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

Analisi del dataset “Australia, Rain Tomorrow”

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

Anno accademico 2021/2022

Indice

1 Data Understanding

1.1 Introduzione

Il dataset utilizzato per la realizzazione di questo progetto, chiamato "*Australia, Rain Tomorrow*" , è reperibile sulla piattaforma Kaggle al link (https://www.kaggle.com/filhypedeeplearning/australia-rain-tomorrow). I dati contenuti al suo interno rappresentano delle osservazioni meteorologiche, raccolte dal Australian Bureau of Meteorology (BOM), di diverse *location* dell'Australia per ogni giorno dal Dicembre 2008 a Giugno 2017. Le osservazioni meteorologiche sono date da una lista di elementi meteorologici rilevati durante le giornate, come ad esempio *MinTemp*, *MaxTemp*. In particolare, gli attributi binari *RainToday* e *RainTomorrow*, indicano rispettivamente la presenza (***Yes***) o meno (***No***) di precipitazioni durante la data considerata e per il giorno successivo ad essa. Gli obiettivi del progetto sono:

- realizzazione di uno studio non supervisionato, tramite un algoritmo di clustering, dell'intero dataset, al fine di individuare le caratteristiche comuni delle osservazioni meteorologiche.
- suddivisione del dataset sia in regioni politiche sia per coordinate geografiche.
- previsione della variabile target *RainTomorrow*, sia nel dataset originale che nei cluster ottenuti, attraverso la realizzazione di diversi modelli di classificazione.
- previsione della variabile *Risk_MM*, che indica il livello delle precipitazioni del giorno successivo in millimetri, attraverso alcuni modelli di regressione.

1.2 Distribuzioni e analisi del dataset

Il dataset è composto da 142193 records e da 24 features contenenti informazioni su alcuni elementi meteorologici osservati in diversi momenti delle giornate. Le features sono brevemente descritte in Tabella 1.1 e in Tabella 1.2.

Nome	Descrizione	Tipologia	Valori
Date	Data dell'osservazione	Attributo Temporale	{"01/12/2008",...}
Location	Luogo dell'osservazione	Categorico	{Sydney, Canberra…}
WindGustDir	Direzione della raffica	Categorico	{WSW, WNW, W,..., E}
WindDir9am	Direzione media del vento 10 min prima delle 9 am	Categorico	{WSW, WNW, W,..., E}
WindDir3pm	Direzione del vento alle 3 pm	Categorico	{WSW, WNW, W,..., E}
RainToday	Presenza o meno di pioggia nella data corrente	Binario	{Yes, No}
RainTomorrow	Presenza o meno di pioggia nella data successiva a quella corrente ***Variabile Target***	Binario	{Yes, No}

Tabella 1.1: Descrizione attributi categorici e binari

Nome	Descrizione	Tipologia	Range(Min,Max)	Media
MinTemp	Temperatura min fino alle 9 am*	Continuo	{-8.5, 33.9 }	12.18
MaxTemp	Temperatura max fino alle 9 am*	Continuo	{-4.8, 48.1}	23.23
Rainfall	Precipitazioni fino alle 9 am*	Continuo	{0.0, 371.0}	2.35
Evaporation	Evaporazione in pan fino alle 9 am*	Continuo	{0.0, 145.0}	5.47
Sunshine	Livello di luminosità fino a mezzanotte*	Continuo	{0, 14.5}	7.62
WindGustSpeed	Velocità in kmh del vento più forte fino a mezzanotte*	Continuo	{6.0, 135.0}	39.98
WindSpeed9am	Velocità media in km h 10 min prima delle 9 am	Continuo	{0.0, 130.0}	14.0
WindSpeed3pm	Velocità media in km h 10 min prima delle 3 pm	Continuo	{0.0, 87.0}	18.4
Humidity9am	Umidità relativa (in percentuale) alle 9 am	Continuo	{0.0, 100.0}	68.84
Humidity3pm	Umidità relativa (in percentuale) alle 3 pm	Continuo	{0.0, 100.0}	51.48
Pressure9am	Pressione atmosferica (hpa) ridotta al livello medio del mare alle 9 am	Continuo	{980.5, 1041.0}	1017.65
Pressure3pm	Pressione atmosferica (hpa) ridotta al livello medio del mare alle 3 pm	Continuo	{977.1, 1039.6}	1015.25
Cloud9am	Frazione di cielo oscurata dalle nuvole alle 9 am (in *oktas*). 0 indica cielo completamente scoperto e 9 completamente ricoperto di nuvole.	Discreto	{0.0, 9.0}	4.44
Cloud3pm	Frazione di cielo oscurata dalle nuvole alle 3 pm	Discreto	{0.0, 9.0}	4.50
Temp9am	Temperatura (in Celsius) alle 9 am	Continuo	{-7.2, 40.2}	16.99
Temp3pm	Temperatura (in Celsius) alle 3 pm	Continuo	{-5.4, 46.7}	21.68
RISK_MM	Livello di pioggia in mm registrato durante la giornata successiva. ***Variabile continua target***	Continuo	{0.0, 371.0 }	2.36

Tabella 1.2: Descrizione attributi numerici

* le seguenti misure sono tutte da intendersi come calcolate nelle 24 ore precedenti.

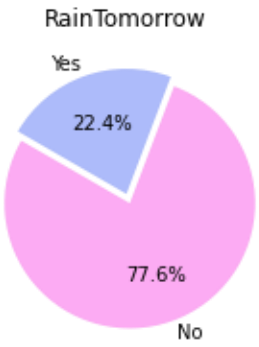

Figura 1.1: Distribuzione Variabile Target

Il dataset si presenta fortemente sbilanciato riguardo la variabile target: solo nel 22.41% dei casi viene registrata pioggia per la giornata successiva (*RainTomorrow* = 'Yes'), come è possibile osservare anche nella Figura 1.2. Anche la variabile *RainToday* ha una distribuzione simile: nel 22.12% delle osservazioni *RainToday* = "Yes", mentre nel 76.89% *RainToday* = "No". Da ulteriori analisi, constatiamo che se nella giornata attuale piove allora ci sono circa il 46.40% di probabilità che piova anche nella giornata successiva.

