate prediction using the NaiveForecaster, which relies solely on historical rainfall data to forecast future values. This method serves as a baseline for evaluating the performance of more sophisticated models.

AutoARIMA with Exogenous Variables

Next, we employed the AutoARIMA model, incorporating all available exogenous variables such as temperature, humidity, and pressure. This approach leverages the powerful capabilities of AutoARIMA to automatically select the best-fitting ARIMA model, enhancing prediction accuracy by considering the influence of external factors.

Multivariate Prediction with ExpandingWindowSplitter

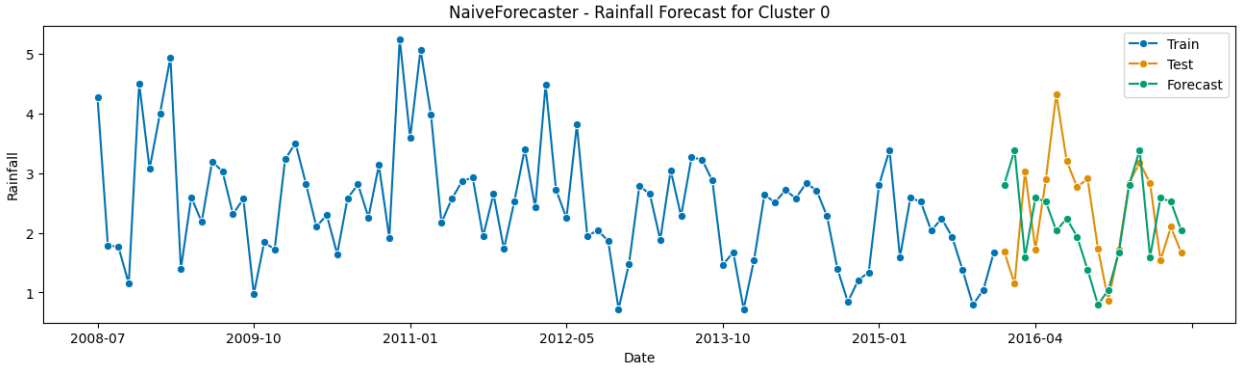
Finally, we implemented a multivariate prediction model using the ExpandingWindowSplitter, which allows for dynamic training on expanding windows of data. This method enables the model to learn from an increasing amount of historical information, potentially capturing complex temporal patterns and interactions among multiple variables.

We found that the AutoARIMA model yielded the most accurate results, outperforming both the NaiveForecaster and the multivariate model.

To prepare the data for sktime modeling, we first calculated the average rainfall per month and set the frequency to monthly. We then split the dataset into a training set (all years prior to 2016) and a test set (2016 and later). The objective is to predict the rainfall for each cluster for the years from 2016. The results are explained in detail in the following.

11.1 *Univariate prediction with NaiveForecaster*

The figures below visualize the results from our first modeling approach. The green lines represent the prediction results, the yellow lines the test data. It can easily be seen that the model does a good job in predicting the rough trend of the monthly rainfall occurrence, but the single predictions are still far off.



