clima_australia

September 3, 2021

TRABAJO PRÁCTICO FINAL - Aprendizaje Estadístico

Alumnos:

- Vera, Gonzalo
- Skrauba, Axel

Contents

1 Objetivo

Realizar el análisis pertinente sobre el set de datos "weatherAUS.csv"*, de acuerdo al contenido abordado en la materia. Finalmente, decidir si se puede crear algún modelo con estas variables para poder predecir si va a llover al día siguiente.

---

*Contexto: *Set de datos de variables climáticas en Australia.*

Este conjunto de datos contiene observaciones meteorológicas diarias de numerosas estaciones meteorológicas australianas. La variable de destino **RainTomorrow** significa: ¿Llovió al día siguiente? *Yes* o *No*.

2 Análisis conjunto de datos

Contamos con 24 variables en el dataset. La variable de respuesta en este caso es RainTomorrow y buscamos predecirla mediante las demás predictoras. Excepto RISK_MM ya que provoca data leakage.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 142193 entries, 0 to 142192
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           142193 non-null  object
 1   Location       142193 non-null  object
 2   MinTemp        141556 non-null  float64
 3   MaxTemp        141871 non-null  float64
 4   Rainfall       140787 non-null  float64
 5   Evaporation    81350 non-null   float64
 6   Sunshine       74377 non-null   float64
 7   WindGustDir    132863 non-null  object
 8   WindGustSpeed  132923 non-null  float64
 9   WindDir9am     132180 non-null  object
 10  WindDir3pm     138415 non-null  object
 11  WindSpeed9am   140845 non-null  float64
 12  WindSpeed3pm   139563 non-null  float64
 13  Humidity9am    140419 non-null  float64
 14  Humidity3pm    138583 non-null  float64
 15  Pressure9am    128179 non-null  float64
 16  Pressure3pm    128212 non-null  float64
 17  Cloud9am       88536 non-null   float64
 18  Cloud3pm       85099 non-null   float64
 19  Temp9am        141289 non-null  float64
 20  Temp3pm        139467 non-null  float64
 21  RainToday      140787 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.0+ MB
```

En la tabla anterior, se tiene una descripción resumida de las variables del set de datos. La mayoría son del tipo numéricas (float64) pero existen otras categóricas (object). En principio, se tienen 16 variables numéricas y 7 categóricas. Las categóricas son las que serán analizadas y codificadas para su utilización en los diferentes modelos a abordar.

3 Descripción de variables

Las tabla siguiente muestra el significado de cada una de las variables para referencia.

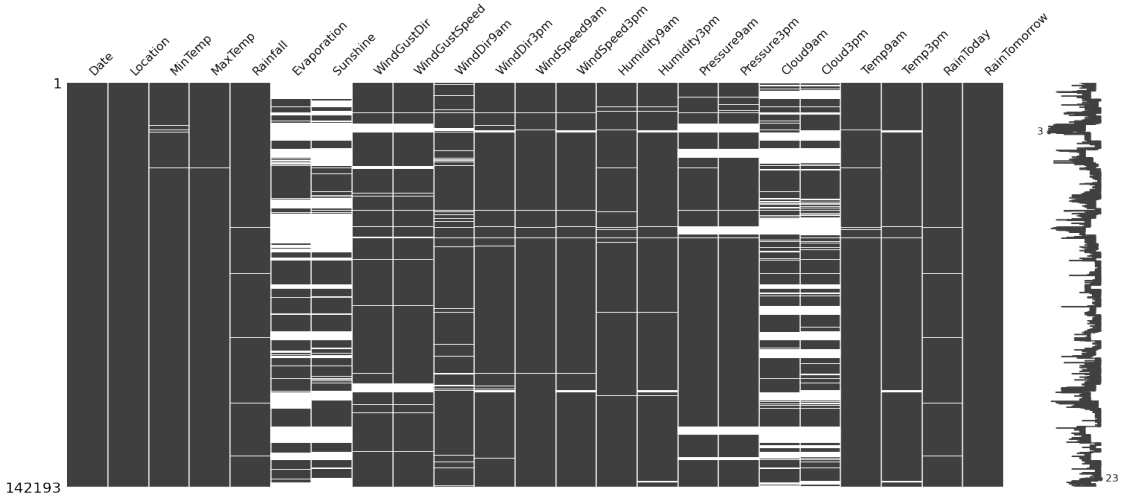
Name	Description
Date	The date of observation (a Date object).
Location	The common name of the location of the weather...
MinTemp	The minimum temperature in degrees celsius.
MaxTemp	The maximum temperature in degrees celsius.
Rainfall	The amount of rainfall recorded for the day in...
Evaporation	The so-called Class A pan evaporation (mm) in ...
Sunshine	The number of hours of bright sunshine in the ...
WindGustDir	The direction of the strongest wind gust in th...
WindGustSpeed	The speed (km/h) of the strongest wind gust in...
Temp9am	Temperature (degrees C) at 9am.
RelHumid9am	Relative humidity (percent) at 9am.
Cloud9am	Fraction of sky obscured by cloud at 9am. This...
WindSpeed9am	Wind speed (km/hr) averaged over 10 minutes pr...