Inoltre, osservando le distribuzioni delle variabili *Humidity9am* e *Humidity3pm* in Figura 1.3, notiamo che quando *RainTomorrow* = "Yes" l'umidità registrata durante il pomeriggio precedente assume valori più alti, mentre ciò non accade per quella registrata la mattina prima. Le variabili che registrano informazioni riguardanti le temperature (*Temp3pm*, *Temp9am*, *MaxTemp* e *MinTemp*) seguono tutte una distribuzione a campana (Figura 1.2). Per quanto riguarda, invece, tutte le variabili che danno informazioni sul vento (*WindGustDir*, *WindDir9am*, *WindDir3pm*), la direzione *"WSW"* è quella più frequentemente registrata. Infine, per quanto riguarda la variabile target continua *RISK_MM*, quest'ultima assume i suoi valori più alti quando piove per almeno due giorni consecutivi.

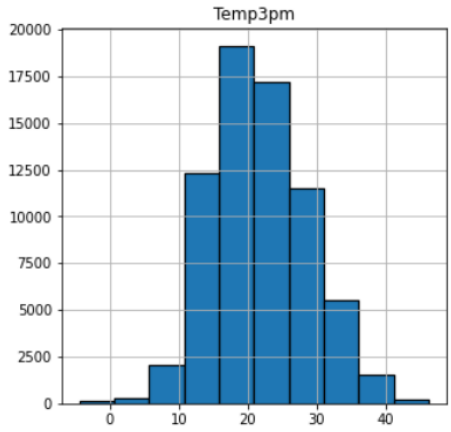

Figura 1.2: Distribuzione Temp3pm

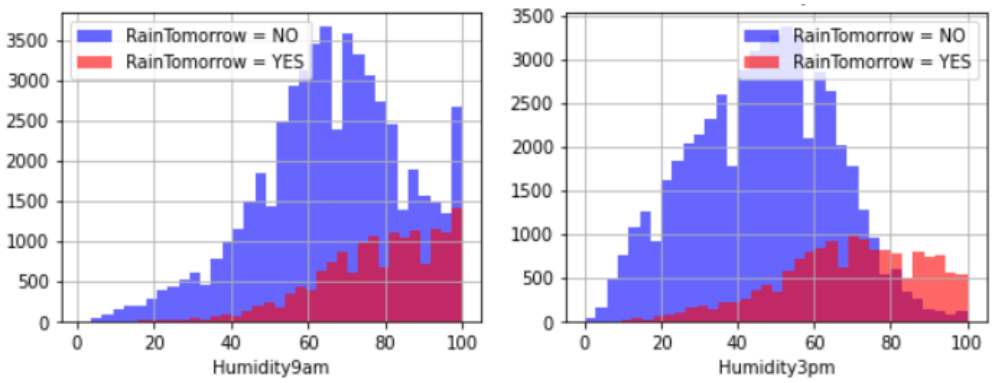

Figura 1.3 : Distribuzione Humidity9am e Humidity3pm sulla variabile Target

1.3 Analisi dell'attributo *Location*

L'attributo *Location* è stato fin da subito oggetto di analisi poiché la sua distribuzione è di fondamentale importanza per valutare la fattibilità di alcuni degli obiettivi del progetto. Esso presenta 49 diversi luoghi. Al fine di comprendere dove si trovino questi luoghi, abbiamo deciso di ricavare le coordinate di essi, poiché queste non venivano fornite nel dataset iniziale. A tal proposito, è stato necessario pulire i valori dell'attributo *location,* in quanto alcuni di questi erano inseriti senza spaziatura e per questo risultava impossibile ricavare le coordinate (ad esempio, il valore "BadgerysCreek" invece di Badgerys Creek). Una volta ricavate le coordinate, è possibile visualizzarle sulla mappa (Figura 1.4, destra), dove possiamo subito notare che i diversi luoghi appaiono nel dataset con una frequenza simile (dimensione delle bolle), a parte alcune come *Nihil, Uluru e Katherine* (Figura 1.4, sinistra). In particolar modo, si noti che il plot in Figura 1.4 (sinistra) è ricavato prendendo in considerazione un sample del dataset, ma che ogni diversa *location* ha circa 3K record.

Inoltre, possiamo notare come la maggior parte delle location si trova sulla costa sudorientale, mentre nella parte centrale vi sono poche osservazioni, probabilmente dovuto alla concentrazione demografica dell'Australia.