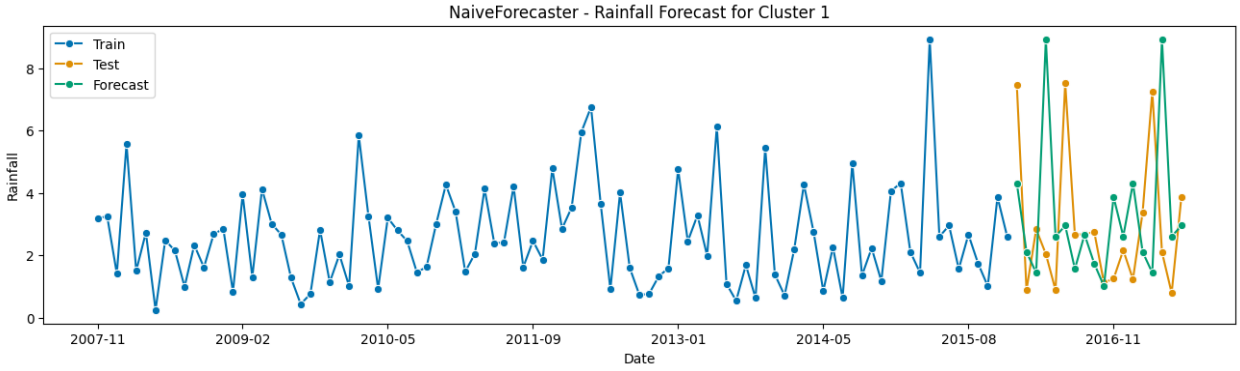
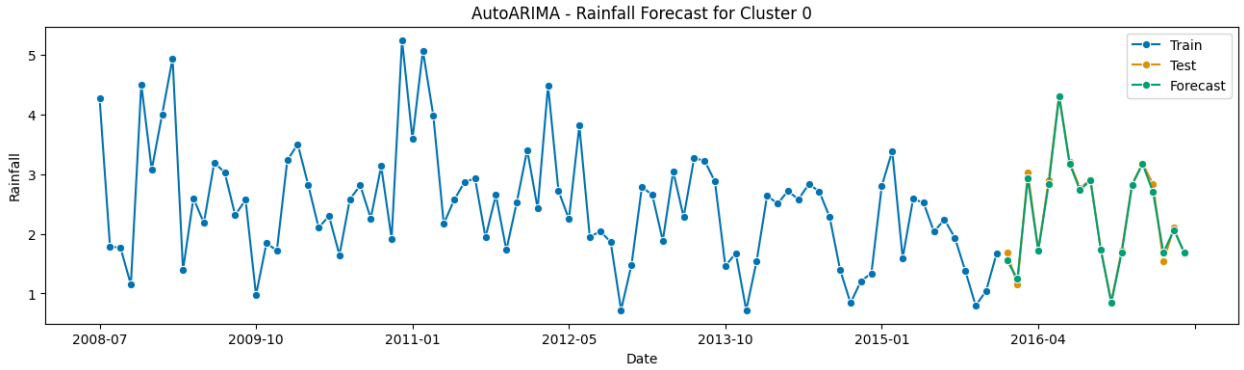
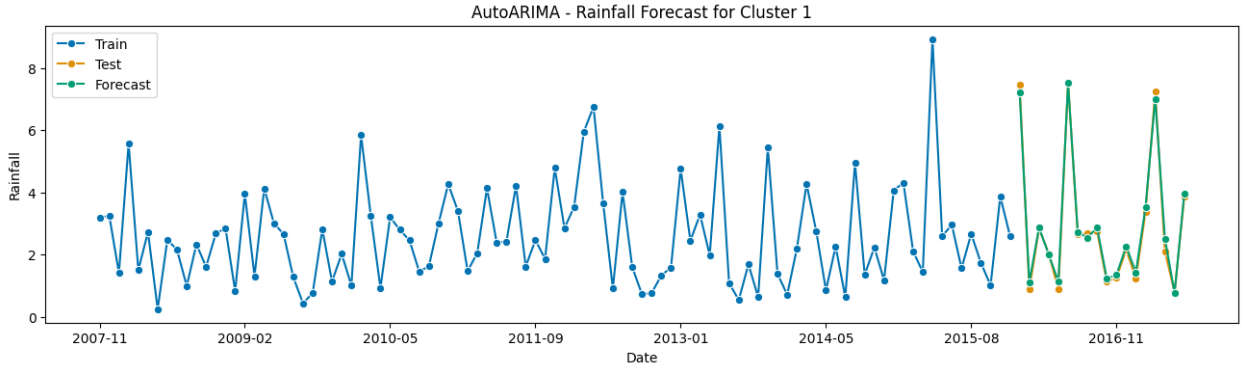


