AUSTRALIA PRECIPITATION FORECASTS: A CLASSIFICATION PROBLEM

Giulia Lumia, Marco D'Antoni

ABSTRACT

This report is aimed to show the outcome of a classification analysis carried out in order to make a prediction on whether the next day there will be some kind of precipitation, based on today's weather conditions. Several svm-based binary classifiers were implemented and compared.
Besides, the effect of having multivariate outliers was highlighted.
The selection of the best binary classifier was made maximising a purpose-specific scoring function, by the mean of a nested cross validation method. The final decision fell on the soft-margin linear svm.

INTRODUCTION

Australia is one of the largest countries on Earth, lying between the Pacific and Indian oceans in the southern hemisphere. Its geographical configuration makes this region particularly likely to violent floods and rainfalls. In this context having a machine learning instrument capable of predicting next precipitations might be useful to both experts and public authorities in order to publish meteorological alerts for the population.
All things considered, the main purpose of this analysis is that one of obtaining a classifier capable of returning correct forecasts while minimizing the error of predicting a false absence of rainfall i.e. when the classifier says that tomorrow won't rain but then it rains.
The analysis is made up of two main parts:

1. The first part includes the **DATA PREPROCESSING** along with an **EXPLORATORY ANALYSIS**;
2. The second part includes the **IMPLEMENTATION OF THE SVM CLASSIFIERS** as well as a **COMPARISON** among them.

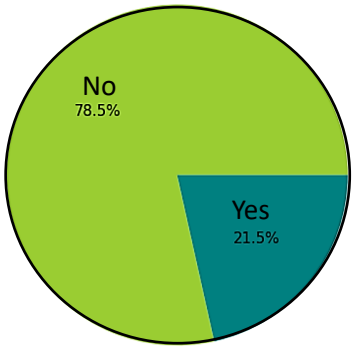
1. DATA PREPROCESSING AND EXPLORATORY ANALYSIS

The dataset used comes from the Bureau of Meteorology of the Australian government's database.
The 9.999 records were collected by 26 of the several weather stations present within national territory.
In the original dataset the daily weather observations were described by different variables:

- **Date**, describing the day under observation;
- **Location**, describing the location of the weather station;
- **MinTemp** (degrees Celsius), describing the minimum temperature in the 24 hours to 9am;
- **MaxTemp** (degrees Celsius), describing the maximum temperature in the 24 hours to 9am;
- **Rainfall**, describing the millimetres of precipitations in the 24 hours to 9am;
- **Evaporation** (millimetres), describing the class A pan evaporation in the 24 hours to 9am;
- **Sunshine**, describing the number of bright sunshine hours in the 24 hours to midnight;
- **WindGustDir**, the direction of the strongest wind gust in the 24 hours to midnight;
- **WindGustSpeed**, the speed (Km/h) of the strongest wind gust in the 24 hours to midnight;
- **WindDir9am, WindDir3pm**, wind direction averaged over 10 minutes prior to 9am / 3pm;
- **WindSpeed9am, WindSpeed3pm**, wind speed (Km/h) averaged over 10 minutes prior to 9am / 3pm;
- **Pressure9am, Pressure3pm** (hPa), atmospheric pressure reduced to mean sea level at 9am / 3pm;
- **Humidity9am, Humidity3pm** (percent), relative humidity at 9am / 3pm;
- **Cloud9am, Cloud3pm** (clouds eighths), fraction of sky obscured by clouds at 9am / 3pm;
- **Temp9am, Temp3pm** (degrees Celsius), temperature at 9am / 3pm**;**
- **RainToday**, describing whether today rained or not;
- **RainTomorrow**, describing whether tomorrow rained or not.

Taking into account the meaning of each variable a congruence check was carried out. It highlighted the presence of 89 records with a minimum daily temperature greater than the temperature at 3pm.
Due to the absence of additional metadata regarding these anomalies, these records were removed.
With respect to the other variables there are no anomalies.

PLOT 1.1: RainTomorrow distribution.