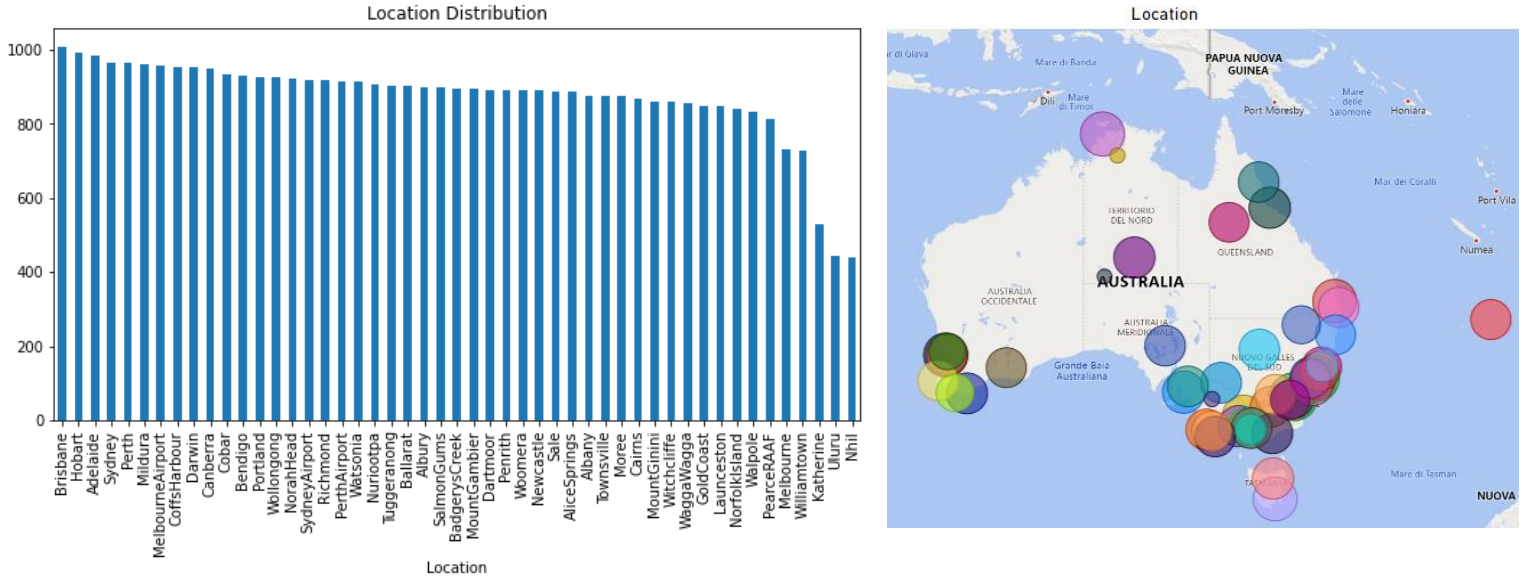

Figura 1.4 : Attributo Location - Bar Chart e Mappa

1.4 Correlazioni

Per ottenere le correlazioni degli attributi sono stati usati due indici di correlazione. Attraverso l'indice di Pearson è stata calcolata la correlazione tra le features del dataset, rappresentata dalla heatmap in Figura 1.2.

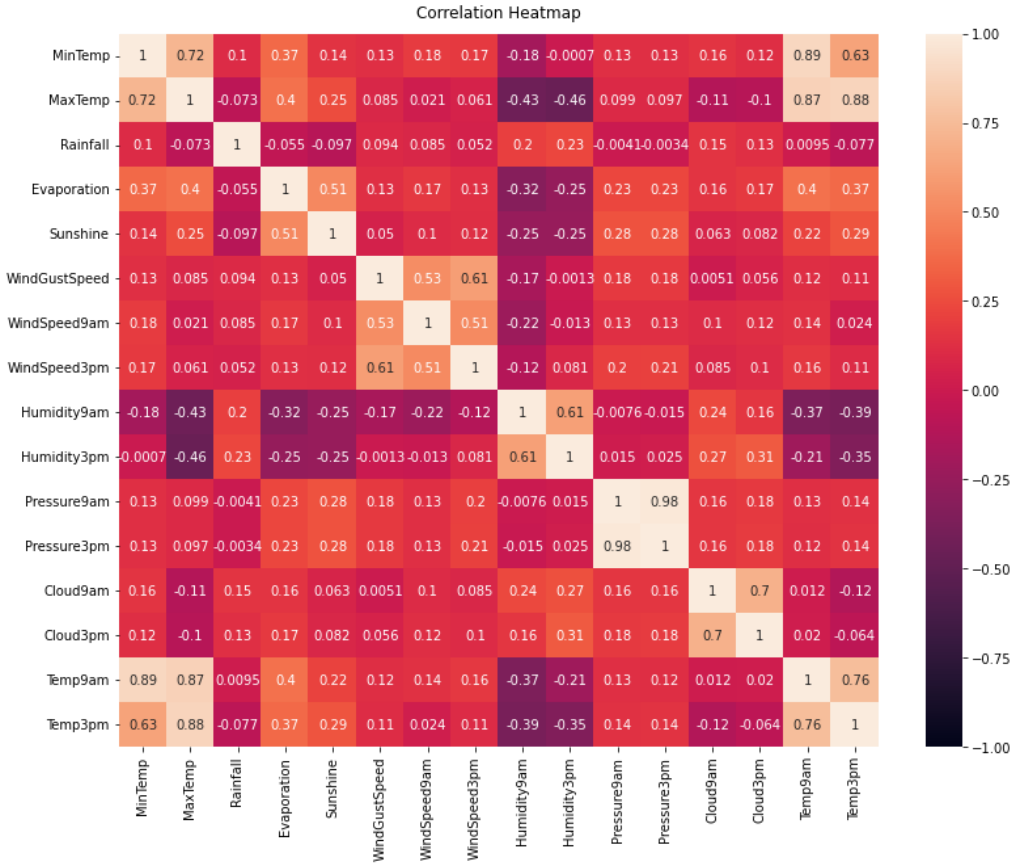

Figura 1.5: Correlazione tra attributi

Possiamo notare come gli attributi che presentano maggiore correlazione sono quelli che indicano le temperature e gli attributi che misurano la stessa cosa ma in diverse fasce orarie (es. *Pressure9am* e *Pressure3pm*). Invece, il Point-Biserial Correlation Coefficient[1] è stato usato per vedere la correlazione dei vari attributi numerici con l'attributo binario *RainTomorrow*. Usando questo coefficiente, possiamo notare come gli attributi più correlati alla variabile target sono: *Rainfall*, *Sunshine*, *WindGustSpeed*, *Humidity9am*, *Humidity3pm*, *Pressure9am*, *Pressure3pm*, *Cloud9am*, *Cloud3pm*.

2 Data Preparazioni

2.1 Creazione degli attributi *Month*, *Season* e *Region*

Abbiamo deciso di ricavare gli attributi *Month* (con valori '01', '02', '03', …, '12') e *Season* ('Winter', 'Fall','Spring','Summer') dalla variabile *Date,* attraverso l'uso delle *regular expressions*, in quanto abbiamo ritenuto interessante andare ad analizzare il comportamento della pioggia durante i vari mesi e le stagioni, anche in relazione al successivo task di clustering. I mesi con la percentuale più alta di *RainTomorrow* = 'Yes' sono risultati appunto quelli estivi, come potevamo aspettarci. È stata integrata anche la variabile Region, che rappresenta la regione politica di ogni *location* presente sul dataset, tramite un ulteriore csv (*cities_australia.csv*). Anche questo attributo è stato aggiunto per il task di clustering.

2.2 Missing values: sostituzione e confronto tra le metodologie

Il dataset presenta diversi valori mancanti. Gli attributi con più valori mancanti risultano essere *Sunshine* con il 48% delle osservazioni mancanti, *Evaporation* con il 43%, *Cloud3pm* con il 40% e *Cloud9am* con il 38%. In Figura 2.1 è riportato il totale di valori mancanti per ogni attributo.
Dato il numero elevato di *missing values*, si è deciso di escludere l'eliminazione dei record contenenti valori mancanti, e di procedere invece con la sostituzione degli stessi. Considerata la notevole influenza di questa fase sul resto dell'analisi, si è deciso di sondare diversi approcci.