Pressure9am	Atmospheric pressure (hpa) reduced to mean sea...
Temp3pm	Temperature (degrees C) at 3pm.
RelHumid3pm	Relative humidity (percent) at 3pm.
Cloud3pm	Fraction of sky obscured by cloud (in "oktas":...
WindSpeed3pm	Wind speed (km/hr) averaged over 10 minutes pr...
Pressure3pm	Atmospheric pressure (hpa) reduced to mean sea...
ChangeTemp	Change in temperature.
ChangeTempDir	Direction of change in temperature.
ChangeTempMag	Magnitude of change in temperature.
ChangeWindDirect	Direction of wind change.
MaxWindPeriod	Period of maximum wind.
RainToday	Integer: 1 if precipitation (mm) in the 24 hou...
TempRange	Difference between minimum and maximum tempera...
PressureChange	Change in pressure.
RISK_MM	The amount of rain. A kind of measure of the "...
RainTomorrow	The target variable. Did it rain tomorrow

4 Manejo de datos faltantes

En la tabla siguiente se presentan las variables con datos faltantes, indicando el porcentaje computado de los mismos. La tabla se encuentra ordenada, siendo el primer valor el que carece de un mayor número de valores.

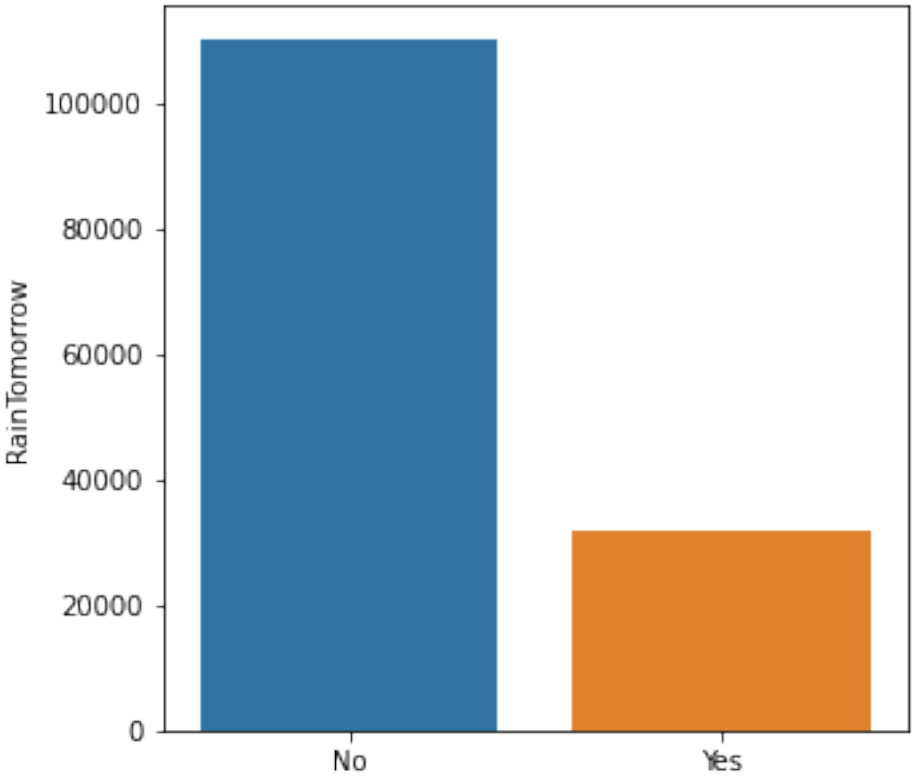
	Porcentaje Nulos
Sunshine	47.692924
Evaporation	42.789026
Cloud3pm	40.152469
Cloud9am	37.735332
Pressure9am	9.855619
Pressure3pm	9.832411
WindDir9am	7.041838
WindGustDir	6.561504
WindGustSpeed	6.519308
WindDir3pm	2.656952

La siguiente gráfica, representa lo equivalente a la tabla anterior con el porcentaje de valores faltantes por variables. La gráfica permite una identificación visual rápida, además de permitir detectar algún patrón en lo que respecta a la ausencia de los datos.



Habría que determinar si el mecanismo de pérdida de datos es completamente aleatorio o no. Las columnas mayormente afectadas son Sunshine (47%), Evaporation (42%), Cloud3pm (40%) y Cloud9am (37%). En las mismas, las pérdidas se agrupan en grandes bloques. Habría que contrastar esto con otras columnas, tal vez exista alguna relación. Por ejemplo, alguna ciudad no podía medir esos parámetros por no contar con los instrumentos. Dado el caso, se podría utilizar alguna técnica para la imputación de los mismos. En algunas zonas, se percibe que la pérdida de datos se da para una misma fila en casi todas las columnas. Estos podrían descartarse directamente.

5 Variable respuesta



Como se observa en la figura, la clase se encuentra desbalanceada. Los días que "No" llueven son la mayoría. Para este inconveniente, se podría optar por una gestión de desequilibrio de clases.

Algunas estrategias posibles son el balanceo del conjunto de datos realizando un sub sampleo de los casos negativos o creando registros positivos ficticios. Otra forma de compensar el efecto del desequilibro es la modificación del threshold en el clasificador, por ejemplo, en regresión logística o discriminantes lineales.

6 Tratamiento de variables

Las variables predictoras no pueden ser ingresadas en los algoritmos de modelado sin antes ser adecuadas para su uso. Entre otras cosas, se deben codificar las variables categóricas y analizar la correlación simple entre variables numéricas.

6.1 Variables categóricas