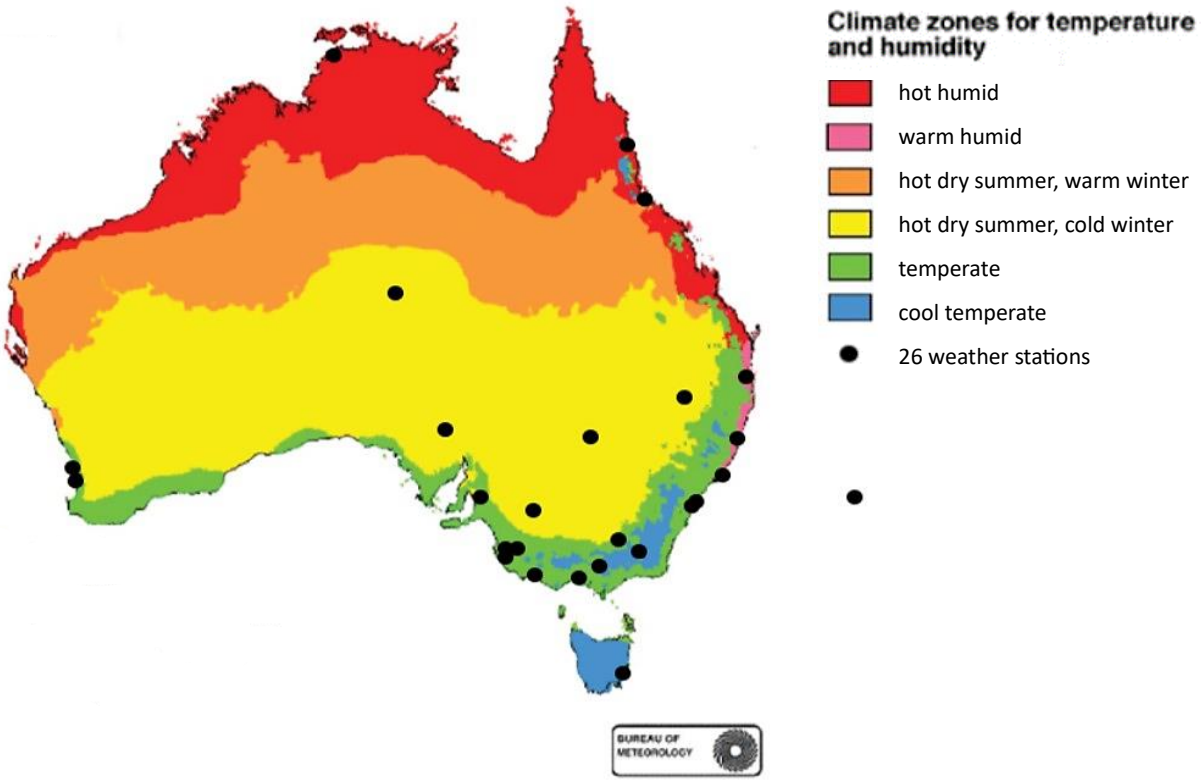
The dataset considered is unbalanced with respect to the response variable RainTomorrow, that is the variable we want our classifier to predict. This particular situation must be seriously taken into account while performing a classification analysis. In fact, it causes the classifier to not train well with respect to the minority class (RainTomorrow = Yes in our case).

The underlying assumption of this analysis consists of the dependence of tomorrow's weather conditions only on today's ones i.e. whether there will be a precipitation at day t+1 depends only on day t and not on all the previous days. Starting from the date variable, which is no longer useful alone, it is possible to derive the corresponding season:

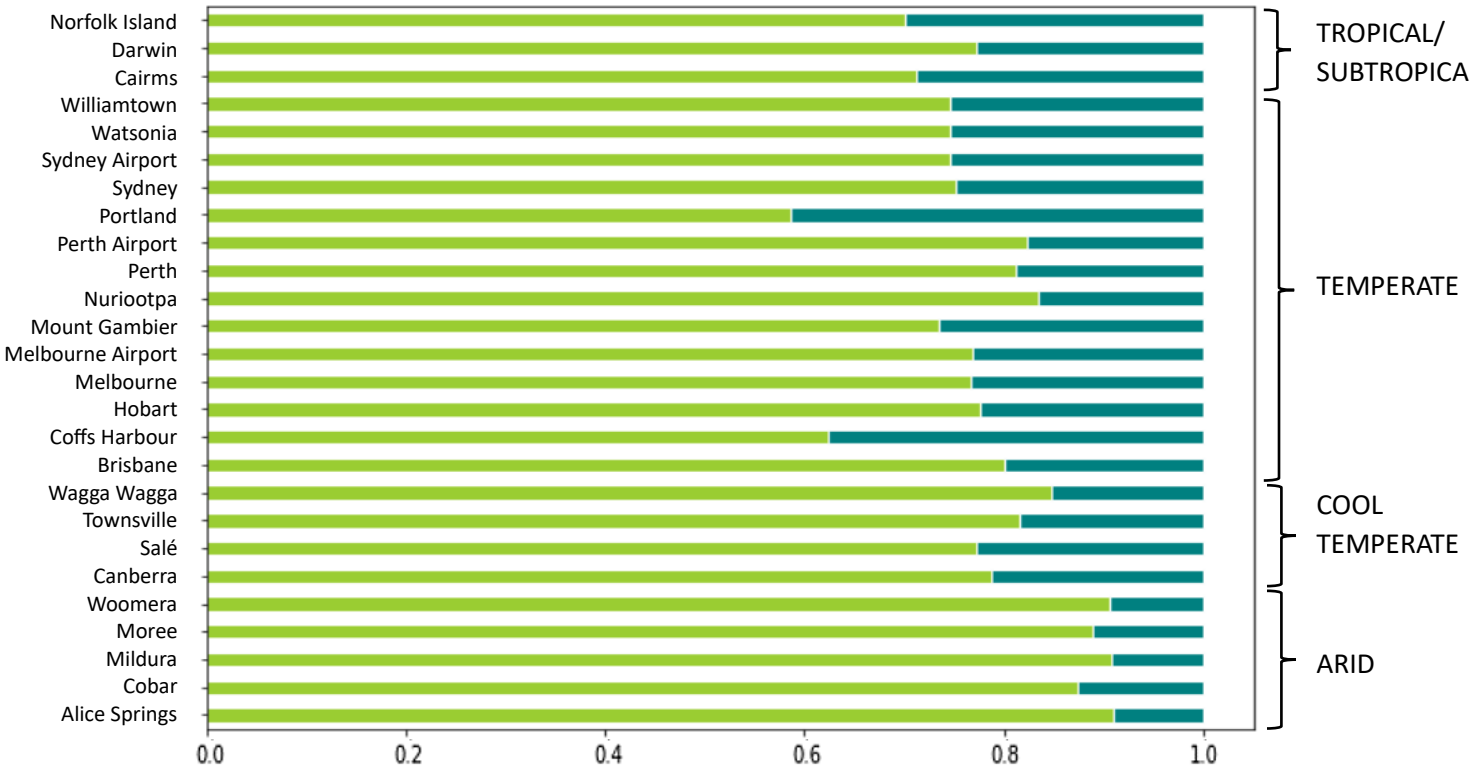
- 21/03 - 21/06 → Autumn
- 22/06 - 22/09 → Winter
- 23/09 - 21/12 → Spring
- 22/12 – 20/03 → Summer

As mentioned before, data were collected by 26 of the several weather stations present within Australian territory. The Bureau of Meteorology itself identified 6 different climate zones on which these stations are located (PLOT 1.2). Since the response distribution in each weather station is similar among stations located in the same climate zone (PLOT 1.3), in order to reduce the dimensionality, the location variable was replaced by the climate variable.

PLOT 1.2: Spatial distribution of the 26 weather station for climate zone.



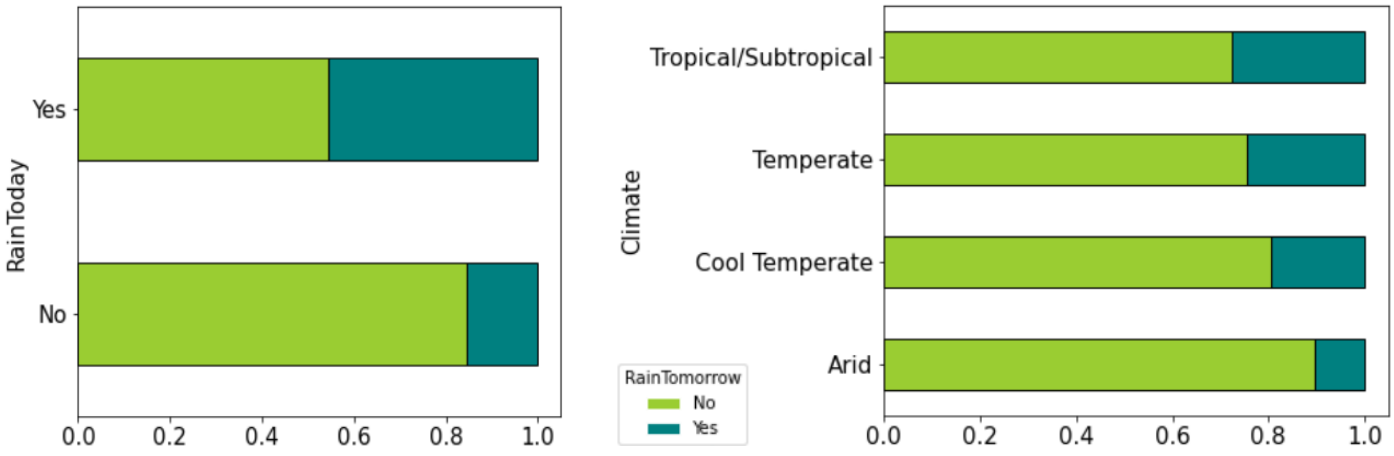
PLOT 1.3: Conditioned RainTomorrow distribution with respect to the location and the climate zone.



