Università di Pisa
Dipartimento di Informatica

Corso di Laurea Magistrale in
Data Science and Business Informatics

Progetto per il corso di
Distributed Data Analysis and Mining

Dataset analysis “Australia, Rain Tomorrow”

A cura di:
Michele Andreucci, 628505
Mario Bianchi, 616658
Francesco Santucciu, 599665
Martina Trigilia, 532155

Anno accademico 2021/2022

1 Data Understanding

1.1 Introduction

The dataset used for this project is called "*Australia, Rain Tomorrow"* , it can be found on Kaggle (https://www.kaggle.com/filhypedeeplearning/australia-rain-tomorrow). The data represents meteorological records, collected by Australian Bureau of Meteorology (BOM), from different *location* of 'Australia for each day from December 2008 to June 2017. The observed data are a list of meteorological features collected during the day, such as *MinTemp*, *MaxTemp*. The attributes *RainToday* e *RainTomorrow*, indicate the presence (***Yes***) or not (***No***) of rainfall during the considered date and for the day following it.The goals for this project are:

- Unsupervised analysis through a clustering algorithm, for the entire dataset, in order to finte some common characteristics of the meteorological observations.
- Dividing the dataset both in political regions and geographical coordinates.
- Prediction of the target variable *RainTomorrow*, both with the original dataset and with the obtained clusters, through the realization of different classification models.
- Forecasting of the variable *Risk_MM*, which indicates the rainfall level of the following day in millimeters, through some regression models.

1.2 Distributions and dataset analysis

The dataset is composed by 142193 records e by 24 features which contains informations about some meteorological elements osserved at different times of the day. The features are briefly described in the following tables 1.1 and 1.2.

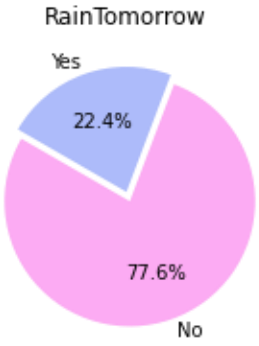
Name	Description	Typology	Values
Date	Observation date	Temporal attribute	{"01/12/2008",...}
Location	Observation place	Categorical	{Sydney, Canberra…}
WindGustDir	Gust direction	Categorical	{WSW, WNW, W,..., E}
WindDir9am	Average wind direction 10 min Before 9 am	Categorical	{WSW, WNW, W,..., E}
WindDir3pm	Wind direction at 3 pm	Categorical	{WSW, WNW, W,..., E}
RainToday	Rainfall presence or absence in that day	Binary	{Yes, No}
RainTomorrow	Presence or absence of rain on the date following the current on ***Target Variabile***	Binary	{Yes, No}

Table 1.1: Categorical and binary attributes description

Nome	Descrizione	Tipologia	Range(Min,Max)	Media
MinTemp	Min temprerature until 9 am*	Continuous	{-8.5, 33.9 }	12.18
MaxTemp	Max temperature until alle 9 am*	Continuous	{-4.8, 48.1}	23.23
Rainfall	Rainfall until 9 am*	Continuous	{0.0, 371.0}	2.35
Evaporation	Evaporation until 9 am*	Continuous	{0.0, 145.0}	5.47
Sunshine	Level of illumination until midnight*	Continuous	{0, 14.5}	7.62
WindGustSpeed	Speed in km/h of the strongest wind until midnight*	Continuous	{6.0, 135.0}	39.98
WindSpeed9am	Average speed in km/h 10 min before 9 am	Continuous	{0.0, 130.0}	14.0
WindSpeed3pm	Average speed in km/h 10 min before 3 pm	Continuous	{0.0, 87.0}	18.4
Humidity9am	Relative humidity (in percentage) until 9 am	Continuous	{0.0, 100.0}	68.84
Humidity3pm	Relative humidity (in percentage) until 3 am	Continuous	{0.0, 100.0}	51.48
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea level at 9 am	Continuous	{980.5, 1041.0}	1017.65
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea level at 9 am	Continuous	{977.1, 1039.6}	1015.25
Cloud9am	Part of the sky obscured by clouds at 9 am. 0 indicates completely clear sky and 9 completely covered with clouds	Discrete	{0.0, 9.0}	4.44
Cloud3pm	Part of the sky obscured by clouds at 3 pm	Discrete	{0.0, 9.0}	4.50
Temp9am	Temperature (in Celsius) at 9 am	Continuous	{-7.2, 40.2}	16.99
Temp3pm	Temperature (in Celsius) at 3 pm	Continuous	{-5.4, 46.7}	21.68
RISK_MM	Rainfall level in mm recorded during the following day. ***Continuous Target Variabile***	Continuous	{0.0, 371.0 }	2.36

Table 1.2: Numerical attributes description

*The following measures are all to be understood as calculated in the previous 24 hours



The dataset is unbalanced with respect to the target variable: only 22.41% of the cases rain is recored for the following day (*RainTomorrow* = 'Yes'), as it possible to see in the Figure 1.2. The variable *RainToday* has a similar distribution too: in 22.12% of the observations *RainToday* = "Yes", while for 76.89% *RainToday* = "No".From further analysis, we find that if it rains today, then there is about a 46.40% probability that it will rain the following day as well.

Figure 1.1 :Target variable distribution

Moreover, observing the distributions of the variables *Humidity9am* and *Humidity3pm* in Figure 1.3, we notice that when *RainTomorrow* = "Yes" the recored humidity during the previous afternoon has higher values, while this does not happen for the one recored for the moring before. The variables which record the informations about the temperature (*Temp3pm*, *Temp9am*, *MaxTemp* e *MinTemp*) follow a shape-bell distribution (Figura 1.2).As regards, however, all the variables which give informations about wind (*WindGustDir*, *WindDir9am*, *WindDir3pm*), the direction *"WSW"* is the most frequent one. In the end, regarding the coninuous target variable *RISK_MM*, this last one has its highest values when it rains for at least two consecutive days.