- **Date**: Esta variable como se presenta no aporta información al modelo. La predicción de lluvia al día siguiente no se ve afectada por su fecha exacta pero si es interesante la época del año que se está atravezando. Cambiarla a una variable discreta ordinal tampoco tiene sentido porque la distancia entre el mes 12 y el 1 es la misma que entre 4 y 5. Por lo que

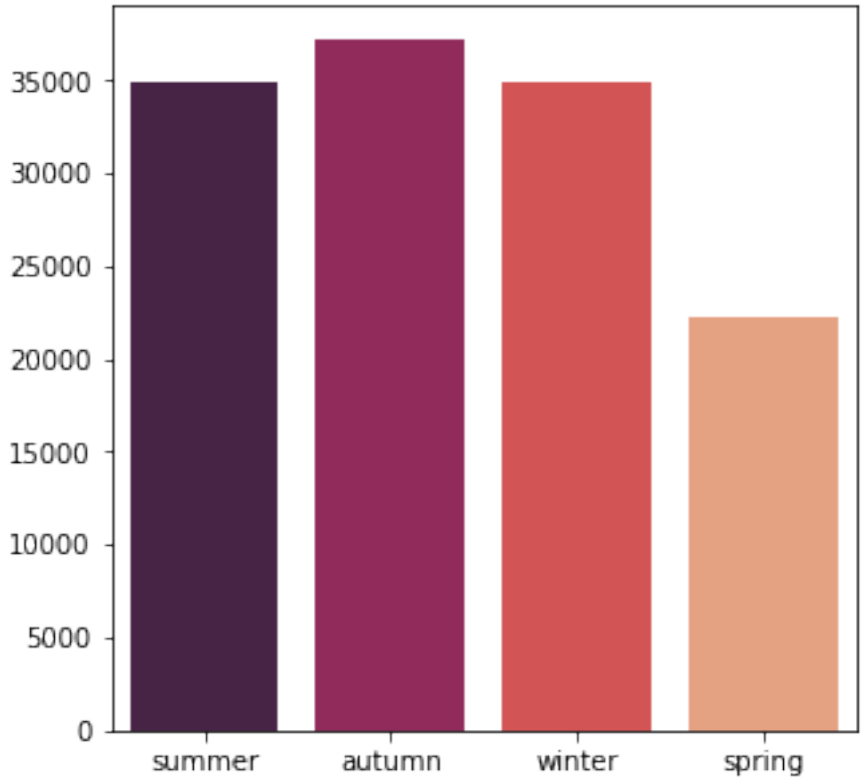
proponemos cambiar a una variable categórica que indique la estación, o sea; Verano, Otoño, Invierno y Primavera.

- **Location**: Las condiciones climáticas pueden variar de ciudad a ciudad por lo que se incluye en el modelo mediante una codificacion one-hot encoding.
- **WindGustDir, WinDir9am, WinDir3pm**: Son variables categóricas en principio. Indica la dirección del viento en 16 puntos de compás. Como provienen de un círculo completo es posible transformar estos puntos en grados. Según http://www.nciwormshead.org.uk/training/compass/68-compass-rose. Sin embargo, la distancia no se cumple en la circunferencia completa. Al igual que en la fecha, la distancia entre 359º y 0º es 1º y no es interpretado de esa manera por el modelo. A tales efecto, proponemos dividir cada direccion en dos columnas asociadas a las proyecciones horizontal y vertical mediante el seno y el coseno. Las proyecciones son variables numéricas que guardan una relación de distancia entre las diferentes direcciones.

A continuación, se procesan las variables categóricas de acuerdo al análisis realizado previamente.

6.1.1 Date

Transformamos esta variable a una categórica que represente la estación del año y la codificamos por one-hot encoding. El resultado de este procedimiento, se percibe en la gráfica y tabla de esta sección.

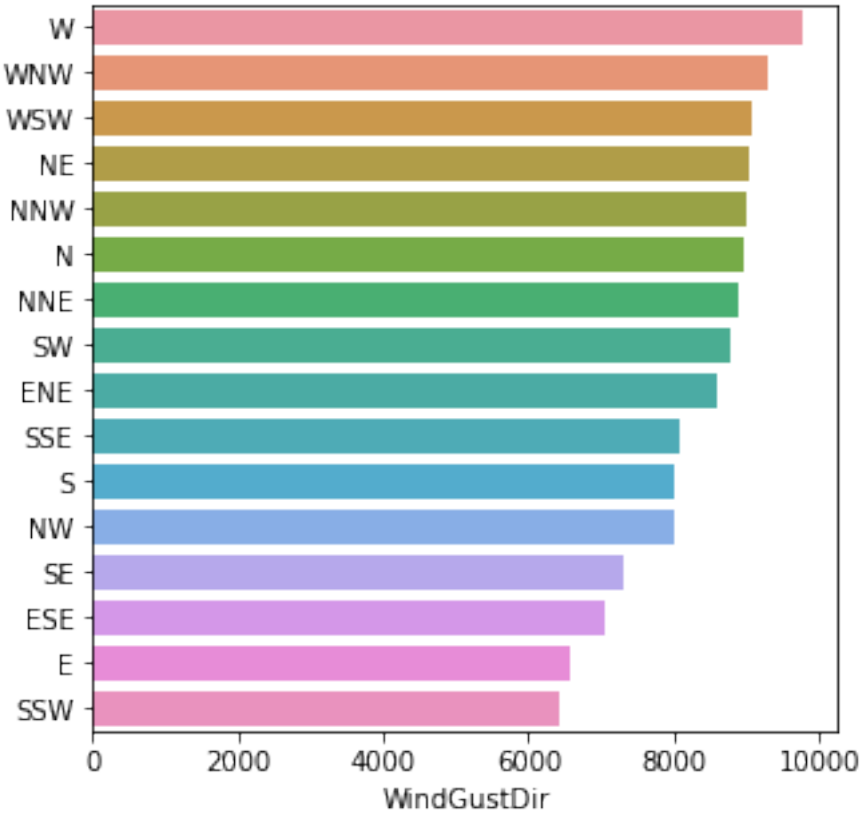


autumn	winter	spring
0	0	1
0	1	0
0	0	0
1	0	0

La ausencia de las 3 será interpretada como verano (summer).

6.1.2 Wind directions

Aquí se muestra una gráfica de la frecuencia con la que se observan las distintas direcciones de viento en la variable WindGustDir.



Para codificar las tres variables 'WindGustDir', 'WindDir9am' y 'WindDir3pm' utilizamos la siguiente conversión a grados sexagesimales.

	Punto de compás	Dirección del viento
0	W	270.0
1	WNW	292.5
2	WSW	247.5
3	NE	45.0
4	NNW	337.5
5	N	0.0
6	NNE	22.5
7	SW	225.0
8	ENE	67.5
9	SSE	157.5
10	S	180.0
11	NW	315.0
12	SE	135.0
13	ESE	112.5
14	E	90.0
15	SSW	202.5

Aplicada la transformación, las columnas quedan de la siguiente manera

WindGustDir	WindDir9am	WindDir3pm
270.0	270.0	292.5
292.5	337.5	247.5
247.5	270.0	247.5
45.0	135.0	90.0
270.0	67.5	315.0

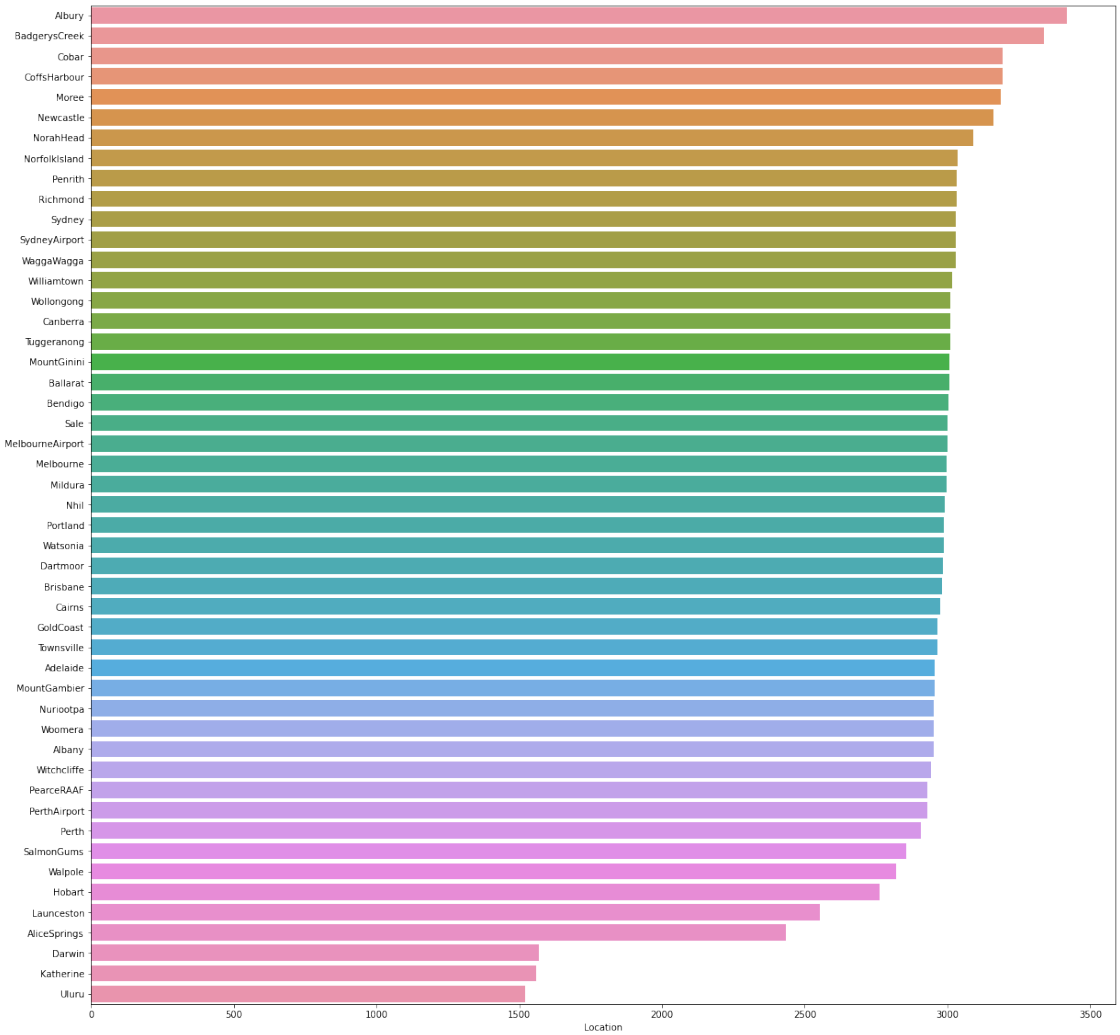
Luego, como las distancias entre grados no son lineales, convertimos cada dirección del viento en dos columnas, correspondientes a sus proyecciones en el eje horizontal y vertical.

WGustDir_sin	WGustDir_cos	WDir3pm_sin	WDir3pm_cos	WDir9am_sin	WDir9am_cos
-1.000000	-1.836970e-16	-0.923880	3.826834e-01	-1.000000	-1.836970e-16
-0.923880	3.826834e-01	-0.923880	-3.826834e-01	-0.382683	9.238795e-01
-0.923880	-3.826834e-01	-0.923880	-3.826834e-01	-1.000000	-1.836970e-16
0.707107	7.071068e-01	1.000000	6.123234e-17	0.707107	-7.071068e-01
-1.000000	-1.836970e-16	-0.707107	7.071068e-01	0.923880	3.826834e-01

6.1.3 Location

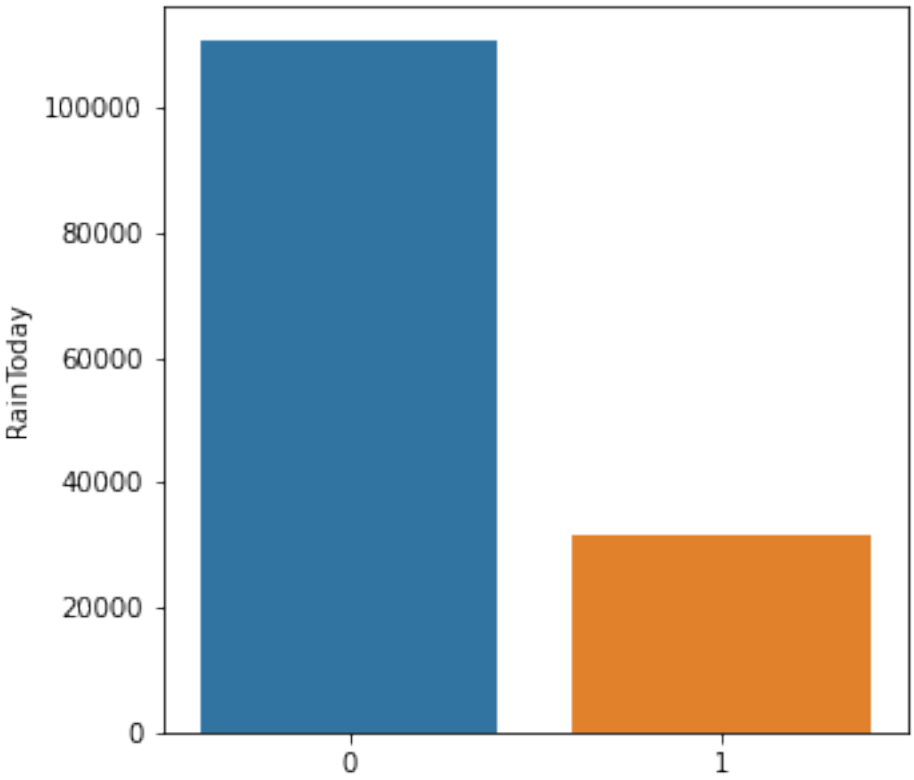
Esta variable es codificada directamente utilizando one-hot-encoding. Como resultado se obtienen 48 columnas que indican si el registro pertenece a una ciudad.

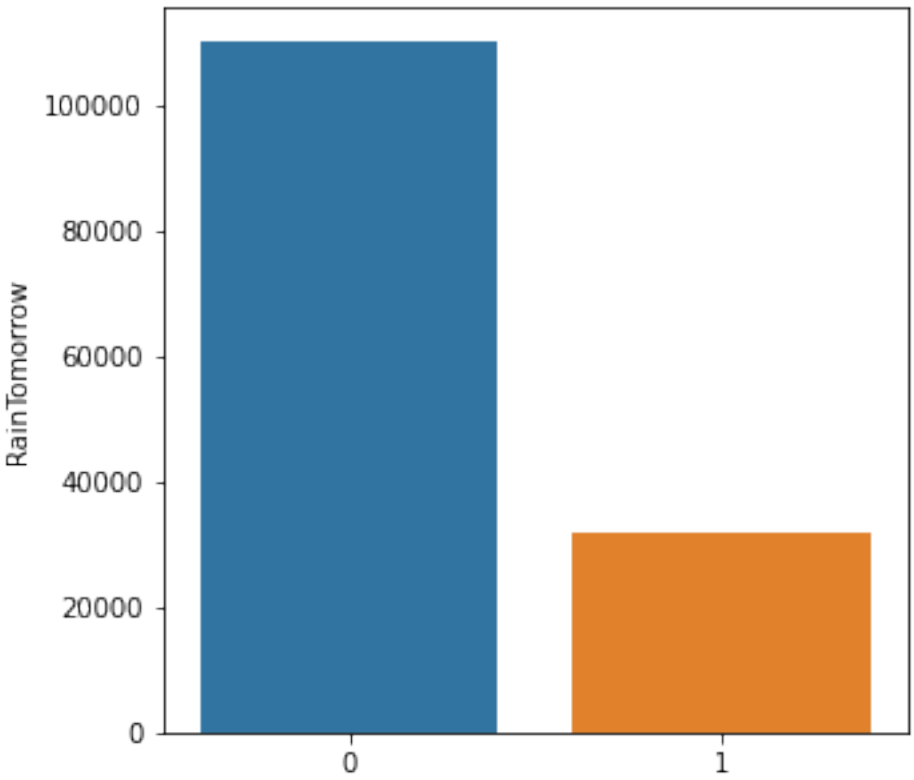