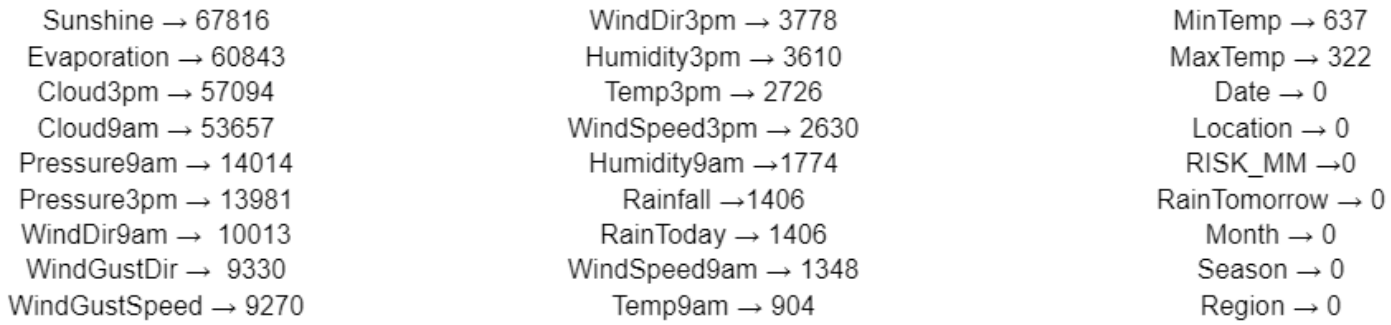

Figura 2.1: Missing Values

1. Il primo approccio prevede il calcolo della media tramite l'utilizzo del raggruppamento per attributi. I primi quattro attributi per numero di valori mancanti sono stati raggruppati in primo luogo per *Location*:

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

i *missing values* risultano distribuiti omogeneamente tra le città, e le città con più valori mancanti per gli attributi in questione ne presentano circa 3000. Successivamente, i valori mancanti sono stati raggruppati per *RainToday* e *RainTomorrow*: essendo attributi binari, si è potuto calcolare facilmente la media per entrambe le classi degli attributi. Di seguito è riportato un esempio per *Cloud9am*.

RainToday	avg(Cloud9am)
No	3.939796797480764
Yes	6.018474088291747
null	5.858288770053476

RainTomorrow	avg(Cloud9am)
No	3.9322820037105752
Yes	6.09999030161963

Il raggruppamento per *Location* ha dimostrato da subito di non essere adatto alla sostituzione dei *missing values*, poiché la tabella di raggruppamento presentava delle osservazioni mancanti. Si è quindi deciso di raggruppare per Month e di raggruppare ulteriormente per l'attributo con la correlazione più elevata. Di seguito viene mostrato un esempio di raggruppamento di *Evaporation* per *Sunshine* e per *Month*.

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

Tuttavia, si è preferito non utilizzare questo metodo di sostituzione, poiché risulta difficile giustificare la sostituzione del valore mancante di una singola giornata con la media del raggruppamento di *Month*, considerate le molteplici condizioni meteorologiche che possono verificarsi durante un mese. In altre parole, un mese è un lasso di tempo troppo ampio per desumere il valore di un attributo relativo ad un solo giorno.

2. Come secondo approccio è stata valutata la possibilità di sostituire i *missing values* tramite la regressione. Anche in questo caso, però, si è deciso di non procedere alla sostituzione, poiché è presente un'elevata correlazione tra i valori mancanti degli attributi per cui attuare la sostituzione (le correlazioni tra *missing values* sono mostrate in Figura 2.2). Il risultato è che per la regressione di un valore non è possibile scegliere la *feature* più utile per la stessa regressione.