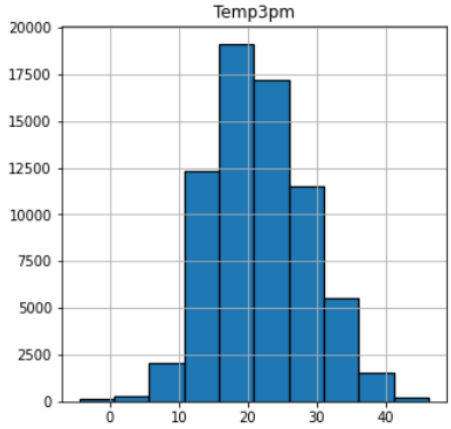


Figure 1.2: Temp3pm distribution

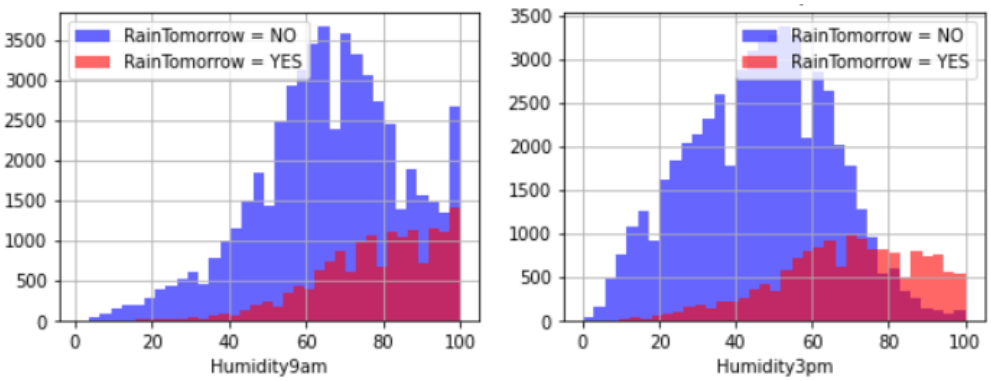


Figure 1.3 : Humidity9am and Humidity3pm distribution w.r.t. target variable

1.3 *Location* attribute analysis

The attribute *Location* was an important object for the analysis because its distribution was fundamental to valuate the feasibility of some of our goals for this project. It presents 49 different places. In order to understand where they find these places, we decided to get the coordinates of them, because they were not in the intial

dataset .For this purpose, it was needed to clean the attribute *location* values, because some of them was inserted without spacing and for this reason it was impossible to get the coordinates (for an instance, the value “BadgerysCreek” instead of Badgerys Creek). Once we got the coordinates, it was possible to visualize them on the map (Figure 1.4 on the right), where we can immediately notice that the different places appear in the dataset with a similar frequency (dimensione delle bolle), except for some others like *Nihil, Uluru and Katherine* (Figure 1.1, on the left). In particular, it notices that in the plot in Figure 1.4 (left) it was gotten taking into account a sample of the dataset, but every different location has almost 3K records. Moreover, we can notice how most of the locations find on the south east coast, while in the central part there are few observations, probably it is due to the demographical concentration of Australia .

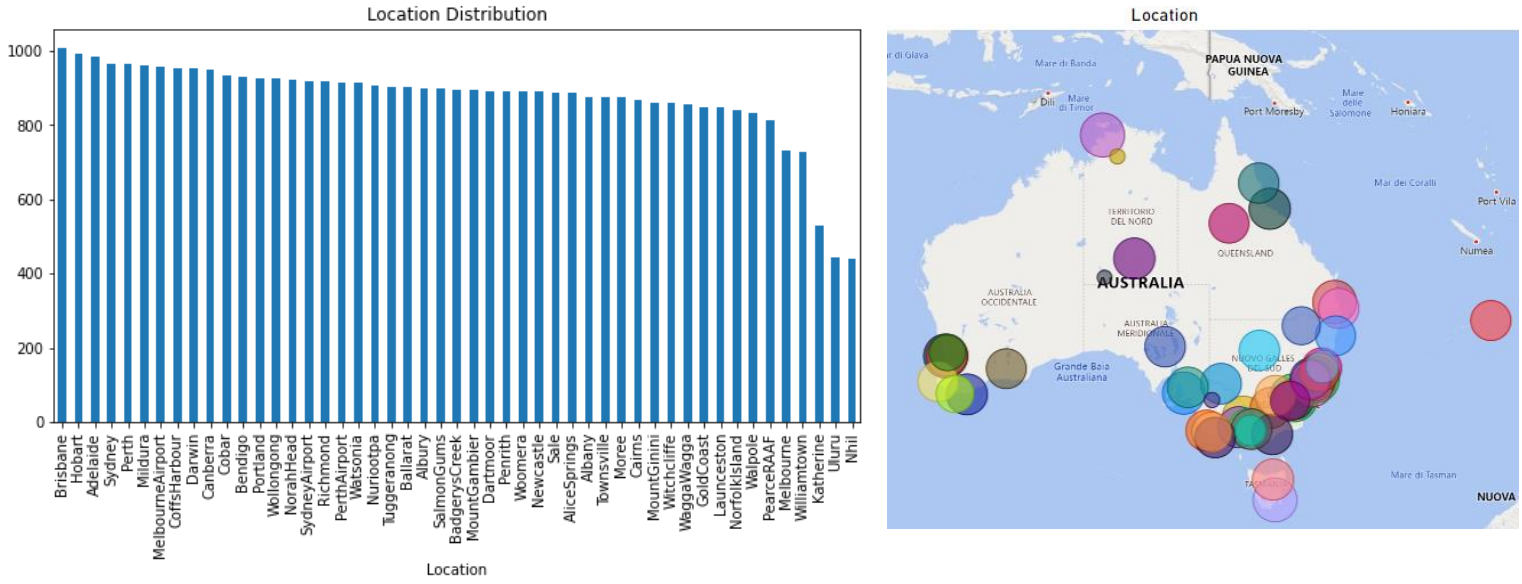


Figure 1.4 : *Location* attribute - Bar Chart and Map

1.4 Correlations

In order to get the correlation between attributes, it was used two correlation indexes . Through the pearson’s coefficent it was passible to calculate the correlation between the features on the dataset, it is represented by the heatmap in Figure 1.2.