```
Ciudades distintas: 49
```



6.1.4 RainToday y RainTomorrow

Estas variables son del tipo binarias, mediante las graficas de barras se observa que la proporción de casos es aproximadamente la misma en ambas clases. Lo cual tiene sentido.


RainToday
100000
80000
60000
40000
20000
0
0
1

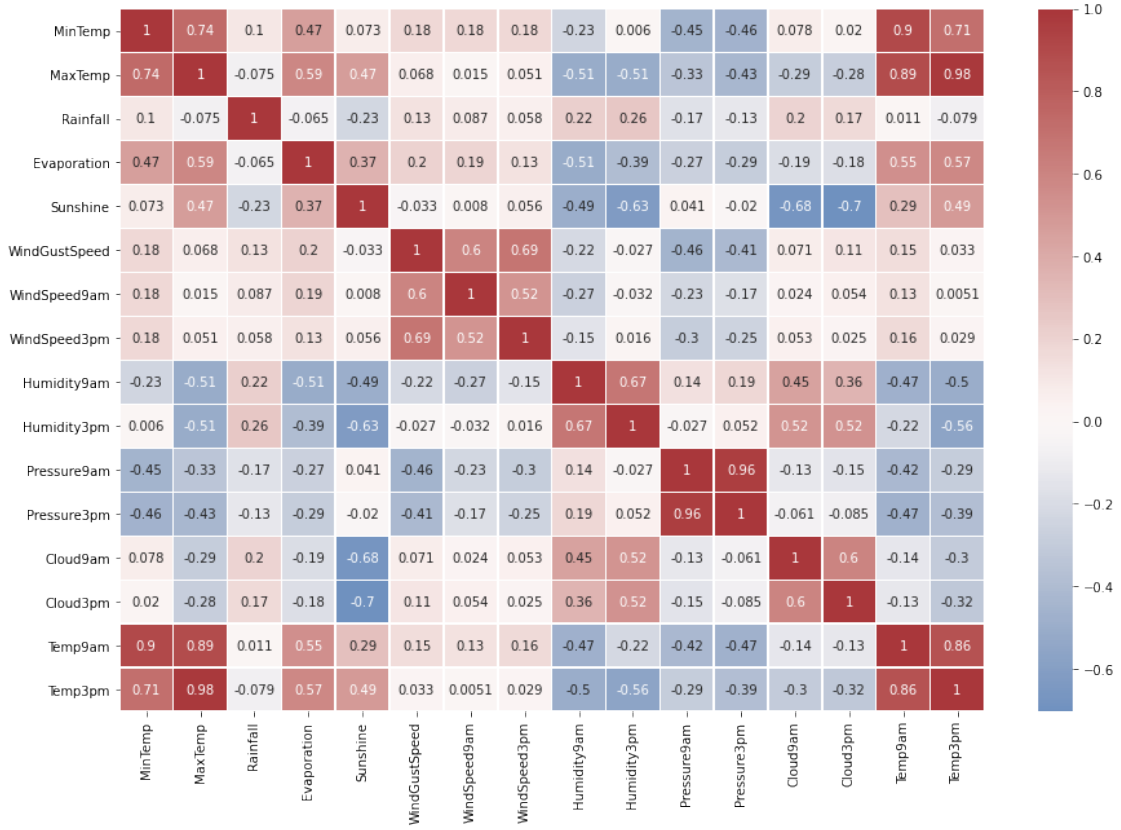

RainTomorrow
100000
80000
60000
40000
20000
0
0
1

6.2 Variables numéricas

Todas las variables numéricas son transformadas mediante un StandardScaler a fines de optimizar el proceso para la obtención de los modelos.

6.2.1 Correlación

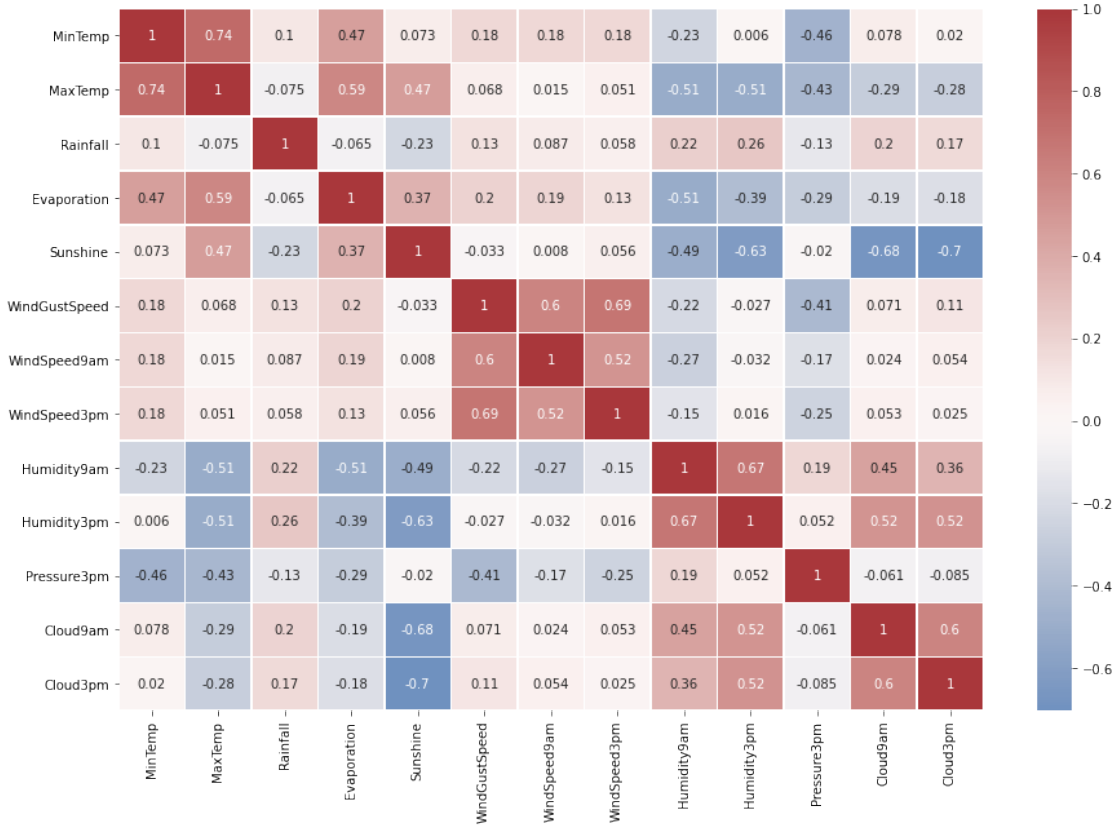
A continuación, se inspeccionan las variables en búsqueda de correlaciones:



- Existe una colinearidad muy alta entre 'Temp3pm' y 'Maxtemp'. Por lo que consideramos eliminar la variable 'Temp3pm' ya que tiene más valores vacíos.
- También, Temp9am tiene una alta correlación con MinTemp y MaxTemp, probablemente son linealmente estimadas. Por lo que se decide descartar Temp9m.
- Además, se observa alta correlación entre las mediciones de presión 'Pressure9am' y 'Pressure3pm', se descarta 'Pressure9am' ya que la lectura es realizada más temprano, por lo tanto, tendrá menos relación con la lluvia al siguiente día.

Estas variables serán descartadas para los algoritmos lineales, como regresión lineal y discriminantes lineales, que son afectados por la colinealidad de los predictores. Otros algoritmos iterativos como los árboles de decisión, no se ven tan afectados.

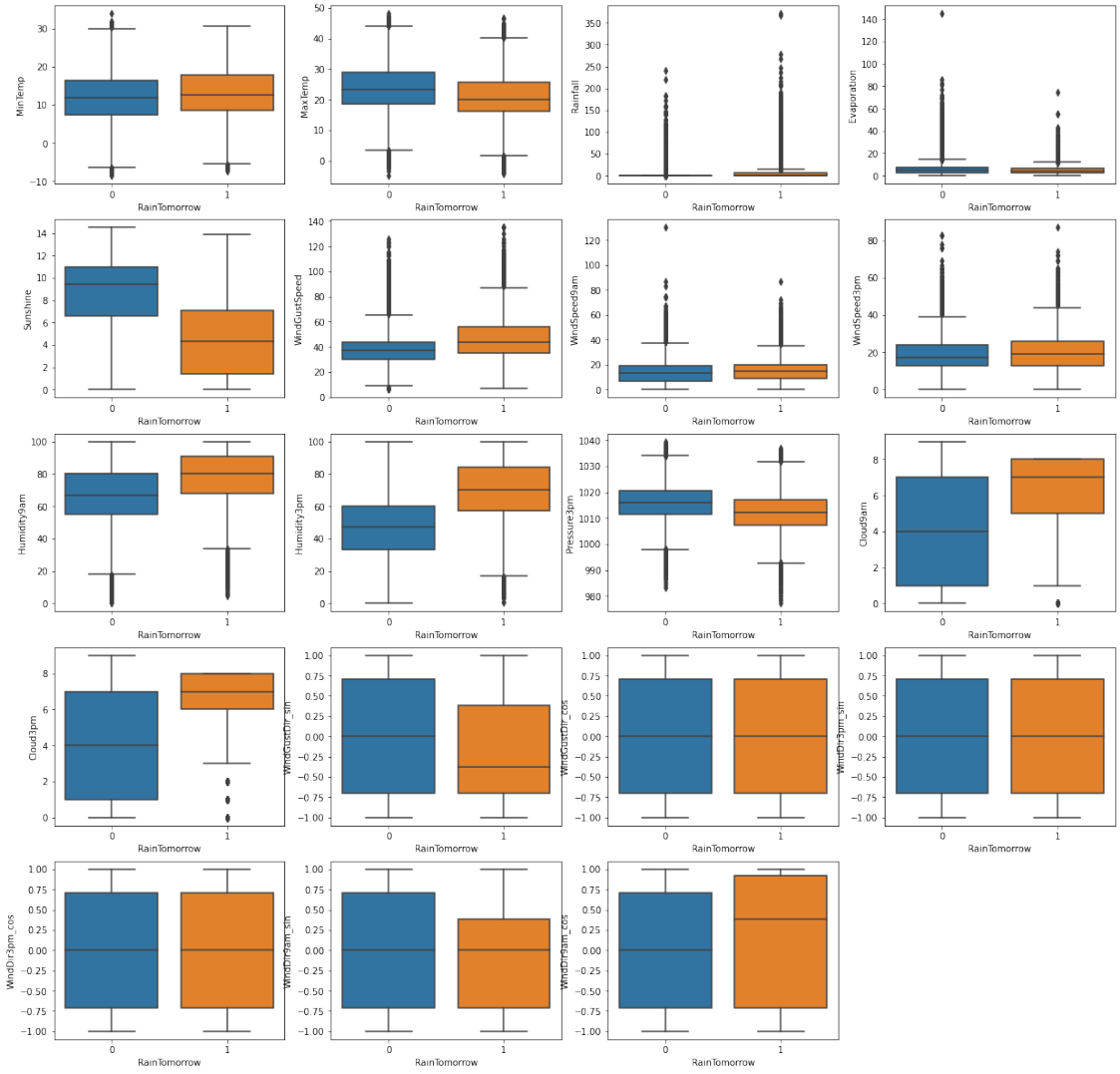
Predictores descorrelacionados



Luego de descartar las variables mencionadas anteriormente, se observa que no existe problemas de multicolinearidad en el set de datos.

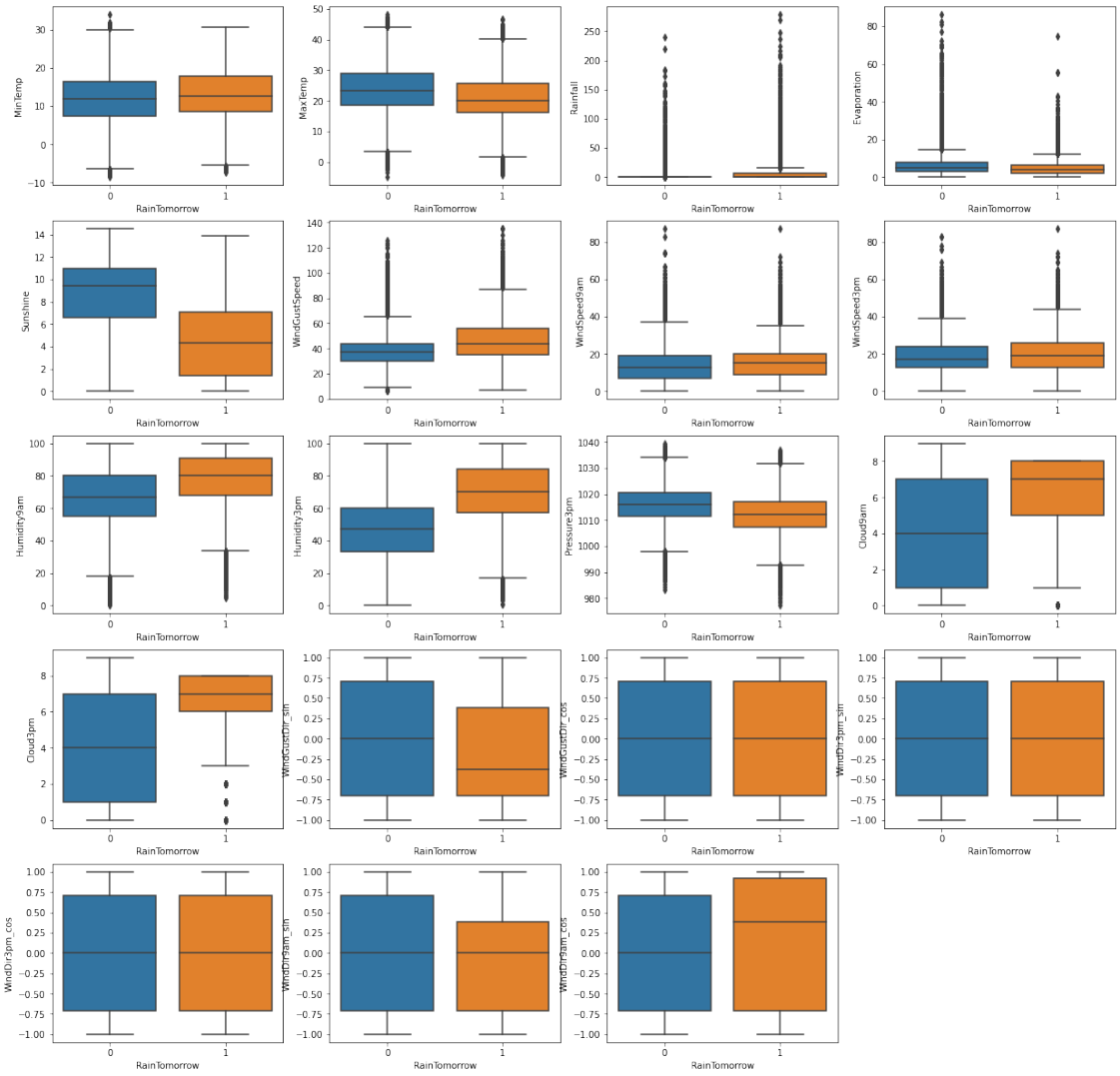
6.2.2 Outliers

Para identificar las variables con outliers, se realizan diagramas de caja y bigote.



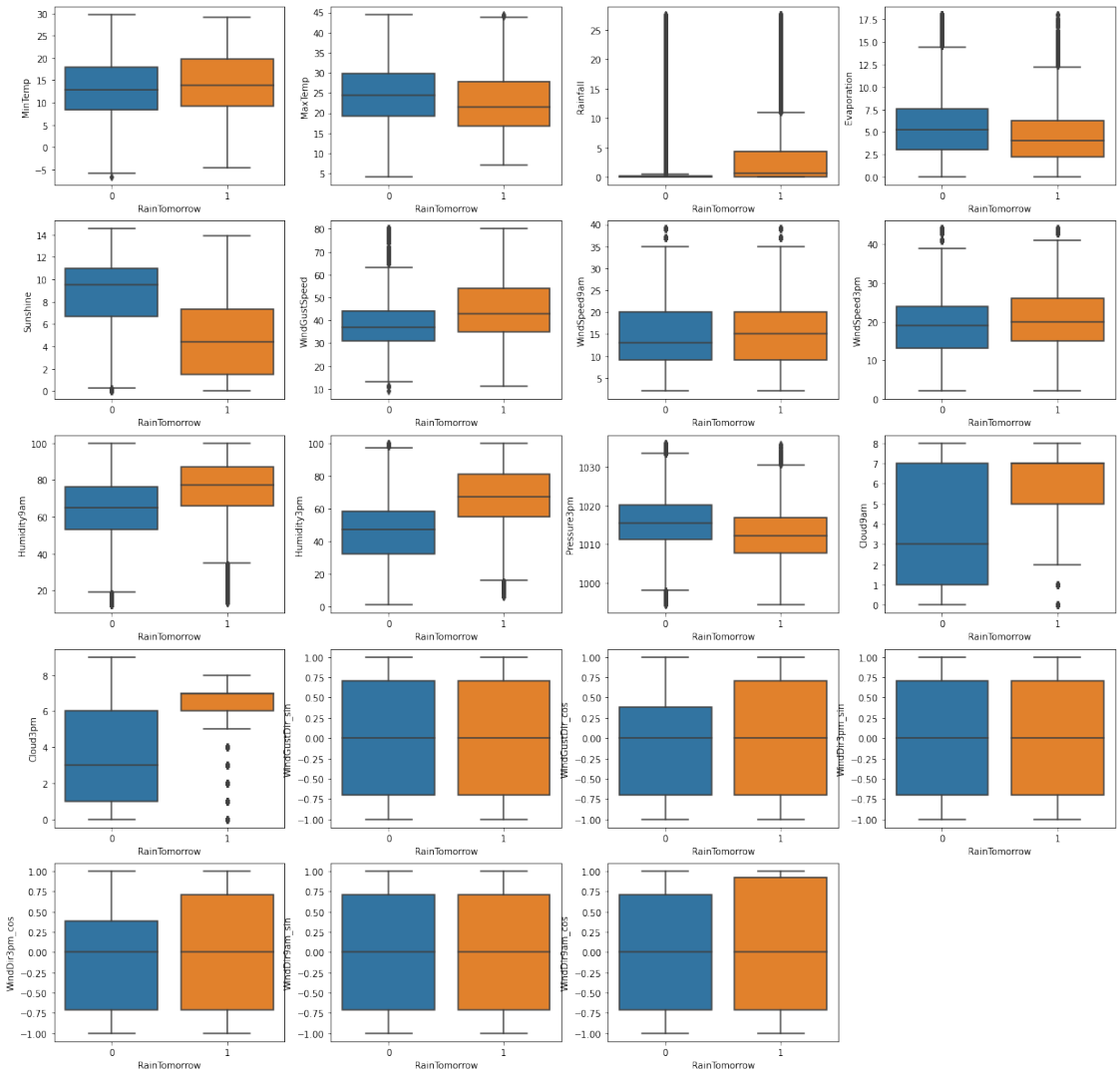
Mediante una inspección visual y siendo muy flexibles podríamos decir que 'Rainfall', 'Evaporation' y 'WindSpeed9am' tienen valores atípicos.

Descarte por inspección visual Descartamos por inspección visual los outliers más evidentes en las variables mencionadas arriba. El resultado es el siguiente



Descarte por zscore

En este caso, descartamos aquellos valores que superen el valor de 3 en el *zscore* de la variable estandarizada.



Se obtienen distribuciones más simétricas. Sobre todo en las variables 'WindSpeed9am', 'WindSpeed3am' y 'Evaporation'.

7 Modelo Base: Dataset original, descarte de registros vacíos

La construcción de modelos la realizamos experimentando con diferentes configuraciones de datos para estudiar la influencia de las transformaciones realizadas en el desempeño de los modelos.

En primer lugar, se optará por la eliminación de las filas con datos faltantes y la consideración de las variables del tipo numéricas. Posteriormente, se incluirán las variables categóricas con las transformaciones mencionadas.

La alternativa siguiente consistirá en la imputación de los datos faltantes, a fines de maximizar la disponibilidad de información del set original. Dentro de la imputación, se emplearán técnicas de regresión, con el objetivo de no modificar la distribución de los datos de cada columna (inconveniente

que podría ocurrir utilizando la media por ejemplo).