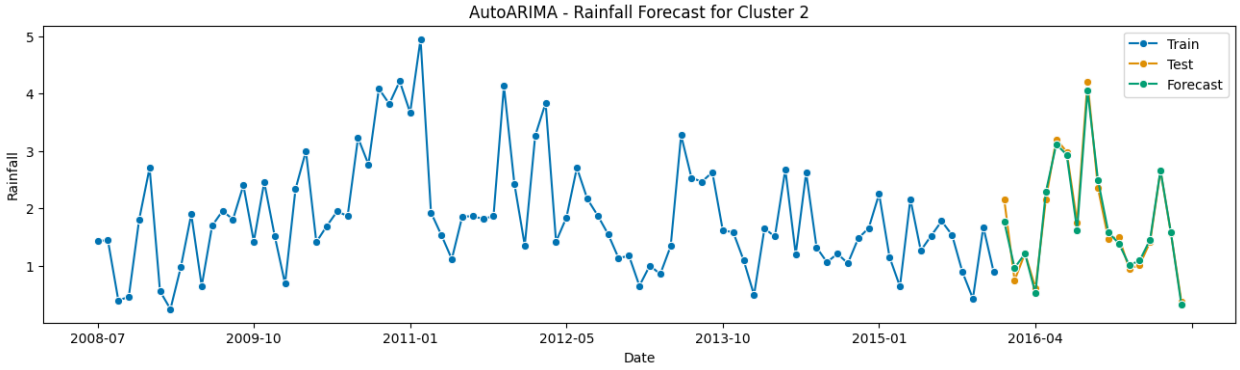
Figure 19: Univariate predictions for clusters 0-4

11.2 AutoARIMA with exogenous variables

As a next step, we looked at a sktime model that includes all exogenous variables. The figures below show the incredibly precise predictions of the model. It is noteworthy that this model uses the exogenous data also for the months where it makes predictions. So, the accuracy of the model shows that the dependencies of the mean values of each variable are captured very well by the model.


AutoARIMA - Rainfall Forecast for Cluster 0
Train
Test
Forecast
Rainfall
Date
2008-07
2009-10
2011-01
2012-05
2013-10
2015-01
2016-04


AutoARIMA - Rainfall Forecast for Cluster 1
Train
Test
Forecast
Rainfall
Date
2007-11
2009-02
2010-05
2011-09
2013-01
2014-05
2015-08
2016-11


AutoARIMA - Rainfall Forecast for Cluster 2
Train
Test
Forecast
Rainfall
Date
2008-07
2009-10
2011-01
2012-05
2013-10
2015-01
2016-04