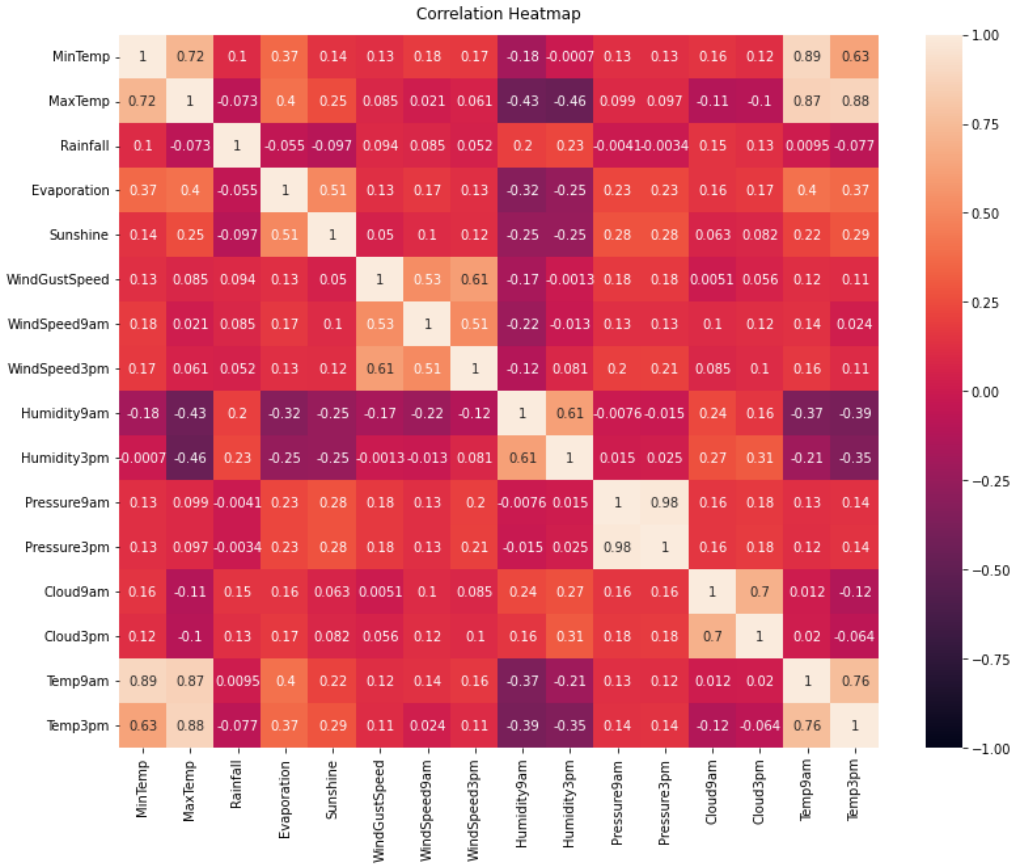


Figure 1.5:
Correlation between attributes

We can notice how the attributes that present a high correlation are the ones that indicate the temperature and the attributes that measure the same thing but in different time solts (es. *Pressure9am* and *Pressure3pm*). On the contrary, the Point-Biserial Correlation Coefficient[1] was used to see the correlation of the various numerical attributes with the binary attribute *RainTomorrow*. Using this coefficient, we can notice how the most correlated attributes with respect to the target variable are: *Rainfall*, *Sunshine*, *WindGustSpeed*, *Humidity9am*, *Humidity3pm*, *Pressure9am*, *Pressure3pm*, *Cloud9am*, *Cloud3pm*.

2 Data Preparation

2.1 Creation of the attributes *Month*, *Season* and *Region*

We decided to get the attributes *Month* (con valori '01', '02', '03', …, '12') and *Season* ('Winter', 'Fall','Spring','Summer') from the variable *Date,* through *regular expressions*, as we thought it interesting to go and analyze the behaviour of the rain during the different moths and seasons. The months with the highest percentage of *RainTomorrow* = 'Yes' are the summer ones, as we could expect. The variable Region was also integrated, which represents the political region of every location in the dataset, through an additional csv (*cities_australia.csv*). This attribute was added for the clustrering task too.

[1] https://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient

2.2 Missing values: imputation and comparison between methodologies

The dataset has a lot of missing values. The attributes with the most missing values turns out to be *Sunshine* with 48% of the missing obsevations, *Evaporation* with 43%, *Cloud3pm* with 40% e *Cloud9am* with 38%. The Figure 1.5 reports the total number of missing values for every attribute.

Given the fact that there are a high number of *missing values*, it was decided to deleting records with missing values, and then proceding with the imputation of them. Considering the delicacy of this phase on the rest of the analysis, it was decided to try different approches.

Sunshine → 67816
Evaporation → 60843
Cloud3pm → 57094
Cloud9am → 53657
Pressure9am → 14014
Pressure3pm → 13981
WindDir9am → 10013
WindGustDir → 9330
WindGustSpeed → 9270

WindDir3pm → 3778
Humidity3pm → 3610
Temp3pm → 2726
WindSpeed3pm → 2630
Humidity9am →1774
Rainfall →1406
RainToday → 1406
WindSpeed9am → 1348
Temp9am → 904

MinTemp → 637
MaxTemp → 322
Date → 0
Location → 0
RISK_MM →0
RainTomorrow → 0
Month → 0
Season → 0
Region → 0

Figure 1.5: Missing Values

1. The first approch involves the average grouping by attributes. The first four attributes with the most missing values were grouped first by *Location*: the missing values are evenly distributed among the cities, and cities with the most missing values for the attributes in question have around 3K. Next, the missing values were grouped by *RainToday* and *RainTomorrow*: being binary attributes, it was easily possible to calculate the average for both classes. An exemple is shown below for *Cloud9am*.

```
+---------+-----------------+
|RainToday|    avg(Cloud9am)|
+---------+-----------------+
|       No|3.939796797480764|
|      Yes|6.018474088291747|
|     null|5.858288770053476|
+---------+-----------------+
```

```
+------------+------------------+
|RainTomorrow|     avg(Cloud9am)|
+------------+------------------+
|          No|3.9322820037105752|
|         Yes|  6.09999030161963|
+------------+------------------+
```

Grouping by *Location* immediately showed that it was not suitable for replacing missing values, as the grouping table had missing observations. It was then decided to group by Month and further group by the attribute with the highest correlation. An example of grouping *Evaporation* by *Sunshine* and by *Month* is shown below.

```
+-----+--------+------------------+
|Month|Sunshine|  avg(Evaporation)|
+-----+--------+------------------+
|   01|    10.3|   8.9438202247191|
|   01|     2.9|           6.86875|
|   01|     8.2| 8.463888888888889|
|   01|     5.9| 6.769230769230769|
|   01|    13.8|7.7700000000000005|
+-----+--------+------------------+
```

However, this method of substitution was not preferred, as it is difficult to justify replacing the missing value of a single day with the average of the *Month* grouping, given the multiple weather conditions that can occure during a month. In other words is too wide a time frame to deduce the value of an attribute related to a single day.

2. As a second approach, the possibility of replacing missing values through regression was evaluated. However, in this case too, the decision was not to proceed with the substitution, as there is a high correlation between the missing values of the attributes for which to perform the substitution (the correlations between *missing values* are shown in Figure 2.1).The result is that for the regression of a value, it is not possible to choose the most useful feature for the same regression.