```
Tamaño del conjunto de datos:
```

```
[ ]: (56420, 23)
```

7.1 Variables numéricas

En esta sección, se presentan los resultados obtenidos considerando únicamente las variables del tipo numéricas del set original, con la eliminación de las entradas con datos faltantes. Los *scores* obtenidos servirán como una línea base, permitiendo definir un punto de comparación para los entrenamientos con procesamiento de las variables.

Scores CV set de entrenamiento

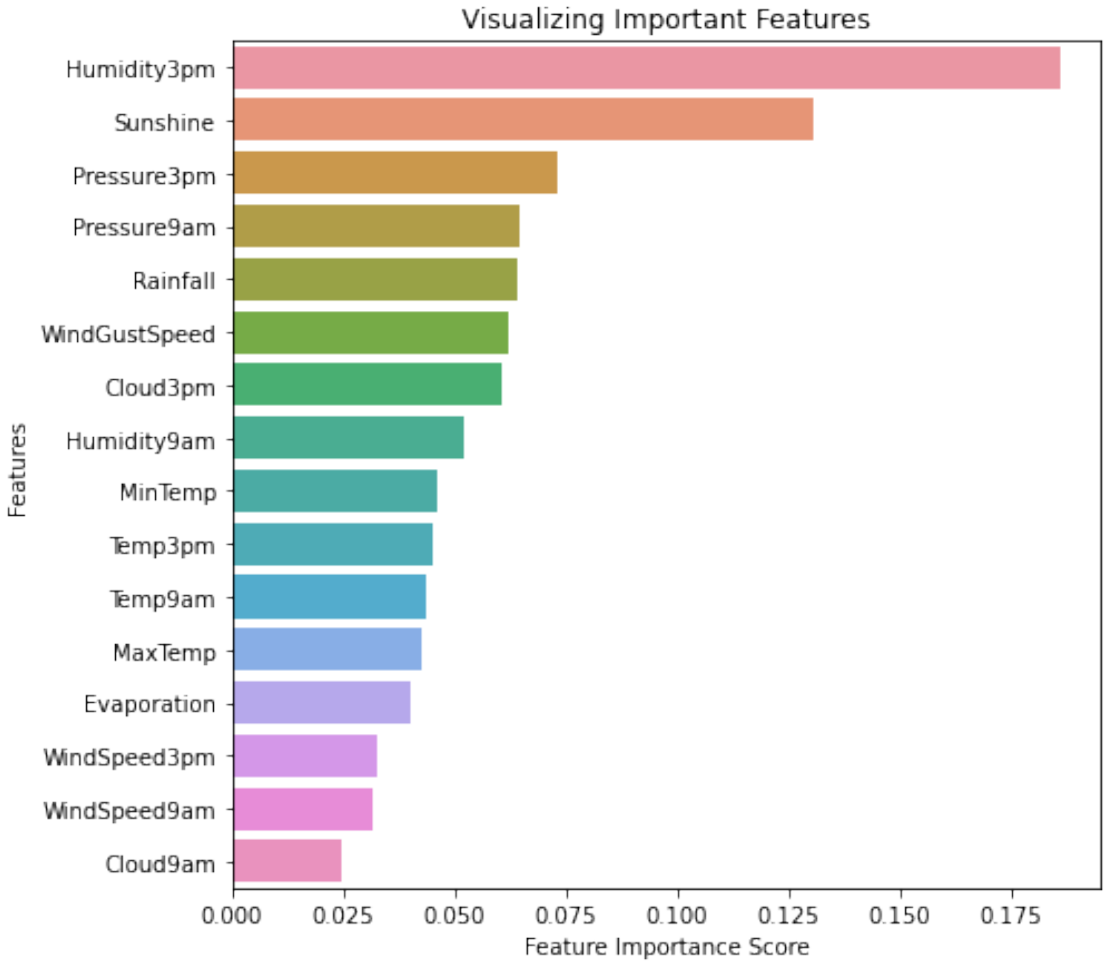
	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.854	0.851	0.801	0.856	0.802
Recall	0.526	0.534	0.682	0.523	0.512
Precision	0.731	0.715	0.536	0.744	0.545
F1	0.612	0.611	0.600	0.612	0.538
AUC	0.883	0.882	0.849	0.887	0.709

Scores set evaluación

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.854	0.853	0.800	0.860	0.804
Recall	0.537	0.548	0.687	0.528	0.527
Precision	0.733	0.722	0.539	0.765	0.562
F1	0.620	0.623	0.604	0.625	0.544
AUC	0.741	0.744	0.760	0.741	0.705

- De forma general, el valor de *Recall* obtenido es bajo para esta alternativa, dejando en evidencia un rendimiento pobre respecto a los falsos negativos. Este inconveniente puede mejorarse con una de las técnicas mencionadas para el balanceo del set de datos.
- El mejor *AUC* corresponde al Algoritmo de clasificación: *Gaussiano Naive-Bayes.*
- Además, la mayoría de modelos presenta una diferencia aproximada del **14%** entre el **AUC** del set para entrenamiento y el de test, dando pautas de la existencia de cierto *overfitting.*

Del modelo Random Forest (RF), se realiza una gráfica para representar la importancia adoptada de las variables que intervienen en el modelo. Cabe resaltar que algunas de las variables con un mayor porcentaje de datos faltantes, se encuentran dentro de las variables con mayor importancia, identificadas por RF.



7.2 Variables numéricas balanceado

El set de datos bajo análisis, como ya se ha visto, presenta el conjunto de datos desequilibrados, por lo tanto el modelo puede no aprender bien respecto a la clase minoritaria (días que "Si" llueve). Este problema, en cierta manera puede verse reflejado en el *Recall* de los últimos resultados.

Como alternativa a este inconveniente, se procede a la generación de nuevas entradas para la clase minoritaria, a fines de lograr una salida más balanceada. Esto se lleva a cabo sobre el set de entrenamiento únicamente. Para las nuevas entradas, se aplica el método *SMOTE-Tomek*, que es una técnica de sobremuestreo de minorías sintéticas que a su vez se combina con submuestreo.

A modo comparativo, se indican las cantidades asociadas a cada clase, previo y posterior a la acción del balanceo. Se utilizó un ratio de **0.8**, es decir, la clase minoritaria pasa a conformarse por un número de entradas que equivale al **80 %** de la clase mayoritaria.