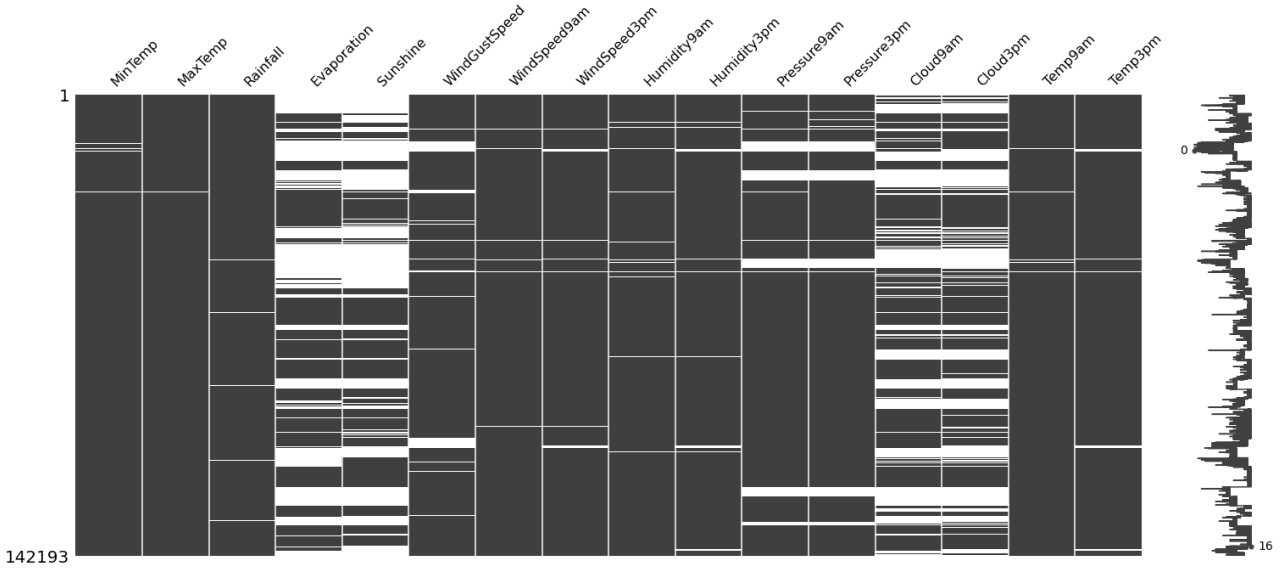

Figura 2.2 - Correlazione tra Missing Values

3. Il metodo scelto per la sostituzione dei *missing values* prevede l'utilizzo della mediana, poiché più resistente agli outliers. Questo approccio è stato utilizzato per i primi 6 attributi per numero di valori mancanti, ossia *Evaporation*, *Sunshine*, *Pressure9am*, *Pressure3pm*, *Cloud9am* e *Cloud3pm*. Per ottenere dei valori più accurati, si è deciso di aggiungere provvisoriamente al dataset 4 variabili binarie:
   - *it-will-rain*: *RainToday* = 0 e *RainTomorrow* = 1 ("oggi non piove, domani pioverà")
   - *no-rain*: *RainToday* = 0 e *RainTomorrow* = 0 ("oggi non piove e domani non pioverà")
   - *it-rain*: *RainToday* = 1 e *RainTomorrow* = 1 ("pioverà sia oggi che domani")
   - *end-rain*: *RainToday* = 1 e *RainTomorrow* = 0 ("oggi piove, ma domani non pioverà")

   Dopo aver ottenuto questi quattro fenomeni, sono state verificate le differenze presenti nelle loro distribuzioni attraverso un test statistico sulla mediana. È stato scoperto che per *Evaporation* il processo di generazione dei valori per *it-rain* e *end-rain* è diverso: questo implica che la mediana da assegnare è diversa, poiché i fenomeni sono differenti tra loro. Per le altre variabili del dataset i valori mancanti sono stati sostituiti con la mediana complessiva, dal momento che per loro i suddetti quattro attributi binari non seguono distribuzioni diverse.

Successivamente si è voluto confrontare il metodo delle mediane con la regressione. È stato scelto un attributo per cui effettuare la regressione, *Sunshine*, ed è stato selezionato un campione di osservazioni che non presentano *missing values* (altrimenti non sarebbe stato possibile effettuare la regressione): per questa ragione sono state escluse le variabili *Evaporation*, *Cloud9am*, *Cloud3pm*, *Pressure9am*, *Pressure3pm*. Con il dataframe ottenuto, composto da 84525 record, sono state ottenute le predizioni per i valori di *Sunshine*. La regressione ha ottenuto un R^2 pari a 0,21 e un RMSE pari a 2,5. Di seguito è riportato un esempio dei risultati ottenuti.

```
+--------------------+--------+------------------+------------------+
|            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|
+--------------------+--------+------------------+------------------+
```

È bene specificare che la colonna *Sunshine* è composta sia da valori originali che da valori ottenuti con la sostituzione. Sono state poi osservate le statistiche relative alla colonna *difference*:

```
+-------+--------------------+
|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|
+-------+--------------------+
```

I risultati mostrano delle differenze significative, considerato che il range di attributi per la variabile *Sunshine* è [0; 14,5] e che lo scarto medio è di circa 2 ore di sole. Allo stesso tempo non è possibile stabilire quale dei due metodi sia effettivamente il migliore, non avendo modo di reperire i valori corrispondenti ai *missing values*. Inoltre, è bene ricordare che le *feature* utilizzate per la regressione non includono gli attributi con la

correlazione maggiore con la variabile da predire, ossia gli attributi che generalmente garantiscono una predizione migliore. Per queste ragioni il metodo di sostituzione utilizzato è stato ritenuto valido.

3 Clustering e Classificazione

Come accennato nell'introduzione, il task di classificazione è stato approcciato con diverse metodologie. Innanzitutto, è stata fatta una classificazione standard sull'intero dataset. Successivamente, il dataset è stato suddiviso in clusters nei seguenti modi:

1. Tramite il K-means usando alcune variabili numeriche.
2. Tramite il K-means usando solo le coordinate geografiche.
3. Suddividendo a priori il dataset in regioni politiche in base alla variabile *Region.*

Gli stessi algoritmi usati per la classificazione standard sono stati applicati ai diversi clusters ottenuti e i risultati confrontati tra di loro.

3.1 Classificazione standard della variabile RainTomorrow

Per il compito di classificazione della variabile *RainTomorrow* è stato usato l'intero dataset, senza distinzione per quanto riguarda i record. Invece, per quanto riguarda la scelta degli attributi, questi sono stati scelti sia osservando le correlazioni con l'attributo *RainTomorrow* (in particolare sono stati selezionati tutti gli attributi con valore assoluto maggiore di 0.2), sia osservando le correlazioni fra gli stessi attributi e scartando tutti quelli che presentavano un'elevata correlazione: in questo modo è stata evitata la ridondanza fra gli stessi. Per quanto riguarda le scelte fatte per il compito di classificazione, sono state eseguite separatamente tre diverse tecniche di *scaling* dei dati. In primo luogo, sono stati utilizzati il *MinMax Scaler* e lo *Standard Scaler*. Successivamente abbiamo utilizzato la PCA con k = 3, sia come *scaling* che come tecnica di *feature reduction*. Per quanto riguarda i modelli usati, per tutte e tre le tecniche di preprocessing abbiamo applicato i seguenti algoritmi di machine learning: Logistic regression, Decision tree classification, Random forest classification, Linear SVM classification.
Come ultima parte della classificazione generale, sono stati eseguiti gli stessi quattro algoritmi di classificazione. Per la scelta degli iperparametri è stata eseguita una *grid-search* con una *cross-validation*. La tecnica usata per la CV è stata la k-fold con k = 5. Per la logistic regression gli iperparametri per cui è stato effettuato il tuning sono il numero massimo di iterazioni e l'*elasticNet* per la decisione della penalità del modello. Per il decision tree è stato effettuato il tuning per il numero massimo di intervalli e la profondità dell'albero; per il Random forest, il numero di alberi e la profondità massima degli alberi, mentre per il linear SVM il numero massimo interazioni e il *fitIntercept.*Successivamente, abbiamo deciso di provare a classificare tenendo tutte le features continue del dataset, ad accezione di RISK_MM, usando la PCA con k = 10, come metodo di *feature reduction*. Nelle Tabelle 3.1, 3.2, 3.3, 3.4 vengono mostrati i risultati per le metriche scelte.

(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

Tabella 3.1 - Risultati StandarScaler

(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 - Risultati MinMax

Si può notare che usando k=10 i risultati migliorano, sia rispetto all'uso di pca con k=3 che con l'uso delle altre tecniche di scaling.