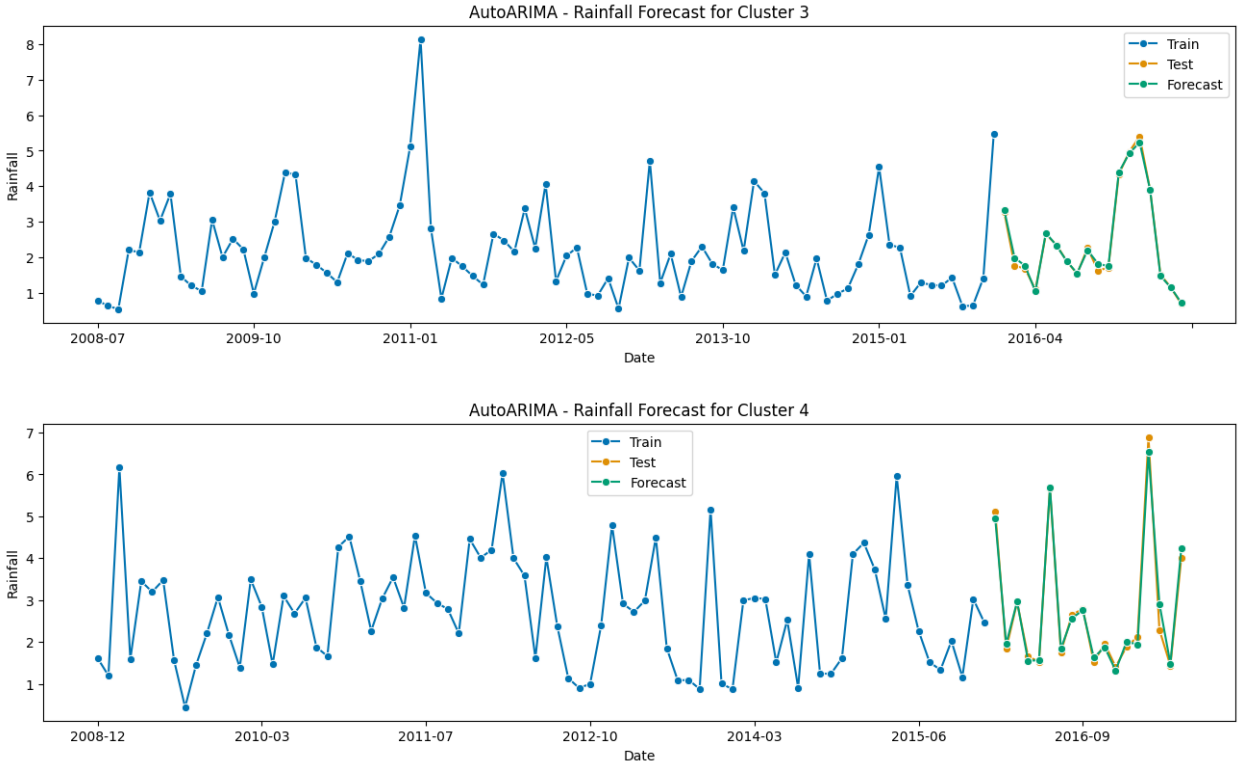
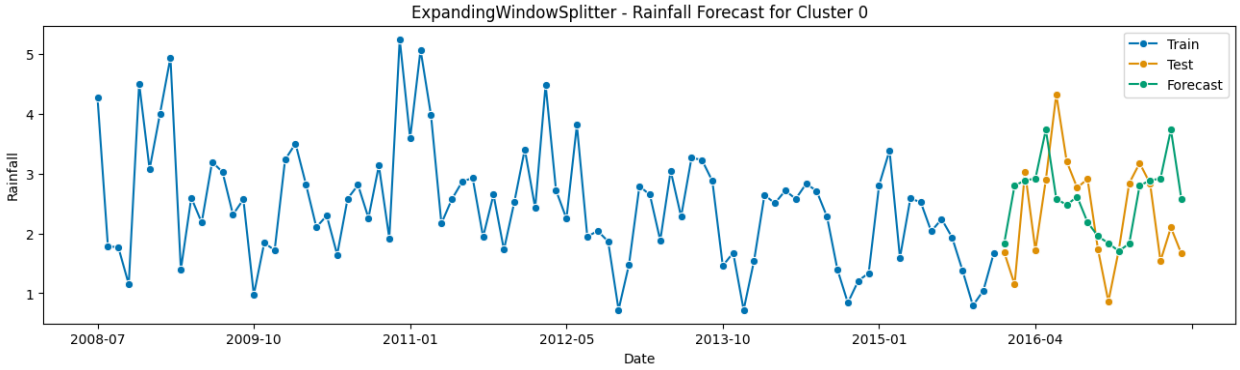
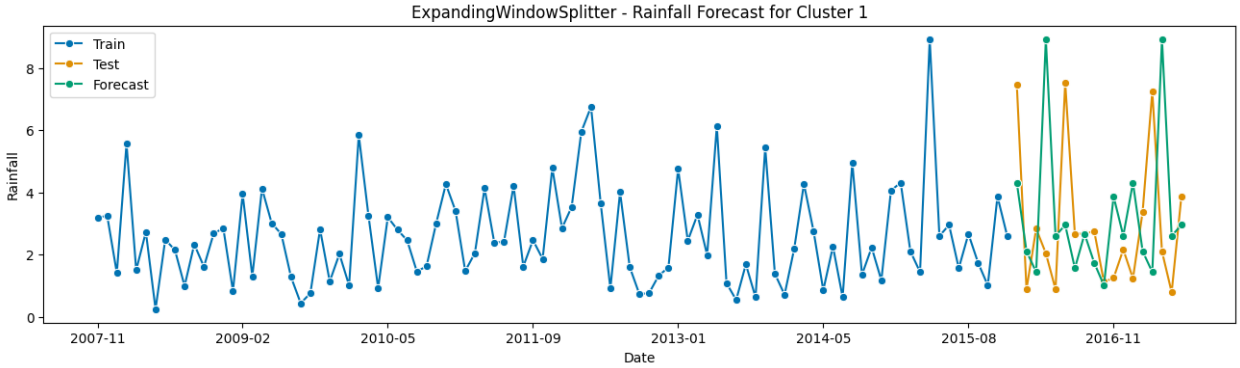