```
Distribución antes:
0    22035
1     6175
Distribución despues:
```

```
0    21910
1    17503
```

Con estas especificaciones, se obtienen los resultados presentados a continuación. Correspondientes tanto al set de entrenamiento como el de pruebas o evaluación.

Scores CV set de entrenamiento

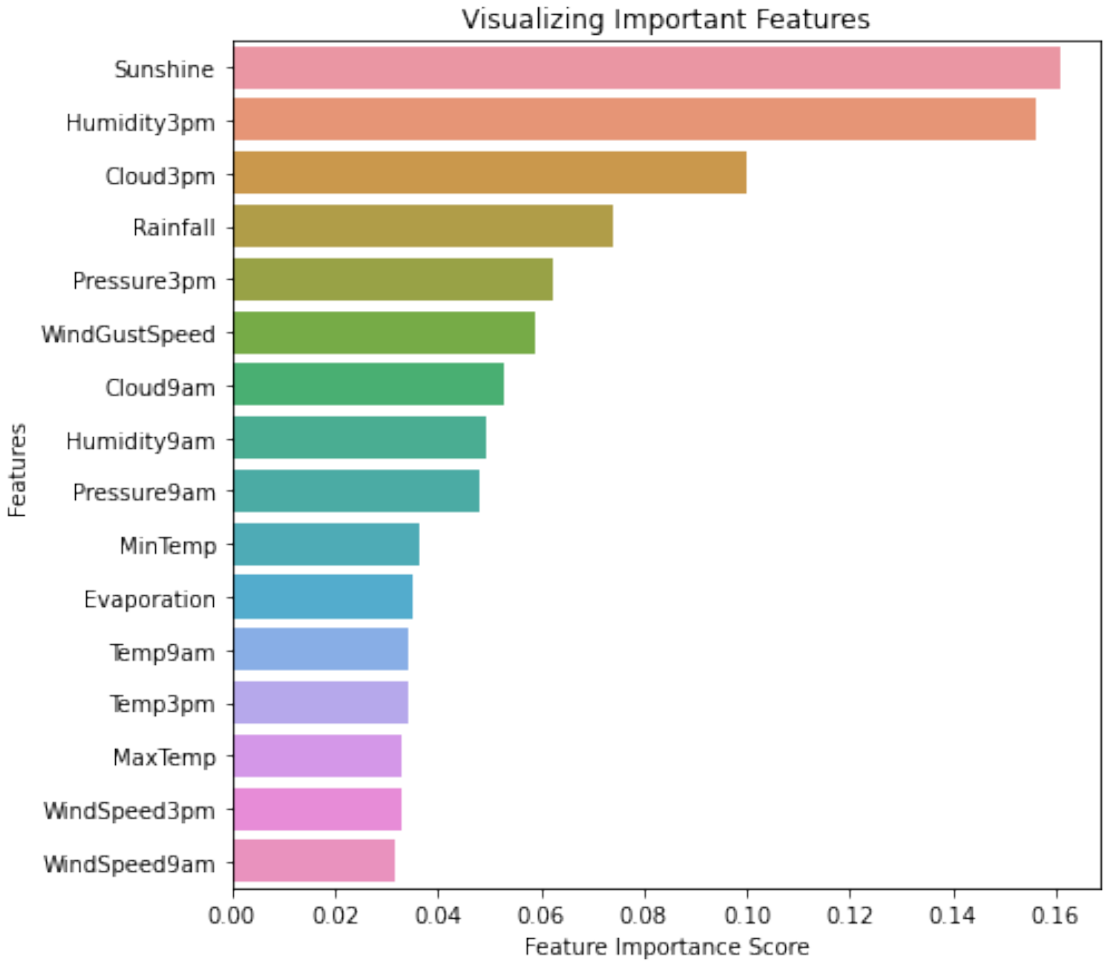
	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.803	0.800	0.775	0.896	0.820
Recall	0.754	0.743	0.765	0.883	0.810
Precision	0.792	0.793	0.738	0.882	0.794
F1	0.773	0.767	0.751	0.879	0.804
AUC	0.889	0.888	0.855	0.964	0.842

Scores set evaluación

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.818	0.820	0.774	0.851	0.787
Recall	0.754	0.742	0.755	0.634	0.607
Precision	0.568	0.572	0.493	0.674	0.517
F1	0.648	0.646	0.597	0.653	0.558
AUC	0.795	0.792	0.767	0.773	0.723

- Como se observa en las métricas, con el hecho de balancear el set de entrenamiento, los resultados mejoran considerablemente. Ahora se tiene un *Recall* en el orden de **0.75**, en contraste al anterior de **0.5**.
- El modelo Random Forest presenta cierto grado de *overfitting*, las diferencias entre las métricas de entrenamiento y test son las mayores.
- Los mejores modelos son el de Regresión Logística y el Gaussiano, y les sigue muy de cerca el de Random Forest.

Dentro de las variables o características que el modelo de Random Forest determina más significativas o importantes, se presenta a continuación una gráfica según su orden de importancia:



Resumen

Sin considerar las variables categóricas, se obtiene para los mejores modelos un *ROC* aproximado de: **0.79**

7.3 Considerando variables categóricas no balanceado

En esta prueba, se incluyen las variables categóricas del set de datos. Para su utilización preliminar, se recurrió a una codificación sencilla en dónde se mapea cada elemento de una categoría con un valor entero. En forma resumida, este método tranforma una variable categórica en una ordinal. Para la columna de "*date*", en esta transformación se consideraron los meses únicamente.

Este tipo de codificación no es el adecuado para los datos en el contexto que se tiene porque establablece un ranking de acuerdo al entero asignado, es decir, una suerte de escala en dónde los valores más altos serán considerados como "más importantes" (se aclaró con mayor detalle, en secciones anteriores, este fenómeno).

Este ensayo se realiza para conformar una base respecto a la mejora de los modelos con la inclusión de estas variables. Posteriormente, se contrastará con lo que se obtiene al codificar estas variables

de acuerdo al análisis individual de cada una, efectuado anteriormente.

Los modelos se comportan de la siguiente manera:

Scores CV set de entrenamiento

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.852	0.852	0.802	0.858	0.802
Recall	0.527	0.544	0.683	0.527	0.508
Precision	0.724	0.711	0.538	0.748	0.557
F1	0.610	0.616	0.602	0.620	0.520
AUC	0.884	0.883	0.849	0.888	0.710

Scores set evaluación

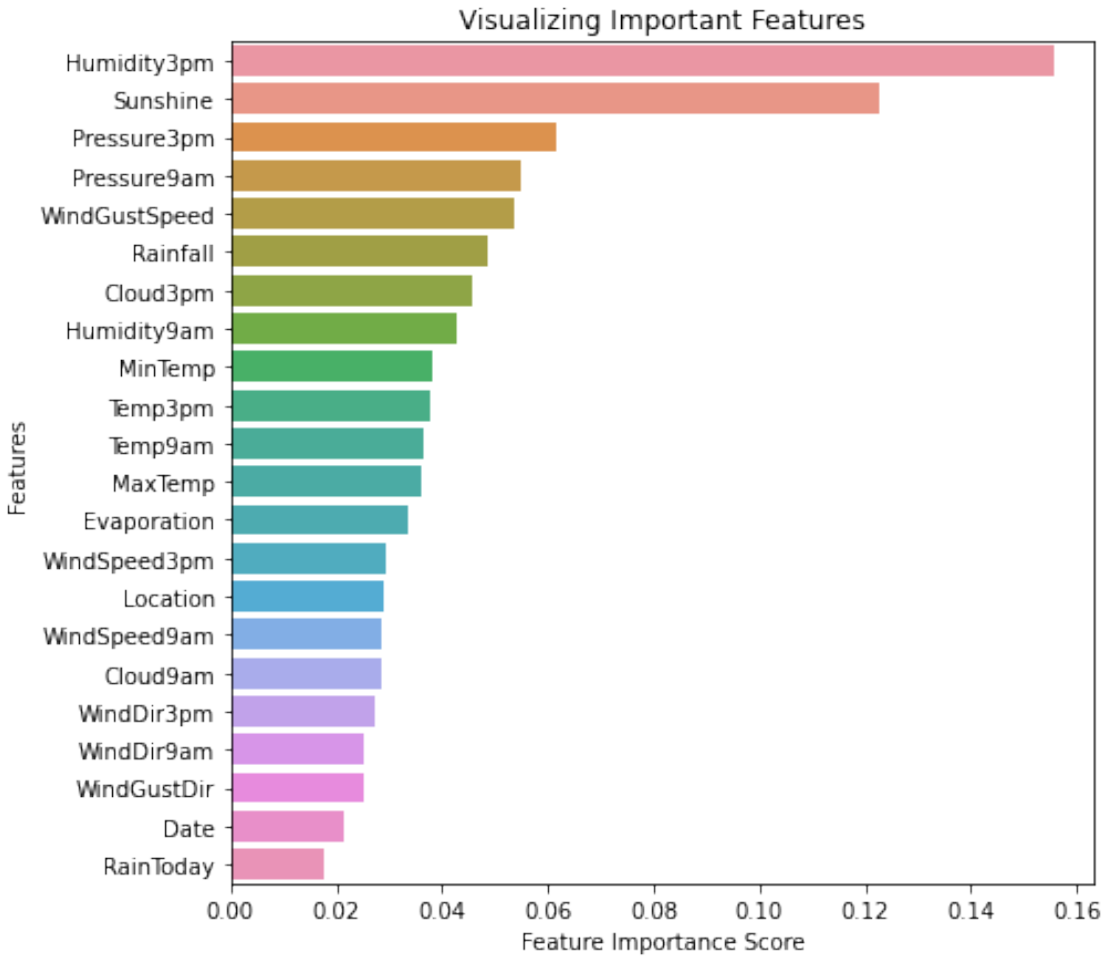
	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.856	0.853	0.802	0.862	0.808
Recall	0.546	0.556	0.690	0.535	0.516
Precision	0.735	0.719	0.542	0.770	0.576
F1	0.627	0.627	0.607	0.631	0.544
AUC	0.745	0.747	0.762	0.745	0.704

Los resultados anteriores corresponden al set de datos sin los datos faltantes, con la codificación de las categoricas a ordinales, sin balanceo. Es decir, que para determinar la influencia de las categóricas, estos resultados pueden compararse con los primeros obtenidos al realizar pruebas sobre el set de datos sin los datos faltantes, y considerando únicamente las variables numéricas.

Al comparar ambas tablas, se deduce que *la adición de los categóricos no mejora los resultados de los modelos*. En principio, no con este tipo de codificación.

Para dar un contraste numérico, en ambos casos el mejor modelo fue el Gaussiano. Sin categóricas se obtuvo un *ROC* de **0.76** y con las categóricas **0.762**.

Finalmente, a modo de representar la importancia de las variables categóricas, se realiza una gráfica de acuerdo a la importancia asumida por el modelo Random Forest. En la gráfica se presentan todas las variables incluidas en los modelos, pero, se aprecia que las categóricas se hallan mayormente en los valores más bajos.



7.4 Análisis de Resultados sobre el modelo base

Considerando las variables categoricas, se obtiene el mejor *ROC* para el modelo *Gaussiano Naive-Bayes*. El mismo corresponde a **0.76**, además, de los mejores modelos, este es el que menos variaciones presenta en los *scores* entre el set de entrenamiento y el de *test*.

8 Manejo de Datos Faltantes

8.1 Quitando columnas con mayoría de datos faltantes

Al quitar *Evaporation*, *Sunshine*, *Cloud3pm* y *Cloud9am*; el set de datos para el entrenamiento será considerablemente mayor al de la *Prueba 1*. Esto no garantiza mejores resultados de todas maneras, ya que *Sunshine* por ej, está 2do dentro de la importancia de las variables en el modelo Random Forest obtenido en la *Prueba 1*.

A continuación se presenta el conteo de los valores de salida (*RainTomorrow*) para los datos con este tratamiento, y los resultantes al realizar el balanceo.

Contrastando numéricamente, en los ensayos anteriores se tenían aproximadamente **22000** entradas y ahora se tienen casi **44000**, es decir, un incremento del doble. Esto se condice con los valores computados respecto a los datos faltantes en las columnas que se han eliminado para esta prueba (Por ej: *Sunshine* con casi **48%** de faltantes).