GRAPHICAL EXPLORATION OF THE DISCRIMINANT POWER OF EACH VARIABLE

The purpose of this step is to highlight the discriminant power of each categorical and numerical variable. For this reason, conditional distribution of the response with respect to each other variable were graphically represented.

PLOT 1.4, 1.5: Conditional distributions of RainTomorrow with respect to RainToday and Climate.

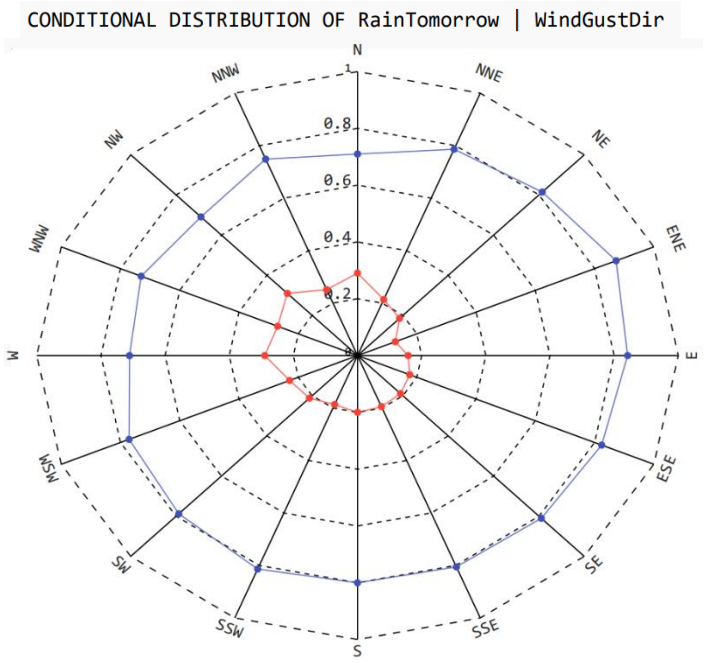


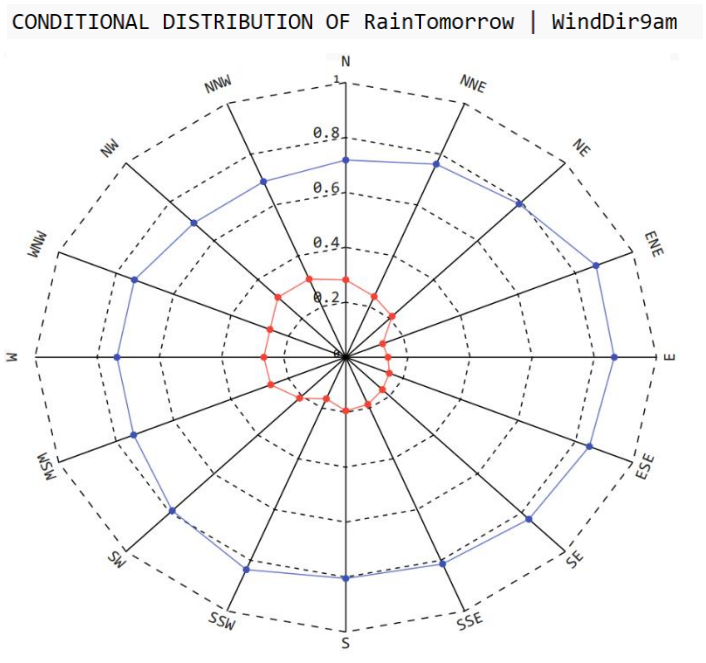
PLOTS 1.4 and 1.5 show that irrespective of the climate zone and today's precipitations, it's always more likely that tomorrow won't rain. According to the particular value of these two variables the probability distribution of the response changes visibly.

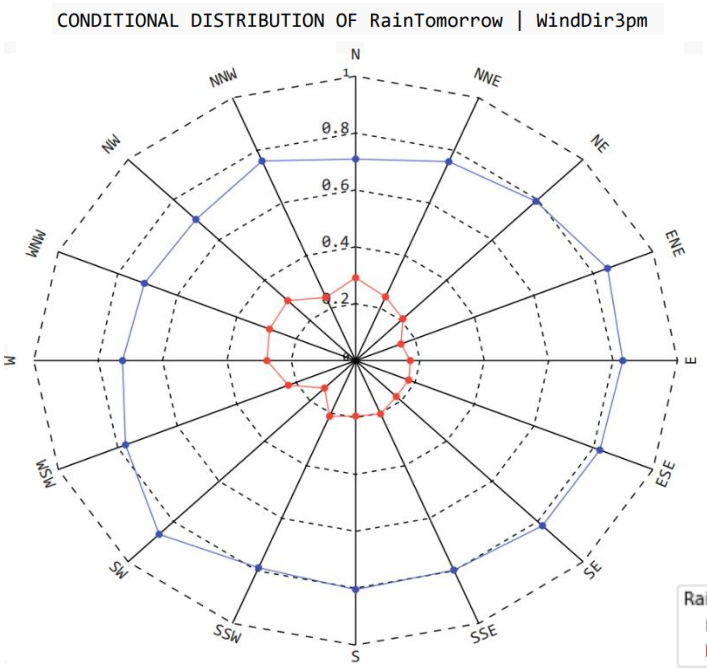
On the one hand it is more likely that tomorrow will rain again if today it rains too (1.4).