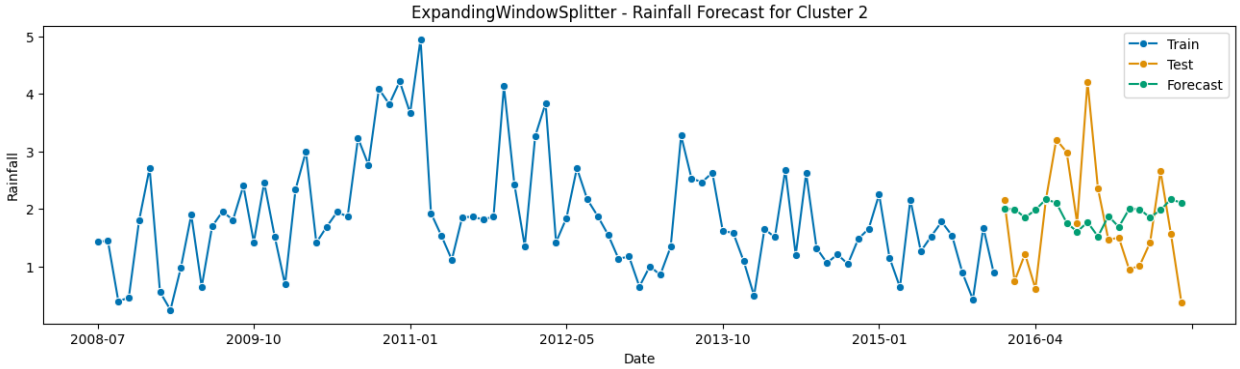
Figure 20: AutoARIMA predictions for clusters 0-4

11.3 Multivariate prediction with ExpandingWindowSplitter

Finally, we deployed a multivariate approach using ExpandingWindowSplitter. As it can be seen from the figures below, the multivariate model didn't deliver the same accurate results as the AutoARIMA model with exogenous variables. This is somewhat to be expected since this prediction does not use the exogenous data of the months in which the rainfall is predicted. This shows again the usefulness of the data used in the AutoARIMA modeling. But the ExpandingWindowSplitter should rather be compared to the naive forecaster which also only uses the information of the training set. Doing this we see that the multivariate prediction seems to be more accurate in the prediction of the trend, but at the same time softens the peaks. This effect can be seen very clearly in Cluster 2. Thus, both models may have their advantages, depending on the use case.


ExpandingWindowSplitter - Rainfall Forecast for Cluster 0
Train
Test
Forecast
Rainfall
Date
2008-07
2009-10
2011-01
2012-05
2013-10
2015-01
2016-04


ExpandingWindowSplitter - Rainfall Forecast for Cluster 1
Train
Test
Forecast
Rainfall
Date
2007-11
2009-02
2010-05
2011-09
2013-01
2014-05
2015-08
2016-11


ExpandingWindowSplitter - Rainfall Forecast for Cluster 2
Train
Test
Forecast
Rainfall
Date
2008-07
2009-10
2011-01
2012-05
2013-10
2015-01
2016-04