(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 - Risultati 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 - Risultati pca, k=10

3.2 K-Means e classificazioni intra-cluster

La definizione stessa di cluster potrebbe portare a pensare che la classificazione svolta all'interno di un cluster debba necessariamente essere più accurata della classificazione globale. In realtà, la suddivisione del dataset porta con sé due problematiche: lo sbilanciamento della variabile target all'interno dei cluster e la riduzione dei record con cui allenare il classificatore. Dopo aver sperimentato K-Means sia nella sua versione standard sia nella declinazione Bisecting K-Means, si è deciso di optare per la prima, poiché garantiva un valore di SSE significativamente più basso. Come per la classificazione generale, gli attributi utilizzati sono *Rainfall*, *Sunshine*, *WingGustSpeed*, *Humidity3pm*, *Pressure9am*, *Cloud3pm* e *RainToday*. È stato necessario escludere *RainToday* e le due variabili target *RainTomorrow* e *RISK_MM*, poiché avrebbero reso poco significativa l'analisi. Successivamente i valori contenuti nei vettori sono stati processati con lo *StandardScaler*.

Per la scelta del K sono stati presi in considerazione sia il *Silhouette Score* che l'SSE. I risultati sono descritti dalla Figura 3.1. Si è deciso di utilizzare k=6.

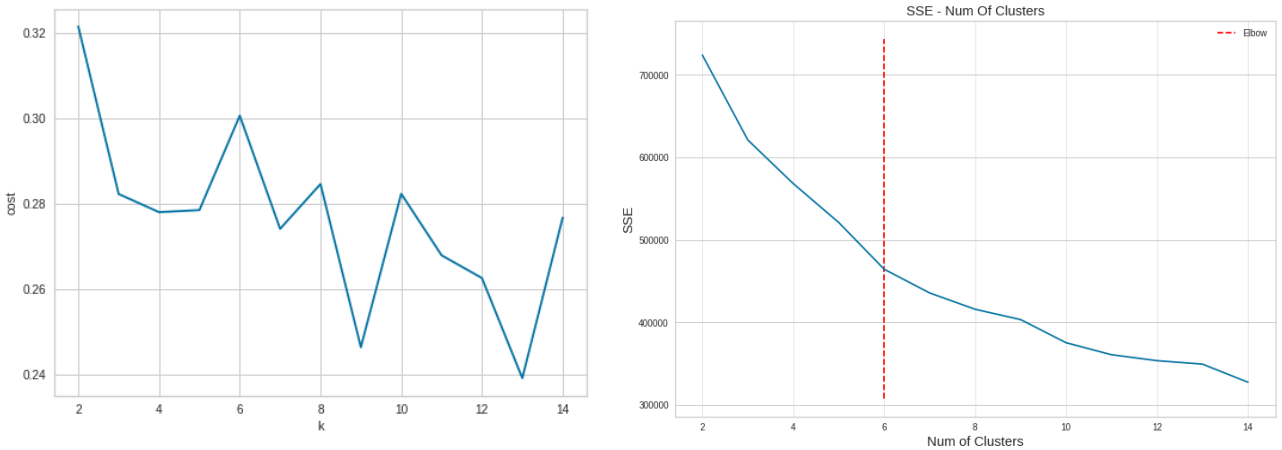

Figura 3.1 - SSE & Silhouette Score

Nella tabella 3.5 , per ognuno dei 6 cluster, è riportata la distribuzione della variabile target.

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

Tabella 3.5 - Distribuzioni Clusters

Come è possibile osservare, la variabile target *RainTomorrow* è distribuita in maniera significativamente diversa in ogni cluster, e la classe 1 di RainTomorrow rappresenta la classe dominante solamente per il cluster 5, con il 74% delle occorrenze. I cluster 0 e 1 sono i più sbilanciati: l'attributo 1 di *RainTomorrow* rappresenta rispettivamente il 4,4% e il 5,4%.
Sono state osservate anche le distribuzioni degli altri attributi all'interno dei cluster. Per l'attributo *RainToday* è valido quanto detto in precedenza per *RainTomorrow*: i cluster 0 e 1 sono nettamente sbilanciati verso lo 0 (ossia verso il 'No'), mentre il cluster 5 è il più bilanciato. La differenza principale è che *RainToday* = 1 non è la classe maggioritaria in nessun cluster. In altri termini, i cluster 0 e 1 sono i cluster per cui il fenomeno della pioggia è più raro, mentre il cluster 5 è quello in cui si verifica più spesso. Successivamente si è cercato di capire quali cluster potessero essere identificati con una stagione dell'anno. In realtà, considerato che l'attributo *Season* non è bilanciato all'interno del dataset, non è stato possibile ottenere un risultato di questo tipo. È stato possibile osservare che *Fall* è la stagione dominante in tutti i cluster ad eccezione del cluster 2, con percentuale massima del 56% nel cluster 0 e percentuale minima del 25% nel cluster 2. Nel cluster 2 la stagione prevalente è *Summer*, con il 40% delle osservazioni. *Spring*, invece, ha il suo valore massimo nel cluster 1, con il 32% delle osservazioni. Il maggior numero di osservazioni di *Winter* si osserva nel cluster 1, con il 10%.

Cercando di riassumere le informazioni ottenute sui cluster:

- nel Cluster 0, composto da 24396 record, il fenomeno della pioggia è molto raro e il 56% dei record sono registrati in autunno;
- nel Cluster 1, composto da 32261 record, il fenomeno della pioggia è molto raro e i record sono distribuiti omogeneamente tra autunno (33%), primavera (32%) ed estate (30%);
- il Cluster 2, composto da 18024 record, è l'unico in cui *Summer* è la stagione dominante (40%);
- i Cluster 3 e 4 sono composti rispettivamente da 29699 e 13428 record, e non presentano elementi particolarmente caratterizzanti;
- il Cluster 5, composto da 12044, oltre ad essere il cluster più piccolo è quello in cui la pioggia è più frequente.

Per la classificazione, in aggiunta alla standardizzazione già citata, si è deciso di svolgere la PCA con K=3, dati i buoni risultati ottenuti nella prima classificazione. Sono stati utilizzati tre classificatori: Decision Tree, Random Forest e Logistic Regression. Per ognuno dei tre sono state effettuate *Grid Search* e *Cross Validation* con *numFolds* pari a 5. In Tabella 3.6 sono riportate le metriche di valutazione delle classificazioni effettuate.