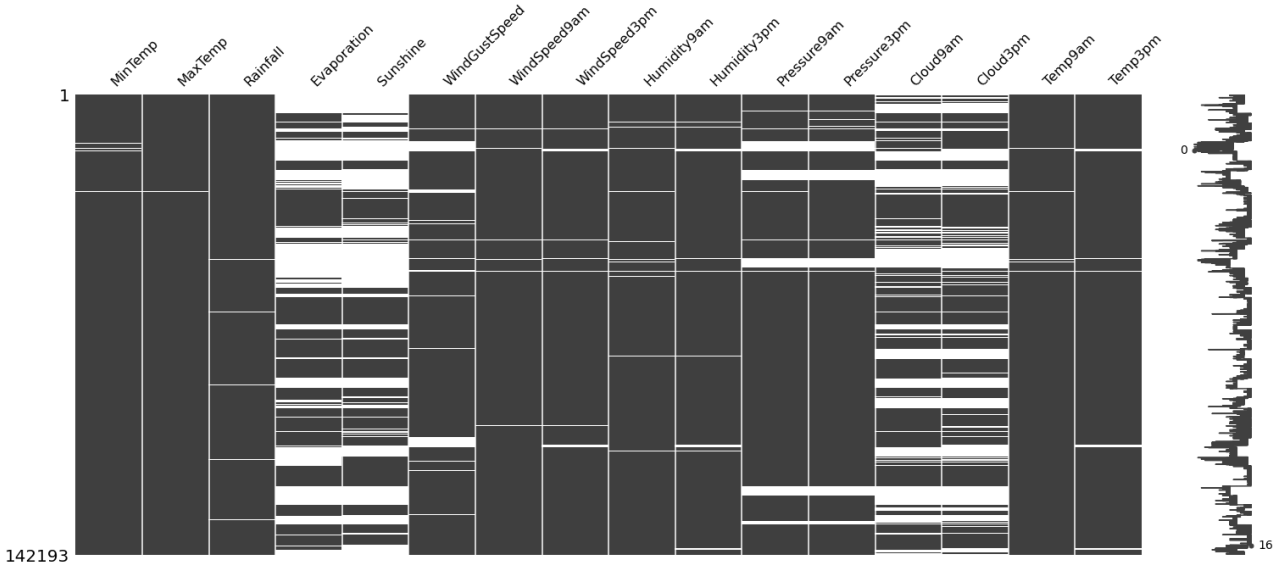


Figure 2.1 - Correlation between missing values

3. The method chosen for replacing missing values involves using the median, as it is more resistant to outliers. This approach was used for the first 6 attributes in terms of number of missing values, namely *Evaporation*, *Sunshine*, *Pressure9am*, *Pressure3pm*, *Cloud9am* e *Cloud3pm*. To obtain more accurate values, it was decided to temporarily add 4 binary variable to the dataset:
   - *it-will-rain*: *RainToday* = 0 e *RainTomorrow* = 1 (“today no rain, tomorrow it will rain”)
   - *no-rain*: *RainToday* = 0 e *RainTomorrow* = 0 (“today no rain and tomorrow it will not rain”)
   - *it-rain*: *RainToday* = 1 e *RainTomorrow* = 1 (“today rains and tomorrow too”)
   - *end-rain*: *RainToday* = 1 e *RainTomorrow* = 0 (“today rains but tomorrow it will not rain”)

   After obtaining these four phenomena, the differences in thei distributions were verified through a median test statistic. It was discovered that for *Evaporation* the process of generating values for *it-rain* and *end-rain* is different: which implies that the median to be assigned is different since the phenomena are different from each other. For the other variables in the dataset; the missing values were replaced with the overall median, since for them the aforementioned four binary attributes do not follow different distributions.

Subsequently, it was wanted to compare the median method with regression. A attribute was chosen for regression,, *Sunshine*, and a sample of observation without missing values was selected (otherwise it would not have been possible to perform the regression): for this reason, the variables *Evaporation*, *Cloud9am*, *Cloud3pm*, *Pressure9am*, *Pressure3pm* were excluded. With the obtained dataset, consisting of 84525 record, the predictions for the *Sunshine* values were obtained. The regression obtained a R^2 of 0,21 and a RMSE of 2,5. An example of the results obtained is reported below.

```
+--------------------+--------+------------------+------------------+
|            features|Sunshine|        prediction|        difference|
+--------------------+--------+------------------+------------------+
|[0.0,13.399999618...|     5.6|7.0110699313355855|1.4110700267030172|
|[1.0,7.4000000953...|     5.6| 7.677187066329452|2.0771871616968838|
|[2.0,12.899999618...|     5.6| 7.607937954908657|2.0079380502760884|
|[3.0,9.1999998092...|     5.6|  8.30139187127823|2.7013919666456623|
|[4.0,17.5,32.2999...|     5.6| 6.887682983888796|1.2876830792562277|
+--------------------+--------+------------------+------------------+
```

It is important to note that the *Sunshine* column is composed of both original values and values obtained through substitution. The statistics for the *difference* column were then observed:

```
+-------+--------------------+
|summary|          difference|
+-------+--------------------+
|  count|               84525|
|   mean|  1.9895800645913424|
| stddev|  1.5151382702108818|
|    min|3.129227160947323...|
|    25%|  0.7564876457468417|
|    50%|   1.651940445529628|
|    75%|  2.9373357384098977|
|    max|   9.208369080744045|
+-------+--------------------+
```

The results show significant differences, considering that the attribute range for the *Sunshine* variable is [0; 14,5]and that the average deviation is about 2 hours of sun. At the same time, it is not possible to detrmine which of the two methods is actually the best, as there is no way to find the corresponding values for the *missing values*. Moreover, it is important to remember that the features used for the regression do not include the attributes with the highest correlation with the variable to be predicted, i.e the attribites that generally guarantee a better prediction. For these reasons the substitution method used was considered valid.

3 Clustering and Classification

As mentioned in the introduction, the classification task was approached using different methodologies. First, a standard classification was performed on the entire dataset. Subsequently, the dataset was divided into clusters in the following ways:

1. Through K-means using some numerical variables.
2. Through K-means using only geographical coordinates.
3. Dividing the dataset into political regions based on the *Region* variable in advance.

The same algorithms used for the standard classifcation were applied to the different clusters obtained, and the results were compared to each other.