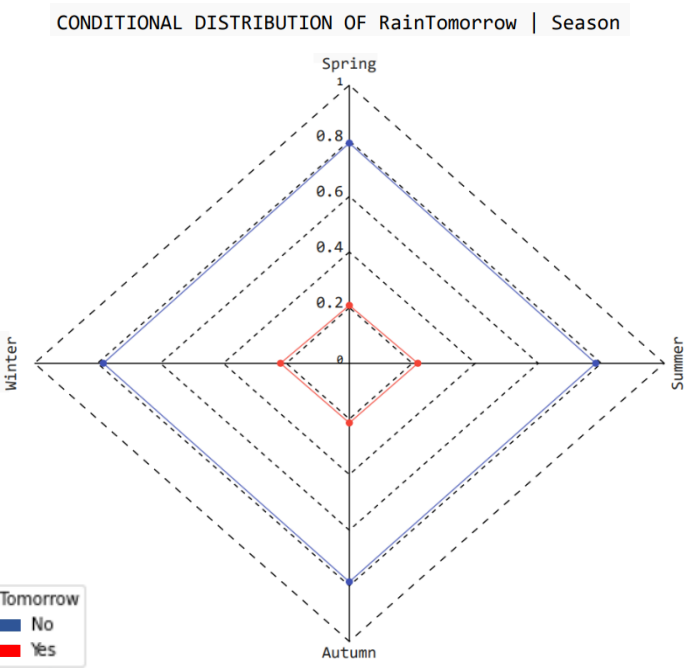
On the other hand the tropical/subtropical region is that one in which it rains more often, while in the arid zone the probability of having a precipitation the next day is the lowest.

PLOTS 1.6, 1.7 1.8, 1.9:





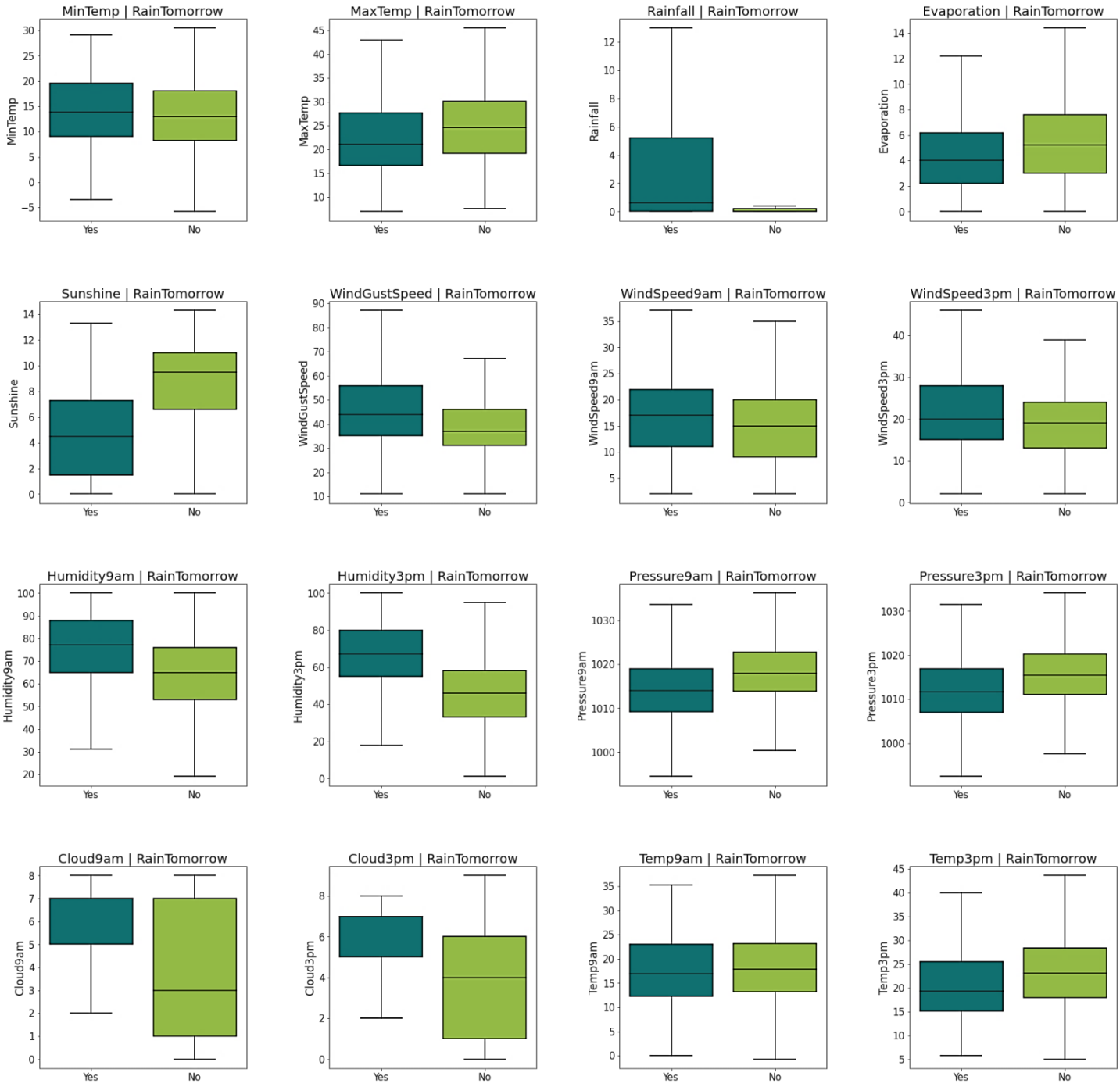




The plots above show how the probability distribution of Rain tomorrow change when conditioned to the value of WindGustDir, WindDir9am, WindDir3pm, Season. The season seems to does not have an effect on the probability of having a precipitation the next day whereas the wind direction does: when the strongest wind gust or the wind at 9am or at 3pm blow from the North, NNW, NW, WNW, W or WSW it is more likely that tomorrow will rain.

The following 16 plots were made to highlight the discriminant power of the numerical variables.
It is easy to notice that while features like those related to the temperature (MinTemp, MaxTemp, Temp9am and Temp3pm) don't seem to have any discriminant power with respect to the values of the response, others, e.g. Sunshine, Humidity at 9am and 3pm, Pressure at 9am and 3pm and Cloud 9am and 3pm may be more helpful in predicting it. As instance, if we know that today there were 10 bright hours it is more likely that tomorrow won't rain.

PLOTS 1.10, 1.11, 1.12, 1.13, 1.14, 1.15, 1.16, 1.17, 1.18, 1.19, 1.20:
Conditional distributions of RainTomorrow with respect to the numerical variables.

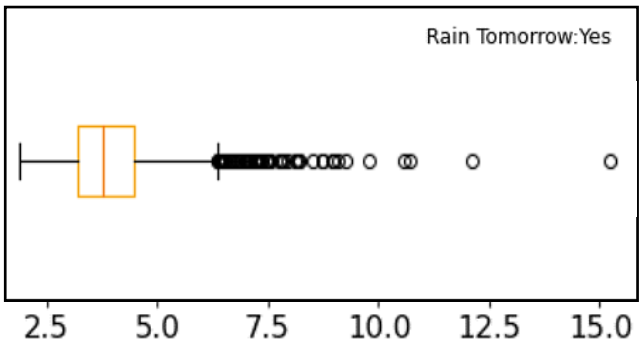


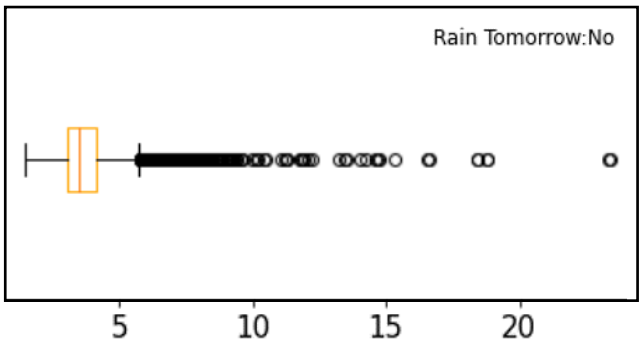
Support vector machine are unable to be applied directly to these variables since they are based on the Euclidean distances. In order to define the distance metrics for categorical variables, a needed preprocessing step is to use dummy variables to represent each value of each categorical variable. Secondly, due to their distinct nature, we need to standardize the numerical variables.

MULTIVARIATE OUTLIERS DETECTION

To detect multivariate outliers the Mahalanobis distance was used. Each observation is assigned a distance value computed considering the distance between the values an observation takes for a set of variables and the mean of the distribution of these variables on the whole dataset.
A response-conditional distribution of these distances can be seen in PLOTS 1.21 and 1.22.

PLOTS 1.21, 1.22: Mahalanobis distances conditional distributions.





Observations having a Mahalanobis distance greater than the right whisk of the boxplot can be considered outliers.