Modello	Misura	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

Tabella 3.6 - Risultati

I risultati sono in linea con quanto detto durante l'analisi dei cluster: i cluster 0 e 1, essendo i più sbilanciati, presentano i risultati peggiori. Il cluster 5, ossia il cluster più bilanciato nonché l'unico con la prevalenza di 1 per *RainTomorrow*, risulta essere il più facile da classificare.

3.3 Clustering geografico e classificazioni intra-cluster

Il dataset è stato suddiviso in clusters geografici, utilizzando l'algoritmo di K-Means con le variabili longitudine e latitudine, ottenute come spiegate nel Paragrafo 1.3 . Per effettuare il clustering, è stato creato un dataframe con le distinte location e le rispettive coordinate. Anche in questo caso sono stati usati sia l'SSE che il Silhouette Score per la scelta del parametro k. Sono state calcolate le variazioni dell'SSE in base al valore di $k \in [2,15]$ con il metodo dell'Elbow, il quale ha indicato k=6 come migliore scelta. La Tabella 3.8 associa il valore di Silhouette Score a quello di SSE per alcuni valori di k. In base a questi risultati, abbiamo deciso di eseguire l'algoritmo con k=6. Il risultato è mostrato visivamente in Figura 3.2. Dopo che K-means ha assegnato ad ogni location il proprio cluster, aggiungiamo l'etichetta ad ogni record del dataset. Possiamo subito notare che all'interno dei clusters la variabile target è equamente distribuita e rispetta la proporzione del dataset originale, ad eccezione del cluster 4: questo, infatti, raggruppa le location al centro dell'Australia, dove in generale piove di meno (Tabella 3.7). Inoltre, è interessante notare come i cluster geografici non

rispecchiano i confini politici definiti dalla variabile Region, fatta eccezione per il cluster 1 (contiene interamente osservazioni del Western Australia) e del 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

Tabella 3.7 - Distribuzioni Clusters

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

Tabella 3.8 - SSE vs Silhouette Score

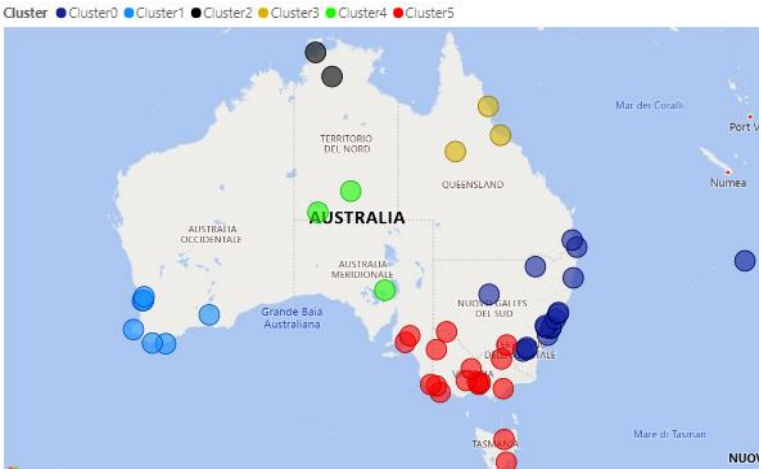

Figura 3.2 - Clusters su mappa

A questo punto sui clusters ottenuti vengono applicati gli stessi modelli di classificazione precedentemente usati nel paragrafo 3.2, ciascuno di essi allenati con i migliori iperparametri trovati dalla *Cross Validation* con *numFold = 5.* Nella Tabella 3.9 sono illustrati i risultati ottenuti. Il Cluster4, il più sbilanciato, è quello che ha risultati peggiori con tutti le metriche usate, fatta eccezione dell'accuracy, il che va a confermare che quest'ultima non è affidabile per la classificazione su un dataset con classe fortemente sbilanciata. Inoltre, i tre classificatori forniscono risultati simili e per questo è difficile affermare quale dei modelli classifica al meglio i cluster geografici.

Modello	Misura	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
Logistic Regression	under ROC	0.66	0.75	0.77	0.72	0.65	0.72
	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

Tabella 3.9 - Risultati

3.4 Classificazioni per regione

Il dataset è stato diviso nelle otto regioni ottenute nella Data Preparation. Come è stato fatto per i due clustering sviluppati in precedenza, i dati sono stati inseriti in vettori con il *VectorAssembler*, sono stati standardizzati con lo *StandardScaler* ed infine è stata utilizzata la PCA con K=3. In Tabella 3.10 è rappresentata la distribuzione della variabile target RainTomorrow negli 8 cluster. Per la classificazione sono stati utilizzati i classificatori Decision Tree, Random Forest e Logistic Regression. Per tutti e tre sono state applicate *Grid Search* e *Cross Validation* con *numFolds* pari a 5. In Tabella 3.11 sono riportati i risultati ottenuti con i tre classificatori.

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

Tabella 3.10 - Distribuzioni Clusters

Modello	Misura	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

Tabella 3.11 - Risultati

3.5 Confronto tra i vari risultati

Per poter effettuare un confronto tra le diverse classificazioni è stato necessario fare alcune considerazioni riguardo il dataset e le metriche.

È evidente che l'accuracy non sia la metrica adatta al confronto, dal momento che il dataset risulta nettamente sbilanciato verso la classe 'No' della variabile target, e che quindi il livello di accuracy potrebbe essere soddisfacente anche nel caso in cui la classe minoritaria non venisse classificata affatto (nel nostro caso una classificazione del genere sull'intero dataset garantirebbe un'accuracy del 77,6%).

In secondo luogo, sono state scartate anche recall e precision. Favorire un modello con una recall più alta vorrebbe dire ammettere implicitamente che la classificazione di *RainTomorrow* = 'Yes' sia più rilevante rispetto a *RainTomorrow* = 'No', poiché la metrica in questione risulta essere più tollerante rispetto ai falsi positivi. Nel nostro caso, qualora un modello classificasse sempre *RainTomorrow* con il valore positivo, lo stesso modello otterrebbe un'ottima accuracy. Il discorso opposto sarebbe vero se fosse preferita la precision, poiché qualora un modello classificasse sempre *RainTomorrorrow* con il valore negativo, la precision sarebbe elevata. Stabilire quale sia la classe più importante vorrebbe dire stabilire se la pioggia sia o meno un fenomeno positivo. Dal momento che un giudizio simile dipende dall'ambito di utilizzo del classificatore, si è preferito utilizzare due metriche più neutrali, ossia l'area under ROC curve (AUC) e l'F-measure.