3.1 Standard classification of the variable RainTomorrow

For the task of classifying the RainTomorrow variable, the entire dataset tas used without distrinction as far as the records are concerned. On the other hand, regarding the choice of attributes, these were selected both by observing the correlation with the *RainTomorrow* attribute (in particular, allattributes with an absolute value greater than 0.2 were selected), and by observing the correlations among the same attributes and discarding all those that had a high correlation: this wat redundancy among the same was avoided. Regarding the choices made for the classification task, three different data scaling techniques were executed separately. Firstly, the *MinMax Scaler* and the *Standard Scaler* were used. Then we used PCA with k = 3, both as *scaling* and as

feature reduction technique. As for the models used, for all three preprocessing we applied the following machine learning algorithms: Logistic regression, Decision tree classification, Random forest classification, Linear SVM classification.

As the final part of the overall classification, the same 4 classification algorithms were executed. For the choice of the hyperparameters, a grid-search was carried out with CV. La CV technique used was k-fold with k = 5. For the logistic regression, the hyperparameters for which tuning was performed were the maximum number of iterations and the *elasticNet* for deciding the penalty of the model. For the decision tree, tuning was performed for the maximum number of intervals and the depth of the tree; for the Random forest, the number of trees and the maximum depth of the trees, while for the linear SVM the maximum number of iteration and the *fitIntercept.*

Subsequently we decided to try to classify keeping all the continuous features of the dataset, except for RISK_MM, using PCA with k = 10, as the *feature reduction* method. The results forthe chosen metrics are shown in tables 3.1, 3.2, 3.3, 3.4. it can be noticed that using k=10, the results improve both compared to using pca with k=3 and for the use of other scaling techniques.

(Standard scaler)	Accuracy	Precision	Recall	AUC
LR	0.83	0.62	0.54	0.72
DT	0.84	0.72	0.46	0.70
RF	0.84	0.71	0.48	0.71
Linear SVM	0.84	0.73	0.46	0.70

Table 3.1 - StandardScaler Results

(MinMax scaler)	Accuracy	Precision	Recall	AUC
LR	0.83	0.62	0.54	0.72
DT	0.84	0.72	0.46	0.70
RF	0.85	0.73	0.48	0.71
Linear SVM	0.84	0.73	0.46	0.70

Tabella 3.2 - MinMax Results

(PCA k = 3)	Accuracy	Precision	Recall	AUC
LR	0.83	0.71	0.37	0.66
DT	0.83	0.68	0.48	0.70
RF	0.84	0.69	0.48	0.71
Linear SVM	0.83	0.70	0.38	0.66

Tabella 3.3 – Results with PCA k=3

(PCA k = 10)	Accuracy	Precision	Recall	AUC
LR	0.84	0.71	0.48	0.71
DT	0.84	0.71	0.48	0.71
RF	0.84	0.72	0.48	0.71

Tabella 3.4 - Risults with PCA k=10

3.2 K-Means and intra-cluster classifications:

The definition of cluster could lead to thinking that the classification carried out within a cluster must necessarily be more accurate than the global classification. In reality, the division of the dataset bring with it two problems:the imbalance of the target variable within the clusters and the reduction of the records to train the classifier. After experimenting with K-Means both in its standard version in its Bisecting K-Means version, it ws decided to opt for the first. As with the overall classification, the attributes used are *Rainfall*, *Sunshine*, *WingGustSpeed*, *Humidity3pm*, *Pressure9am*, *Cloud3pm* and *RainToday*. It was necessary to exclude

RainToday and the two target variables *RainTomorrow* and *RISK_MM*,as they would have madethe analysis insignificant. Subsequently, the values contained in the vectors were processed with *StandardScaler*.
For the choice of K, both *Silhouette Score* and SSE were taken into considerationthe results are described by Figure 3.1. It was decided to use k=6.

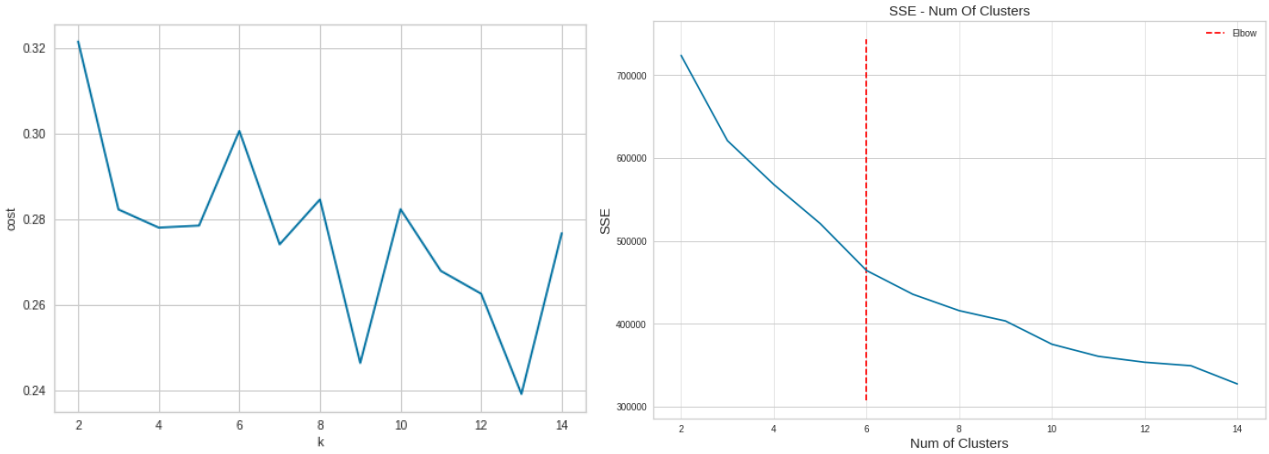


Figure 3.1 - SSE & Silhouette score

In the table 3.5 , for each of the 6 clusters, it was reported the distribution of the target variable.