```
Distribucion antes: Counter({0: 43932, 1: 12530})
Distribucion despu: Counter({0: 43691, 1: 34904})
```

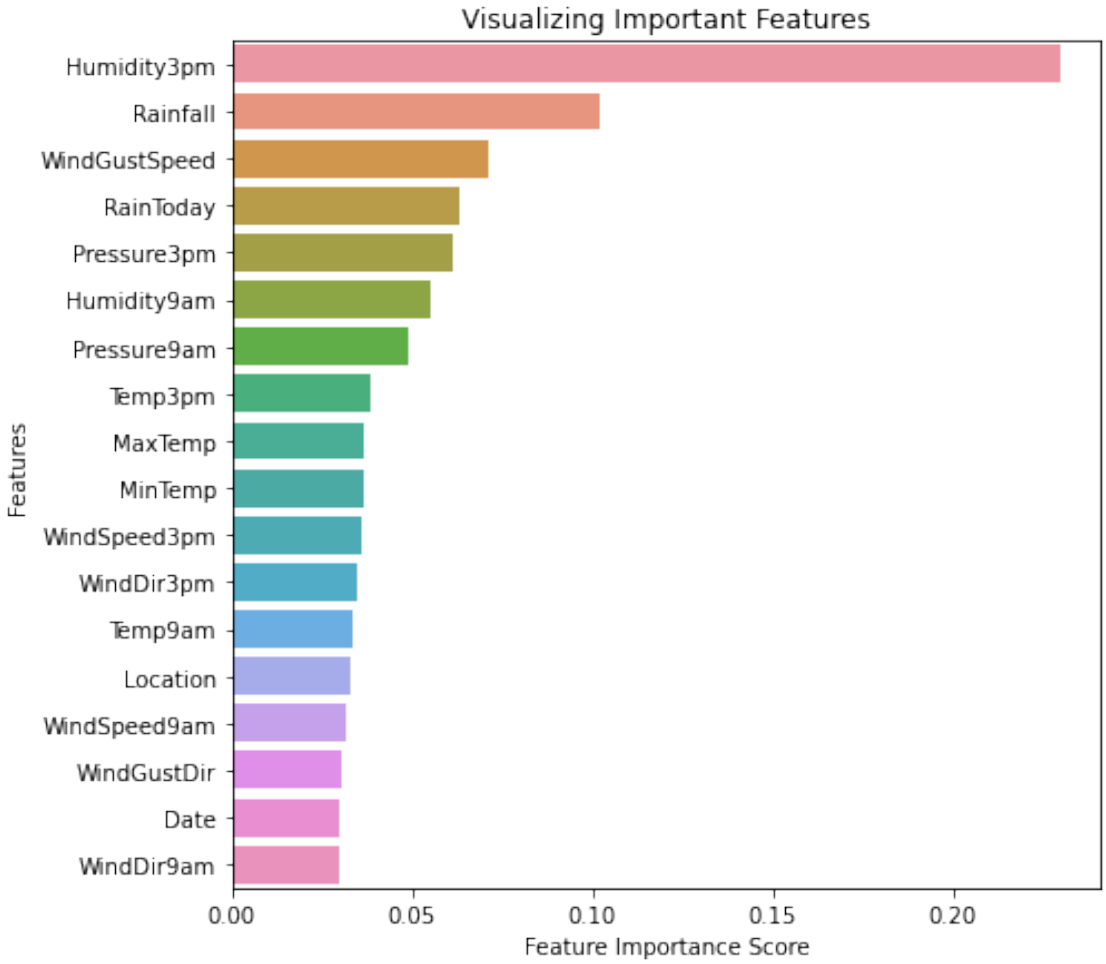
En las tablas siguiente, se tienen los resultados tanto para el set de entramiento como el de test:

Scores CV set de entrenamiento

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.793	0.791	0.749	0.892	0.814
Recall	0.733	0.716	0.625	0.859	0.777
Precision	0.786	0.793	0.766	0.894	0.794
F1	0.758	0.752	0.689	0.872	0.785
AUC	0.871	0.870	0.826	0.963	0.840

Scores set evaluación

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.814	0.818	0.796	0.850	0.783
Recall	0.735	0.718	0.639	0.605	0.600
Precision	0.560	0.570	0.532	0.681	0.508
F1	0.636	0.636	0.580	0.641	0.550
AUC	0.786	0.782	0.739	0.762	0.718



Resumen:

Con el descarte de las columnas con cantidades significativas de datos faltantes, se obtiene para los modelos más significativos un *ROC* de:

- Logistic Regression: 78.5 %
- LDA: 78.3 %
- Random Forest: 76.5 %

NOTA: Si bien existe una leve mejoría en los *scores*, es probable que se deba al incremento de entradas para el proceso de entrenamiento de los modelos. La importancia de las variables asumidas por el modelo Random Forest bajo esta prueba, resulta comparable a la anterior. Por lo tanto, con la eliminación directa de la columna *Sunshine* (que es muy significativa en pruebas anteriores), es probable que se esté perdiendo información significativa del fenómeno bajo análisis. Esto mismo podría estar ocurriendo con las demás variables descartadas por eliminación directa. Lo más conveniente sería optar por un camino que considere la imputación de los faltantes mediante alguna técnica, y, de esta forma, procurar mejores *scores* en los modelos.

8.2 Imputación de datos faltantes

Para evitar modificar la distribución de las diferentes columnas, se opta por utilizar un modelo de regresión para la imputación de los datos. Se ha escogido el algoritmo *MICE* (*Multiple Imputation by Chained Equations*) para este cometido.

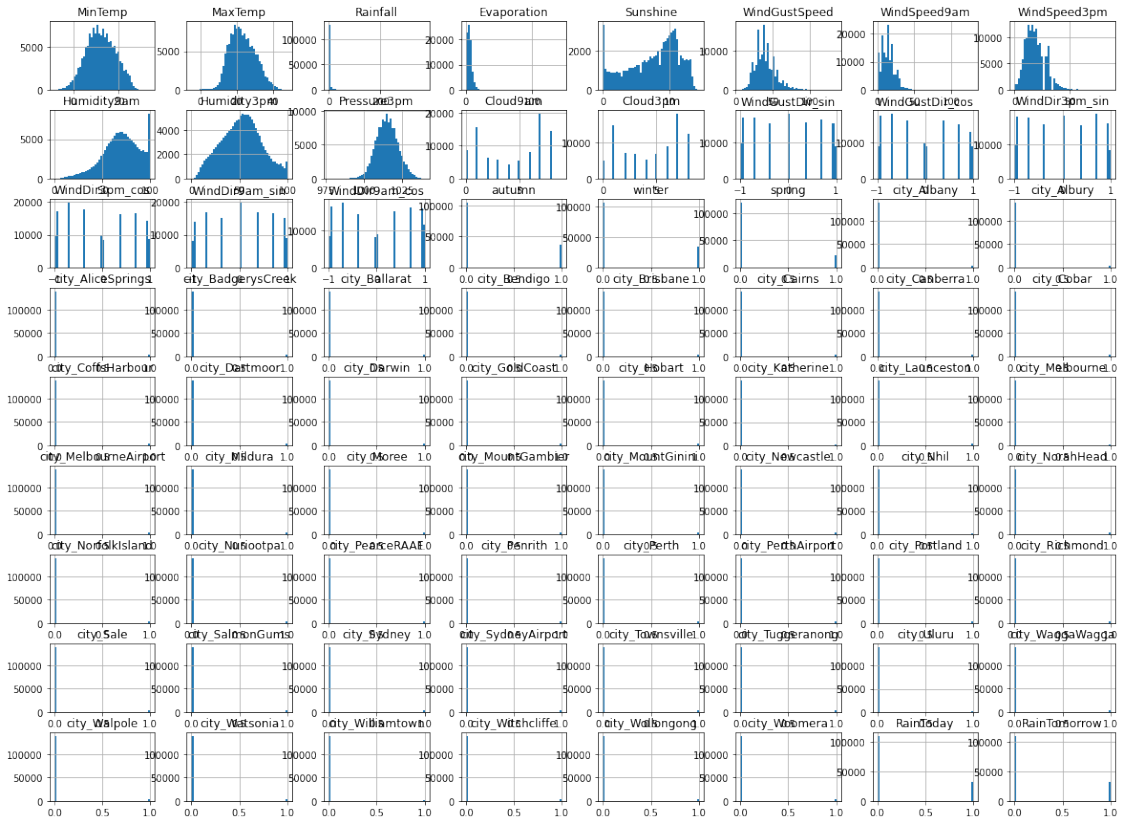
De forma resumida, la imputación múltiple por ecuaciones encadenadas es un método robusto e informativo para tratar los datos faltantes en los conjuntos de datos. El procedimiento imputa los datos faltantes en un conjunto de datos a través de una serie iterativa de modelos predictivos. En cada iteración, cada variable especificada en el conjunto de datos se imputa utilizando las otras variables en el conjunto de datos. Estas iteraciones se ejecutan hasta alcanzar la convergencia.

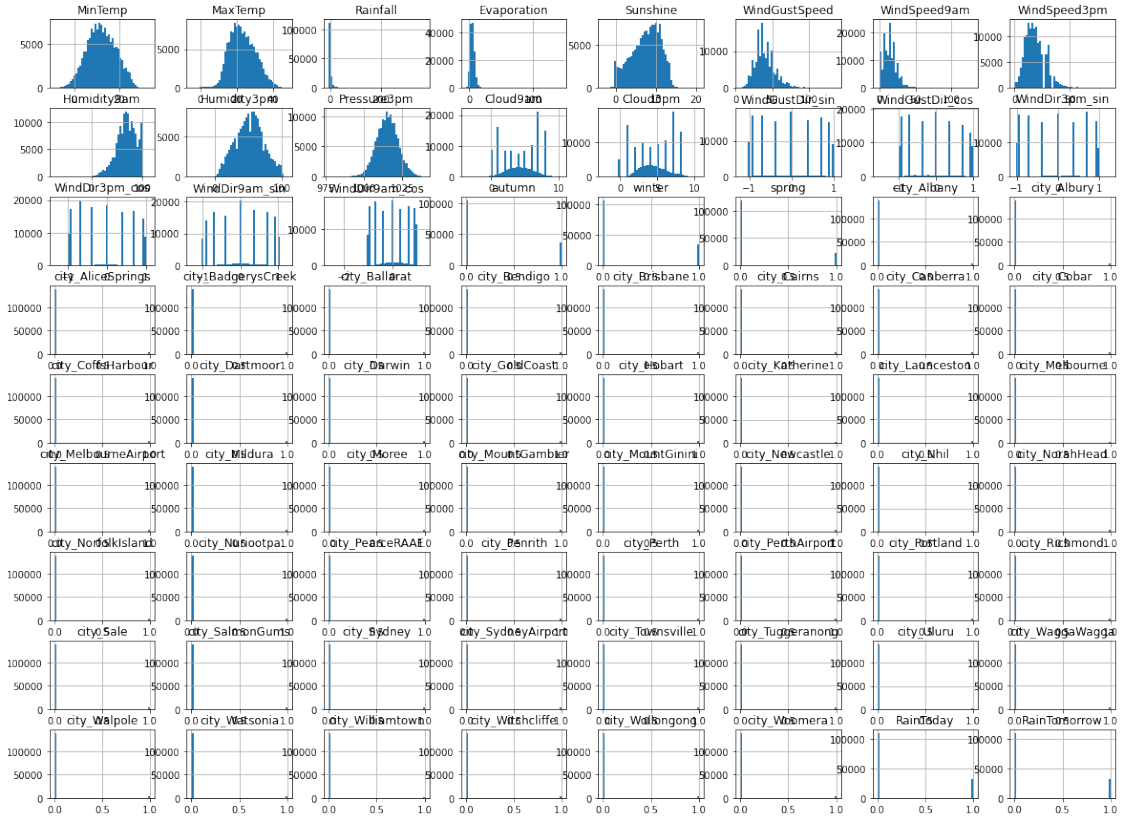
https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html

8.2.1 MICE imputation + SMOTE balance

A continuación, se presetan los histogramas de cada variable incluida en esta prueba. En primer lugar, se observan las distribuciones para las variables originales, y, luego, las resultantes al imputar los faltantes con este método.

Además, con el proceso de imputación y balanceo, ahora se disponen de **54000** entradas para los procesos de entrenamiento y test de resultados.





```
Distribución antes:
0.0    55166
1.0    15930
Distribución despues:
0.0    54820
1.0    43786
```

A continuación, se presentan los resultados obtenidos al utilizar los datos con este tratamiento:

Scores CV set de entrenamiento

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.818	0.816	0.698	0.905	0.827
Recall	0.777	0.767	0.782	0.867	0.779
Precision	0.806	0.809	0.628	0.917	0.819
F1	0.791	0.787	0.697	0.884	0.801
AUC	0.900	0.899	0.780	0.973	0.858

Scores set evaluación

	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.828	0.830	0.650	0.875	0.813
Recall	0.759	0.751	0.725	0.653	0.617
Precision	0.592	0.596	0.361	0.756	0.579
F1	0.665	0.665	0.482	0.701	0.598
AUC	0.804	0.802	0.677	0.796	0.744

Resumen:

Con la imputación de los datos faltantes y el correspondiente balanceo, se obtiene para los modelos más significativos un *ROC* de:

- Logistic Regression: 80.3 %
- LDA: 80.2 %
- Random Forest: 79.0 %

En forma general, se obtuvieron mejoras en los *scores* de cada modelo abordado. Sigue siendo evidente un porcentaje de *overfitting* en algunos de ellos, el más notorio se da para Random Forest, para el cuál el *ROC* es de **0.97** en el entramiento y luego disminuye a **0.79** en las pruebas.

9 Variables Codificadas

En este set de datos aplican las transformaciones explicadas en la sección **Tratamiento de variables**.

Además, se experimenta con el balanceo del conjunto de entrenamiento mediante sub-sampleo de los registros de la clase mayoritaria. A diferencia de los entrenamientos anteriores, se reduce considerablemente, en un **80%** la cantidad de datos en el set de *train*.