2. IMPLEMENTATION AND COMPARISON OF THE SVM CLASSIFIERS

To classify whether tomorrow will rain or not the support vector machine was used. In a binary classification problem, it's based on finding the best hyperplane that separates the data, when they are linearly separable, or a transformation of them through a so called kernel function, when they are not. The hard version of this algorithm finds the best hyperplane by maximising the margins whereas the soft one (suitable when the data are not linearly separable e.g. due to the presence of outliers), a minimisation of the total classification rule violation is included.
Hyperparameters change according to the kernel function and svm version implemented.
In order to both do model selection and hyperparameters tuning a nested cross validation is needed.
In nested cross-validation, there are two loops of k-fold cross-validation, a 10 folds outer loop for model selection and a 5 folds inner loop for the hyperparameters tuning, using an appropriate scoring function: the precision with respect to the "RainTomorrow: No" category was used, since the main objective is that one of minimizing the dangerous error of predicting a false "RainTomorrow: No".
Both inner and outer folds were stratified with respect to the response with the purpose of taking into account the dataset imbalance.

The initial dataset has been split (stratifying) into a 40% of test set and 60% of training set.
To highlight the effect of having some multivariate outliers into the training set, a comparison was made (TABLE 2.1).

TABLE 2.1: Nested cross validations results for the various kernel functions

KERNEL FUNCTION	MULTIVARIATE OUTLIERS			
	YES		NO	
	Average outer precision	Standard Deviation	Average outer precision	Standard Deviation
Linear	**0,868**	0,011	**0,871**	0,008
Polynomial	0,855	0,011	0,858	0,007
RBF	0,857	0,012	0,861	0,008

In both cases the linear kernel SVM seems to have the highest average precision; this value is slightly lower when the training dataset used includes the multivariate outliers detected.
Cross validation allows us to get a distribution of precision values instead of a single estimate, with a certain standard deviation as dispersion parameter.
TABLE 2.2 shows the results of the nested cross validation applied in both the cases considered.

TABLE 2.2: Nested cross validations results for the linear kernel

	MULTIVARIATE OUTLIERS					
	YES			NO		
Fold	C	INNER PRECISION	OUTER PRECISION	C	INNER PRECISION	OUTER PRECISION
1	0,1	0,863	0,888	0,1	0,872	0,877
2	0,1	0,871	0,859	0,316	0,871	0,87
3	0,316	0,871	0,855	0,316	0,867	0,887
4	0,316	0,87	0,872	0,031	0,871	0,862
5	0,316	0,87	0,865	0,316	0,874	0,871
6	0,1	0,864	0,875	0,316	0,87	0,872
7	0,1	0,87	0,875	0,1	0,87	0,877
8	0,316	0,869	0,88	0,316	0,872	0,868
9	0,316	0,872	0,854	0,316	0,871	0,865
10	0,316	0,87	0,86	0,316	0,873	0,862

Taking into account one of the C values tuned into the inner loop with a soft margin linear kernel (TABLE 2.2), on new data we expect to obtain a precision that deviates from the mean value in TABLE 2.1 at most by a value equal to the relative standard deviation.

The selected model in both cases is a linear kernel one (soft version) with C=0.316. It was adapted to the entire datasets used to implement the nested cross validations (with or without outliers), and subsequently tested on the same test set (PLOT 2.1).

PLOT 2.1: Confusion matrices of the two linear soft svm classifiers.