Cluster ID	RainTomorrow = Yes	RainTomorrow = No	Nr Records
0	1074	23322	24396
1	1738	30523	32261
2	6997	11027	18024
3	6233	23466	29699
4	3422	10006	13428
5	8881	3163	12044

Table 3.5 - Clusters distributions

As can be observed, the target variable *RainTomorrow* is distributed significantly differently in each cluster, and the class 1 of RainTomorrow only represents the dominant class in cluster 5, with 74%of occurrences. Cluster 0 and 1 are the most imbalanced: the attribute 1 of *RainTomorrow* represents respectively 4,4% and 5,4%.
The distributions of other attributes within the clusters were also observd. For the attribute *RainToday*, the same applies as previously stated for *RainTomorrow*: cluster 0 and 1 are clearly imbalanced towards 0 (i.e. 'No'), while cluster 5 is the most balanced. The main difference is that *RainToday* = 1 is not the majority class in any cluster. In other words, cluster 0 and 1are the clusters in which the phenomenon of rain is rarer, while cluster 5is the one where it occurs more often. Subsequently, an attempt was made to understand which clusters could be identified with a season of the year. In fact, given that the attribute is not balanced within the dataset, it was not possible to obtain a result of this typeit was possible to observe that is the dominant season in all clusters except *cluster* 2, with a maximum percentage of 56% in cluster 0 and a minimum percentage of 25% in cluster 2. In cluster 2, the predominant season is Summer, with 40% of observations. *Spring* on the other hand, has its highest value in cluster 1, with 32% of observatons. The highest number of observations of *Winter* is observed in cluster 1, with 10%.

Summing up the information obtained on the clusters:

- Cluster 0, consisting of 24396 records, the occurrence of rain is very rare, and 56% are recorded in the fall;
- Cluster 1, consisting of 32261 records, the occurrence of rain is very rare and the records are distributed evenly among Fall (33%), Spring (32%) and Summer (30%);
- Cluster 2, consisting of 18024 records, is the only one where *Summer* is the dominant season (40%);
- Cluster 3 e 4 consist of 29699 and 13428 records,and do not have particularly distinctive elements;
- Cluster 5, consisting of 12044 records, is the smallest cluster and is the one where rain is the most frequent.

For the classification, in addition to the standardization already mentioned, it was decided to perform PCA with K=3 given the good results obtained in the first classification. Three classifiers were used: Decision Tree, Random Forest and Logistic Regression. *Grid Search* and *Cross Validation* with *numFolds* equal to 5 were performed for each of the three. The evaluation metrics of the classification performed are reported in Table 3.6.

Model	Measure	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Decision Tree	under ROC	0.53	0.51	0.70	0.55	0.57	0.57
	under PR	0.10	0.08	0.61	0.27	0.36	0.76
	Accuracy	0.94	0.93	0.73	0.75	0.72	0.76
	Recall	0.09	0.04	0.58	0.21	0.26	0.92
	Precision	0.16	0.11	0.67	0.31	0.43	0.76
	F-measure	0.11	0.06	0.62	0.25	0.32	0.83
Random Forest	under ROC	0.5	0.5	0.70	0.51	0.56	0.57
	under PR	0.04	0.05	0.63	0.34	0.56	0.57
	Accuracy	0.96	0.94	0.73	0.79	0.75	0.76
	Recall	0.0003	0.0001	0.56	0.02	0.16	0.95
	Precision	1	1	0.69	0.54	0.60	0.77
	F-measure	0.0006	0.0003	0.62	0.04	0.25	0.85
Logistic Regression	under ROC	0.50	0.50	0.69	0.50	0.59	0.60
	under PR	0.10	0.06	0.59	0.21	0.39	0.78
	Accuracy	0.95	0.94	0.71	0.79	0.75	0.73
	Recall	0.003	0.0001	0.58	0	0.28	0.88

	Precision	0.18	1	0.64	1	0.47	0.78
	F-measure	0.006	0.0004	0.61	0	0.35	0.83

Table 3.6 - Results

The results are consistent with the analysis of the clusters: the cluster 0 and 1, being the most imbalanced,havethe worst results. The cluster 5, which is the most balanced and the only one with prevalence of 1 for *RainTomorrow*, turns out to be the easiest to classify.

3.3 Geographical clustering and intra-cluster classifications:

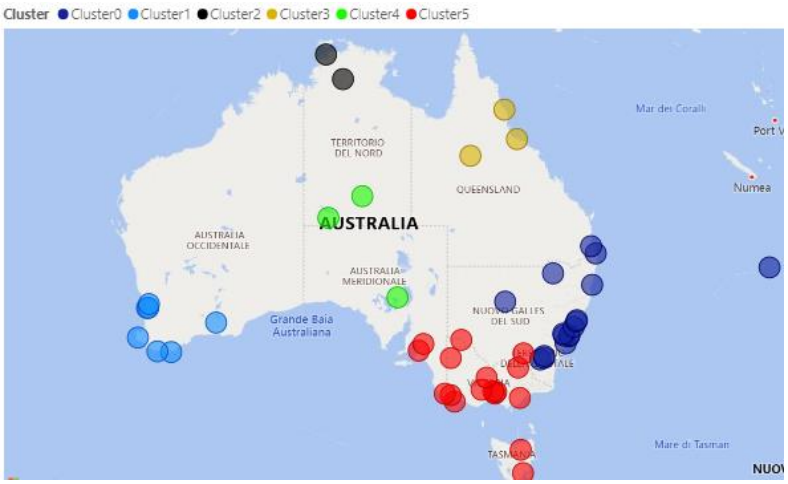
The dataset was divided into geographical clusters using the K-Means algorithm with longitude and latitude variables, obtained as explained in Section 1.2.1. To perform the clustering, a dataframe was created with the distinct locations and their respective coordinates. Both the SSE and Silhouette Score were used for the choice of the k parameter. The variations of the SSE were calculated based on the value of k∈[2,15] with the elbow method, which indicated k=6 as the best choice. Table 3.8 associates the Silhouette Score value to the SSE for some values of k. Based on these results, we decided to run the algorithm with k = 6. The result I is visually shown in Figure 3.2. After K-means assigned each location to its own cluster, we added the label to each record of the dataset. We can immediately see that within the clusters, the target variable is equally distributed and respects the proportion of the original dataset, with the exception of cluster 4: this groups the locations in the center of Australia, where it rains less in general (Tabella 3.7). furthermore, it is interesting to note that the grographical clusters do not reflect the political boundaries defined by the Region variable with the exception of cluster 1 (entirerly contains observations from Western Australia) and cluster2 (Northern Territory).