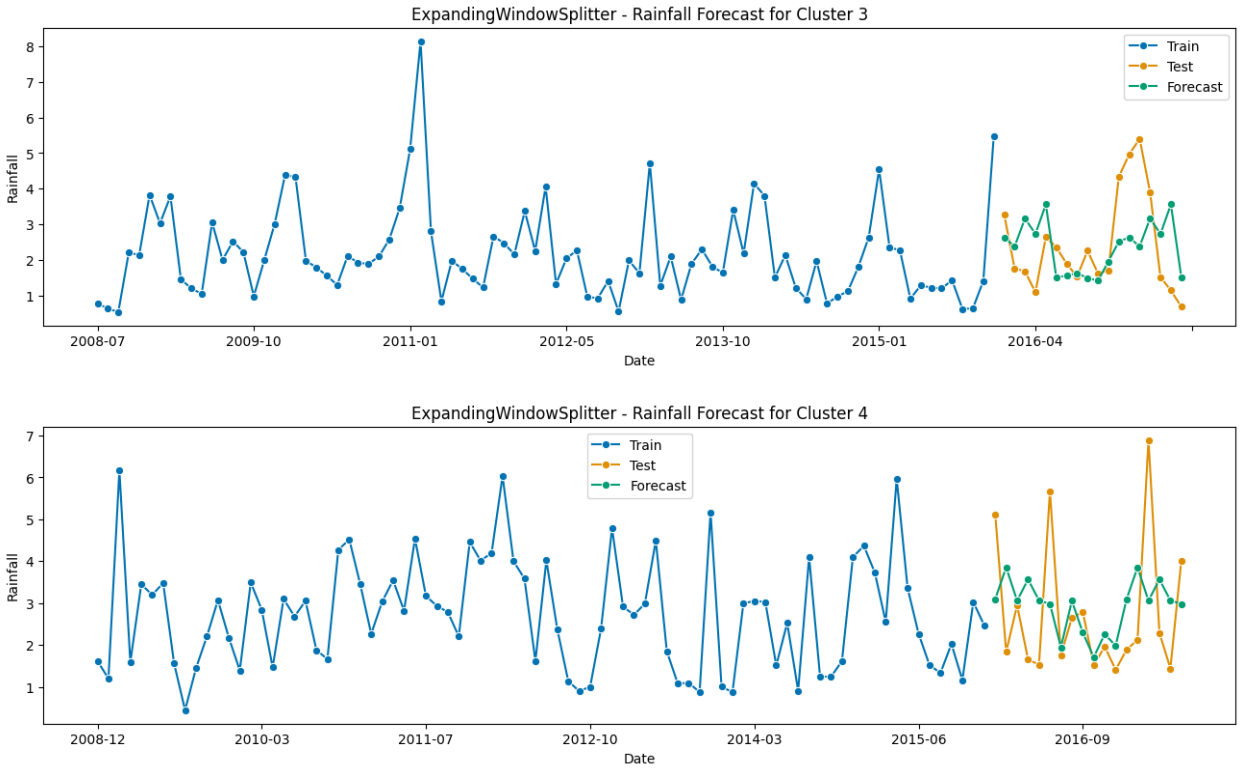


Figure 21: ExpandingWindowSplitter predictions for clusters 0-4

12. Conclusion

Weather forecasting has long been an endeavor for humankind. Rain predictions in particular are a regular part of the news every day all over the world, in addition to temperature and wind predictions. This goes to show that such predictions are of utmost importance for our society, may it be for planning vacations, organizing agriculture or even escaping hurricanes or other life-threatening weather phenomena.

Accordingly, methods for predicting the weather are constantly evolving, going from simple rule of thumbs, over statistical and historical consideration up to complex physical simulations of earth's weather using real-time collected data. And while the physical models of earth's atmosphere are getting better, there is also danger of losing at least some predictability as climate change changes parameters and interactions that have been in place for centuries[2].

[2] https://climate.mit.edu/ask-mit/will-climate-change-make-weather-forecasting-less-accurate

However, accurate climate modeling and weather forecasting also require large computational power, demand high energy inputs and long times for computing the models. Machine learning is already providing a supplementary approach for weather prediction[3] and climate modelling. It is by design somewhat of a middle ground between statistical methods and simulation and may thus be a good additional source of information.

Our project showed that this approach may indeed be worthwhile to pursue. We saw that the XGB model can predict rain with a good accuracy and further tuning of the parameters could improve the results depending on the type of wrong predictions one would like to improve. It would be interesting to analyze similar predictions for other variables such as temperature or wind speed.

Additionally, we saw that the SKTIME models can be very accurate in predicting the mean values of rain in a given month and for many months in the future. The predictions are particularly good, when additional data that can be provided for the model. Such data, however, could also be provided by a physical simulation and one might wonder what capabilities such a combination of physical simulation and machine learning might achieve.

[3] https://www.climateforesight.eu/articles/reacting-to-the-ai-revolution-in-weather-forecasting/