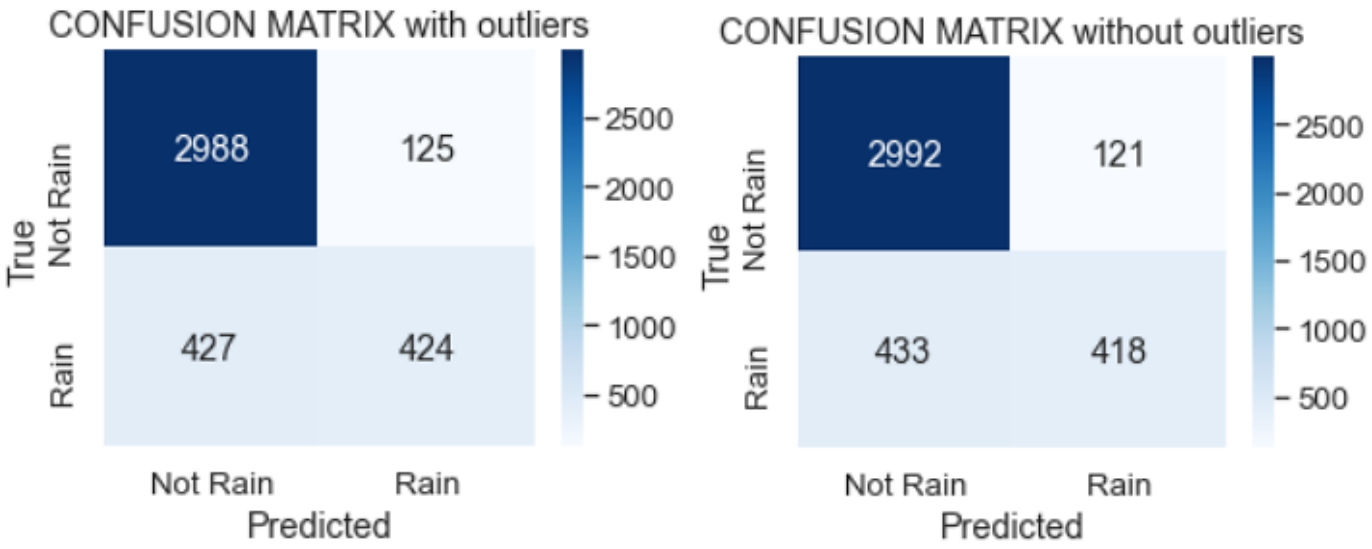


TABLE 2.3: Performance metrics of the two linear soft svm classifiers.

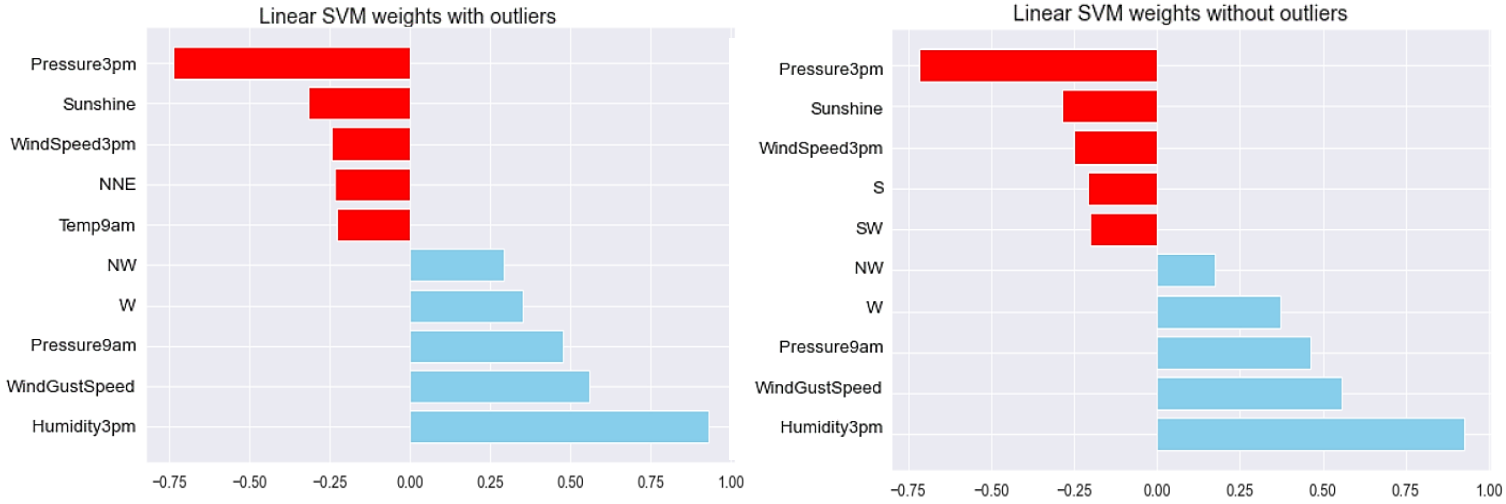
METRICS	MULTIVARIATE OUTLIERS	
	YES	NO
Precision on Not Rain	0,875	0,874
Precision on Rain	0,772	0,776
Accuracy	0,861	0,86

Outliers don't seem to have an effect in the model performance since TABLE 2.3 show similar values.

Since the selected model has a linear kernel, we could give an interpretation to the parameters found by the algorithm, i.e. the weights of the hyperplane associated to each input variabile:

- Its magnitude (absolute value) can be interpreted as the influence of that variable on the function whose sign gives the classification decision;
- Its sign refers to the direction of this influence with respect to a response class rather than the other one.

PLOT 2.2: Variable importance of the two linear soft svm classifiers.



PLOT 2.2 show the 5 most influent variables for each influence direction (positive and negative).
It is possible to notice that within the most influent variables there are those detected during the exploratory analysis.

CONCLUSION

The binary classifier implemented in the context of this analysis is indeed a basic one since it doesn't predict the intensity of the future precipitation. In spite of its simplicity, it can be considered as a useful starting point for future for detailed research developments.