Cluster ID	RainTomorrow = Yes	RainTomorrow = No	Nr Records
0	23.26%	76.73%	50877
1	23.75%	76.24%	20706
2	23.51%	76.48%	4751
3	22.61%	77.38%	8972
4	7.45%	92.54%	7542
5	23.12%	76.87%	49345

Table 3.7 - Clusters distributions

K	SSE	Silhouette Score
5	621.45	0.644
6	461.98	0.666
7	313.15	0.687
8	304.20	0.555

Table 3.8 - SSE vs Silhouette Score



At this point, the same classification models used in paragraph 3.2 are applied to the obtained clusters, each of them trained with the best hyperparameters found from the *Cross Validation* with *numFold = 5.* The results obtained are shown in Table 3.8. Cluster 4 the most imbalanced, is the one with the worst results with all the metrics used, with the exception of Accuracy, which confirms that the latter is not reliable for classification on a dataset with a strongly imbalanced class.furthermore, the three classifiers provide similar results, making it difficult to assert which of the models best classifies the geographic clusters.

Model	Measure	Cluster0	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
Decision Tree	under ROC	0.69	0.75	0.82	0.77	0.67	0.71
	under PR	0.53	0.63	0.59	0.56	0.40	0.58
	Accuracy	0.82	0.86	0.84	0.82	0.93	0.83
	Recall	0.44	0.56	0.77	0.66	0.37	0.49
	Precision	0.65	0.75	0.63	0.62	0.55	0.69
	F-measure	0.53	0.64	0.69	0.64	0.44	0.58
Random Forest	under ROC	0.69	0.74	0.82	0.73	0.61	0.72
	under PR	0.56	0.67	0.54	0.63	0.29	0.60
	Accuracy	0.83	0.86	0.82	0.84	0.92	0.84
	Recall	0.43	0.53	0.82	0.51	0.26	0.49
	Precision	0.69	0.81	0.57	0.77	0.42	0.73
	F-measure	0.53	0.64	0.67	0.61	0.32	0.59
	under ROC	0.66	0.75	0.77	0.72	0.65	0.72

Logistic Regression	under PR	0.57	0.67	0.65	0.65	0.45	0.59
	Accuracy	0.82	0.86	0.86	0.85	0.94	0.84
	Recall	0.36	0.53	0.60	0.48	0.32	0.50
	Precision	0.73	0.80	0.76	0.80	0.64	0.71
	F-measure	0.48	0.64	0.67	0.60	0.43	0.58

Table 3.8 - Results

3.4 Classifications by region

The dataset was divided into 8 regions obtained in the Data Preparation. As was done for the two clustering developed previously, the data was put into vectors with the *VectorAssembler*, standardized with the *StandardScaler* and finally PCA with K=3 was used. Table 3.9 represents the distribution of the target variable RainTomorrow in the 8 clusters.

Cluster ID	RainTomorrow = Yes	RainTomorrow = No	Nr Records
Norfolk Island	30.08%	69.92%	2894
Victoria	23.10%	76.90%	27214
South Australia	19.52%	80.48%	11793
New South Wales	21.21%	78.79%	44084
Western Australia	22.17%	77.83%	16938
Tasmania	23.40%	76.60%	6128
Queensland	24.17%	75.83%	11906
Northern Territory	15.50%	84.50%	8289

For classification, the Decision Tree, Random Forest and Logistic Regression were used. *Grid Search* and *Cross Validation* with *numFolds* equal to 5 were applied to all three. The results obtained with the three classifiers are reported below.

Model	Measure	Norfolk Island	Victoria	New South Wales	Queensland	Northern Territory	South Australia	Western Australia	Tasmania
Decision Tree	under ROC	0.72	0.68	0.67	0.71	0.70	0.72	0.74	0.64
	under PR	0.57	0.56	0.55	0.63	0.54	0.50	0.58	0.43
	Accuracy	0.78	0.82	0.83	0.83	0.88	0.84	0.85	0.76
	Recall	0.57	0.41	0.39	0.47	0.45	0.53	0.56	0.40
	Precision	0.64	0.69	0.70	0.76	0.69	0.60	0.68	0.52
	F-measure	0.60	0.51	0.50	0.58	0.56	0.56	0.62	0.45
Random Forest	under ROC	0.70	0.68	0.68	0.75	0.75	0.69	0.72	0.63
	under PR	0.60	0.58	0.54	0.61	0.52	0.57	0.61	0.55
	Accuracy	0.78	0.82	0.84	0.84	0.88	0.86	0.85	0.80

	Recall	0.5	0.41	0.42	0.56	0.57	0.43	0.50	0.30
	Precision	0.70	0.72	0.67	0.71	0.62	0.73	0.73	0.71
	F-measure	0.58	0.52	0.52	0.63	0.59	0.54	0.60	0.42
Logistic Regression	under ROC	0.68	0.69	0.67	0.71	0.73	0.70	0.71	0.64
	under PR	0.58	0.57	0.54	0.59	0.60	0.56	0.62	0.50
	Accuracy	0.76	0.82	0.83	0.82	0.89	0.85	0.84	0.80
	Recall	0.47	0.45	0.38	0.50	0.49	0.44	0.47	0.36
	Precision	0.67	0.69	0.68	0.70	0.75	0.70	0.75	0.63
	F-measure	0.55	0.55	0.49	0.58	0.59	0.54	0.58	0.45

3.5 Comparison of various results

In order to compare the different classifications, it was necessary to make some considerations regarding the dataset and the metrics.

It is evident that accuracy is not the appropriate metric for comparison, since the dataset is clearlu biased towards the ‘No’class of the target variable, and therefore the accuracy level could be satisfactory even if the minority class was not classified at all (in our case, such a classification of the entire dataset would guarantee an accuracy of 77,6%).

Secondly, recall and precision were also discarded. Favoring a model with a higher recall would mean implicitly admitting that the classification of *RainTomorrow* = ‘Yes’ is morerelevant than *RainTomorrow* = ‘No’, since the metric in question is more tolerant of false positives. In our case, if a model always classified *RainTomorrow*with a positive value, the same model would obtain excellent accuracy. The opposite would be true if precision were preferred, since if a model always classified *RainTomorrorrow* with a positive value, the model always classified *RainTomorrow* with a negative value, *precision* would be high.establishing which class is more important would mean establishing whether rain is a positive phenomenon or not. Since such a judgment depends on the scope of use of the classifier, two more neutral metrics, namely the area under ROC curve (AUC) and l’F-measure were preferred.

In section 3.2, the main problems related to applying classifiers on clusters were synthesized: further imbalance of the target variable and reduction of available observatins. Regarding the clusters obtained with K-Means, the classification results were not satisfactory for any cluster except for cluster 5.

Geographical clustering and dividing the dataset by region turned out to be better for classification. Regarding geographical clusters, the Decision Tree was the classifier that guaranteed the best results. An average AUC between the clusters of 0.735 was obtained, while the average F-measure was 58,7%.