Nel paragrafo 3.2 sono stati sintetizzati i principali problemi legati all'applicazione dei classificatori sui cluster: ulteriore sbilanciamento della variabile target e riduzione delle osservazioni disponibili. Per quanto riguarda i cluster ottenuti con K-Means, i risultati della classificazione non sono stati soddisfacenti per nessun cluster ad eccezione del cluster 5.
Il clustering geografico e la divisione del dataset per regione si sono rivelati migliori per effettuare la classificazione. Per quanto riguarda il clustering geografico, il Decision Tree è il classificatore che ha garantito i risultati migliori. È stata ottenuta una AUC media tra i cluster di 0.735, mentre la media della F-measure si attesta al 58,7%.
Per quanto riguarda invece la classificazione per regione, è il Random Forest il miglior classificatore: AUC media tra i cluster di 0.7 e F-measure media del 55%. Alla luce di ciò, il clustering geografico è stato ritenuto il più utile ai fini della classificazione.
Nonostante ciò, la classificazione generale rimane il metodo migliore in termini di metriche, specialmente se si prende in considerazione il modello PCA con k=10. Con questo metodo, a prescindere dal classificatore utilizzato, è stata ottenuta una AUC pari a 0.7 ed una F-measure del 57,6%.

4 Regressione del livello di precipitazioni

RISK_MM è stata la variabile target continua della nostra analisi. L'obiettivo è quello di utilizzare due algoritmi, Linear Regressor e RandomForest Regressor, per stimare i millimetri di pioggia nel giorno successivo.
Sono state eliminate le variabili categoriche e processate le features utilizzate attraverso lo *Standard Scaler* in modo da ottenere risultati migliori. I risultati della regressione sono stati ottenuti splittando il dataset (70% nel training set e 30% nel test set) e risultano essere molto soddisfacenti per la Linear Regression, restituendo un alto R^2 score e un basso RMSE, al contrario del RandomForest Regressor che ha restituito un R^2 Score abbastanza basso e un alto RMSE, come possiamo vedere in Tabella 4.1. Successivamente, abbiamo applicato la *Cross Validation* al fine di trovare i migliori iperparametri, e risulta un evidente miglioramento della qualità del modello di Linear Regression con un abbassamento dell'errore quadratico medio (RMSE) e un peggioramento delle performance per il RandomForest Regressor, il quale subisce un abbassamento dello score R^2 e un innalzamento dell'errore quadratico medio (Tabella 4.2).

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

Tabella 4.1 - R^2 Score e RMSE

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

Tabella 4.2 - R^2 Score e RMSE CrossValidation

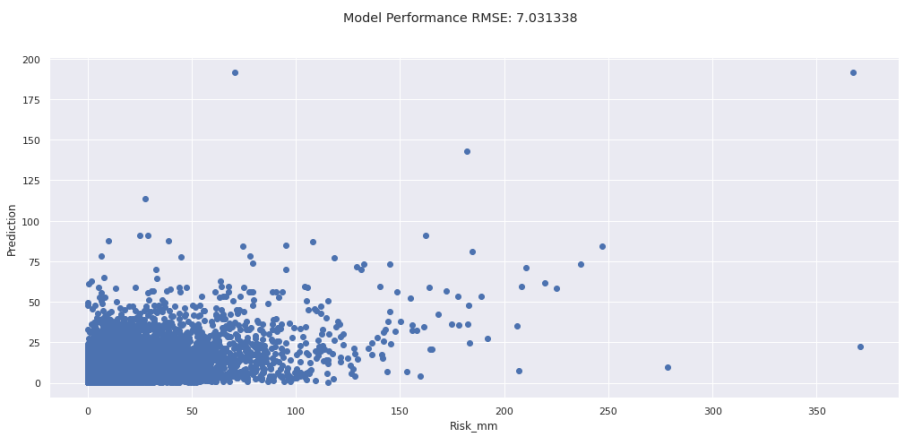

Figura 4.1 - Risk_mm con RFR CrossValidation

Infine, abbiamo sfruttato la funzione ***featureImportance*** del modello di regressione RandomForest che stabilisce una percentuale su quanto influente sia ciascuna feature sulle predizioni del modello. Per isolare il modello che ha performato meglio nella nostra griglia dei parametri, abbiamo utilizzato ***bestModel*** (Figura 4.3). Sembra che Humidity3pm, Rainfall e MinTemp siano i più grandi predittori del nostro Risk_mm finale.

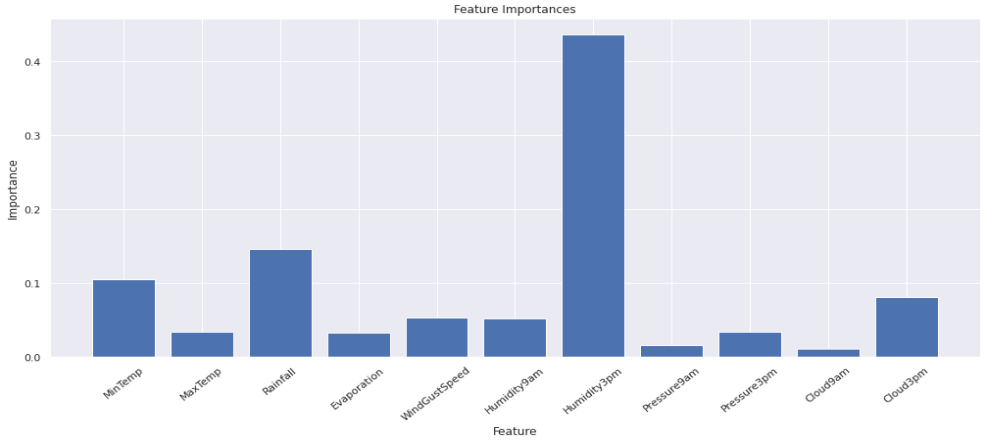

Figura 4.2 – Importanza delle features