As for the classification by region, the Random Forest is the best classifier: average AUC between the clusters of 0.7 and average F-measure of 55%. In light of this, geographical clustering was cinsidered the most useful for classification purposes

However, the overall classification remains the best method in terms of metrics, especially if the PCA model with k=10 is taken into consideration. Withthis method, regardless of the classifier used, an AUC of 0.7 was obtained and an F-measure of 57,6%.

4 Precipitation level regression

RISK_MM was the continuous target variable of our analysis
The goal is to use two algorithms, Linear Regressor and RandomForest Regressor, to estimate the millimeters of rain the following day.

Categorical variables were eliminated and the features used were processed through *Standard Scaler*to obtain better results. The regression results were obtained by splitting the dataset (70% in the training set and 30% in the test set) and are not very sarisfactory for both Linear Regression and RandomForest Regressor, yielding a low R^2 score and a high RMSE, as we can see in Table 4.1.

Model	R^2	RMSE
Linear Regression	0.1849	7.3716
RandomForest Regression	0.2996	7.1175

table 4.1 R^2 score and RMSE

Later, we applied *Cross Validation* in order to find the best hyperparameters, and we observe a slight impovement in the quality of the Linear Regression with a decrease in the RMSE and an increase in the R^2 score, and the RandomForest Regressor, which experiences a decrease in the RMSE at the expense of the R^2 score which remains unchanged (Table 4.2).

Model	cross validation R^2	cross validation RMSE
Linear Regression	0.1949	7.4386
RandomForest Regression	0.2960	7.0313

Table 4.2 R^2 score and RMSE with CV

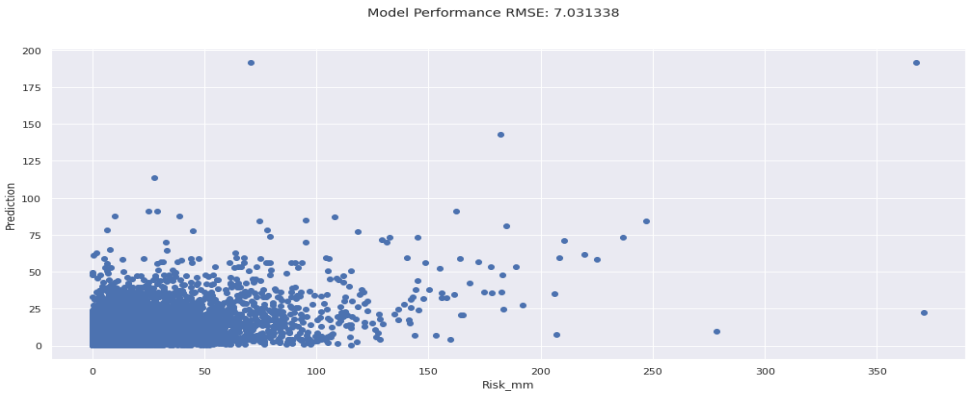


Figure 4.1 Performance of the RandomForestRegressor

Finally, we utilized the ***featureImportance*** function of the RandomForest regression model, which estrablishes a percentage on how influential each feature is on the model's predictions.
To isolate the best-performing model in our parameter grid, we used ***bestModel*** (Figure 4.2)

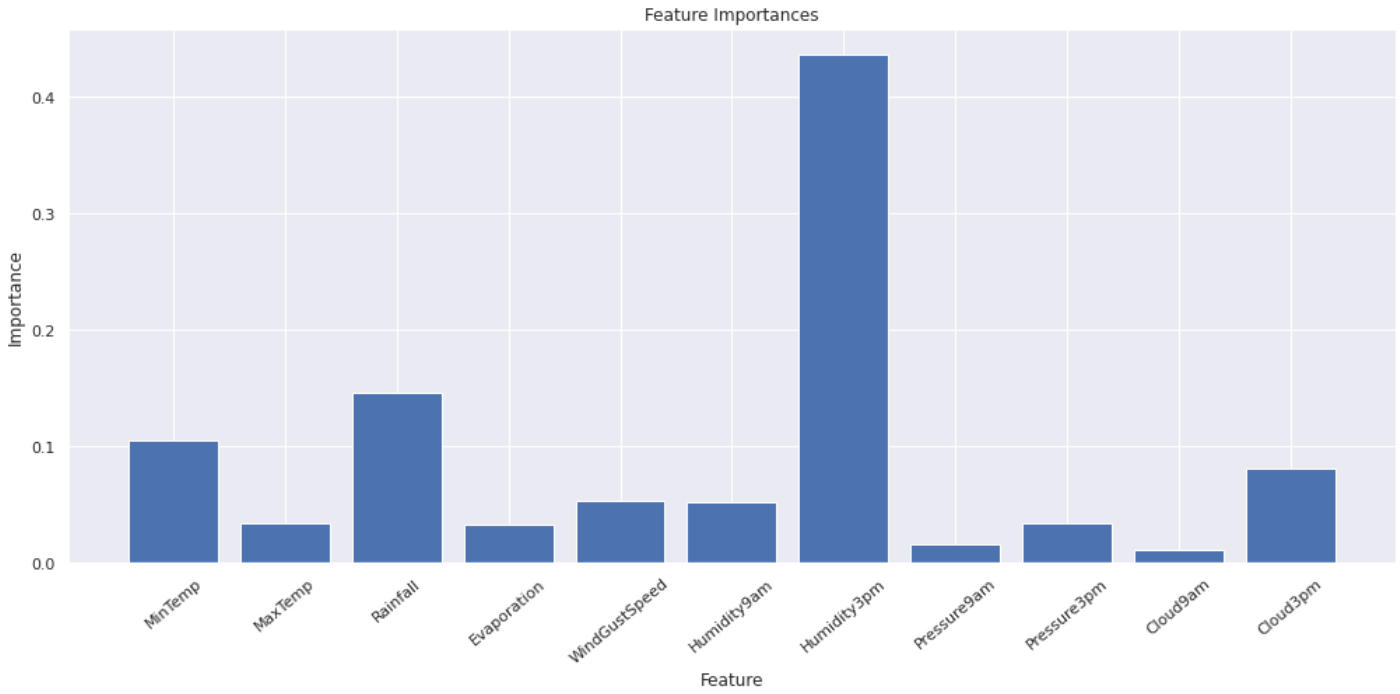


Figure 4.2 - Feature importance

It seems that Humidity3pm, Cloud3pm, Rainfall and MinTemp are the most important predictors for Risk_mm.