```
Proporción set balanceado:
1    9919
0    9919
Name: RainTomorrow, dtype: int64
```

Scores CV set de entrenamiento

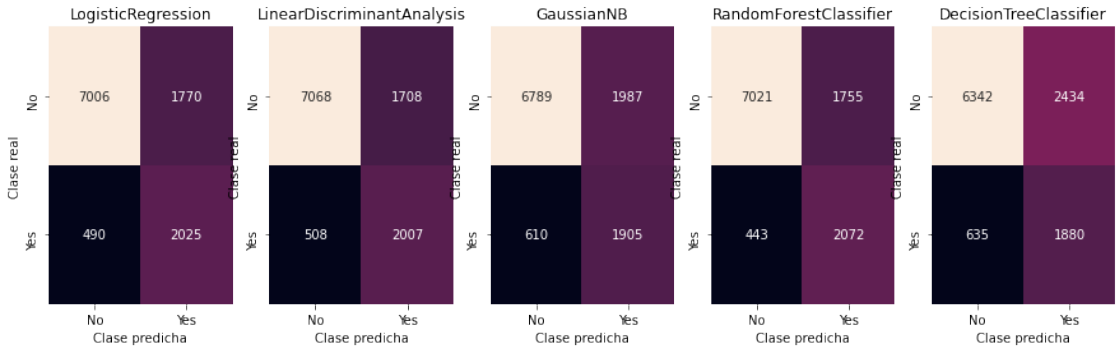
	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.802	0.800	0.768	0.813	0.738
Recall	0.790	0.783	0.755	0.820	0.730
Precision	0.809	0.811	0.776	0.813	0.736
F1	0.799	0.797	0.765	0.817	0.738
AUC	0.886	0.886	0.850	0.897	0.739

Scores set evaluación

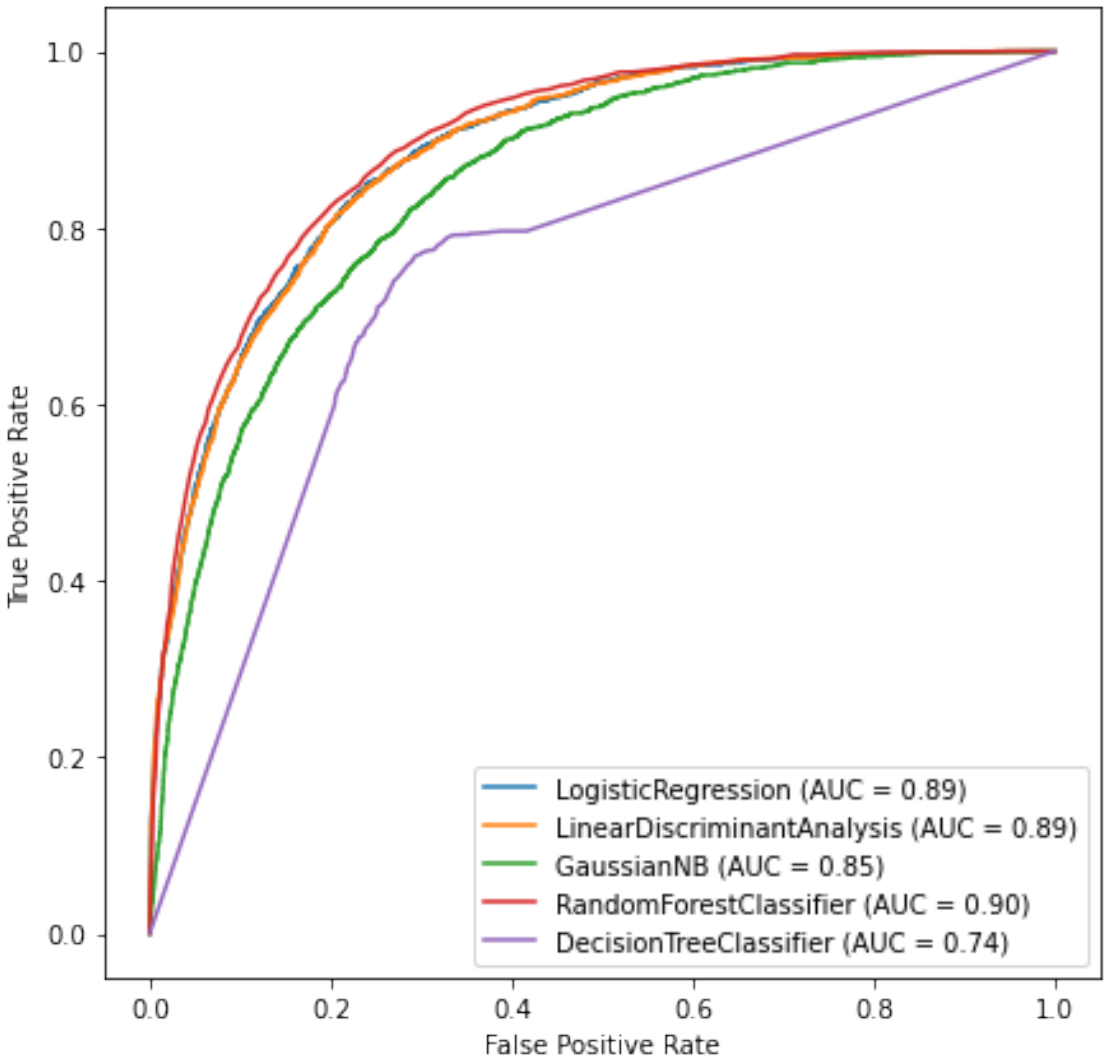
	Logistic	LDA	GaussianNB	RandomForest	DecisionTree
Accuracy	0.800	0.804	0.771	0.804	0.742
Recall	0.807	0.800	0.753	0.820	0.725
Precision	0.534	0.540	0.491	0.539	0.450
F1	0.643	0.645	0.595	0.651	0.556
AUC	0.803	0.802	0.765	0.810	0.736

- Balanceando el data set por sub sampleo de la clase mayoritaria tiene un costo en el *accuracy*.
- Vemos un gran aumento del *recall* en el set de testing en todos los modelos, pasamos de valores del orden de **65%** a **82%**.
- Random Forest se ubica como el mejor clasificador en todas las métricas.
- Ya no se observa *ovefitting*.

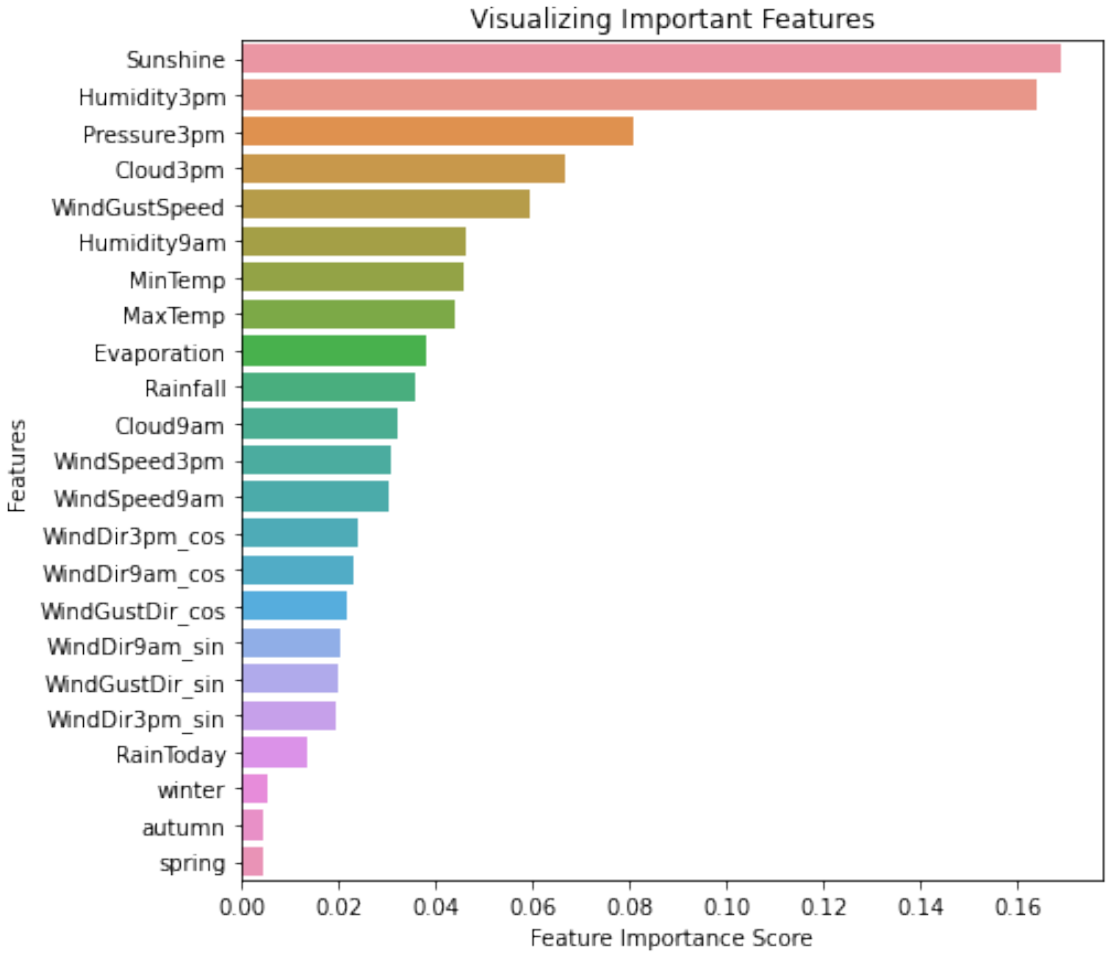
La figura siguiente muestra las matrices de confusión de los modelos entrenados.



- Resultaron solamente **443** falsos negativos para Random Forest.



- Las curvas *ROC* confirman el mejor desempeño de RF.
- *DecisionTree* tiene el peor desempeño.



- La variable más importante en este caso es *Sunshine*, es decir, el valor proporcional de exposición solar en el día. Ya que esta es una variable con un porcentaje alto de datos vacíos, este puede ser el motivo de la corrección de este problema.

9.0.1 Resumen de resultados Variables Codificadas

- A pesar de la gran reducción de tamaño del set de entrenamiento, se obtuvieron resultados muy favorales.
- Siendo el RandomForest el modelo de mejor desempeño en todas las pruebas encuanto al *AUC* con *0.81* según la tabla de test.
- En cuanto al *recall*, métrica muy importante para un clasificador más cauteloso, se observa una mejoría muy grande en cuánto al *recall* del clasificador **0.82**.
- Se soluciona el problema de *overfitting* en el RF. La hipotesis es que al ser un método iterativo, se ve afectado por la generación de datos sintéticos por parte del *Smote-Tomek*.
- Cabe destacar que la regresión logística tiene un desempeño muy bueno, teniendo la ventaja de que el modelo es interpretable.

Haciendo uso de las variables más significativas según el último RF. Creamos un modelo de regresión

con ellas.

10 Regresión Logística: Variables más significativas

Las variables elegidas son las 5 más importantes

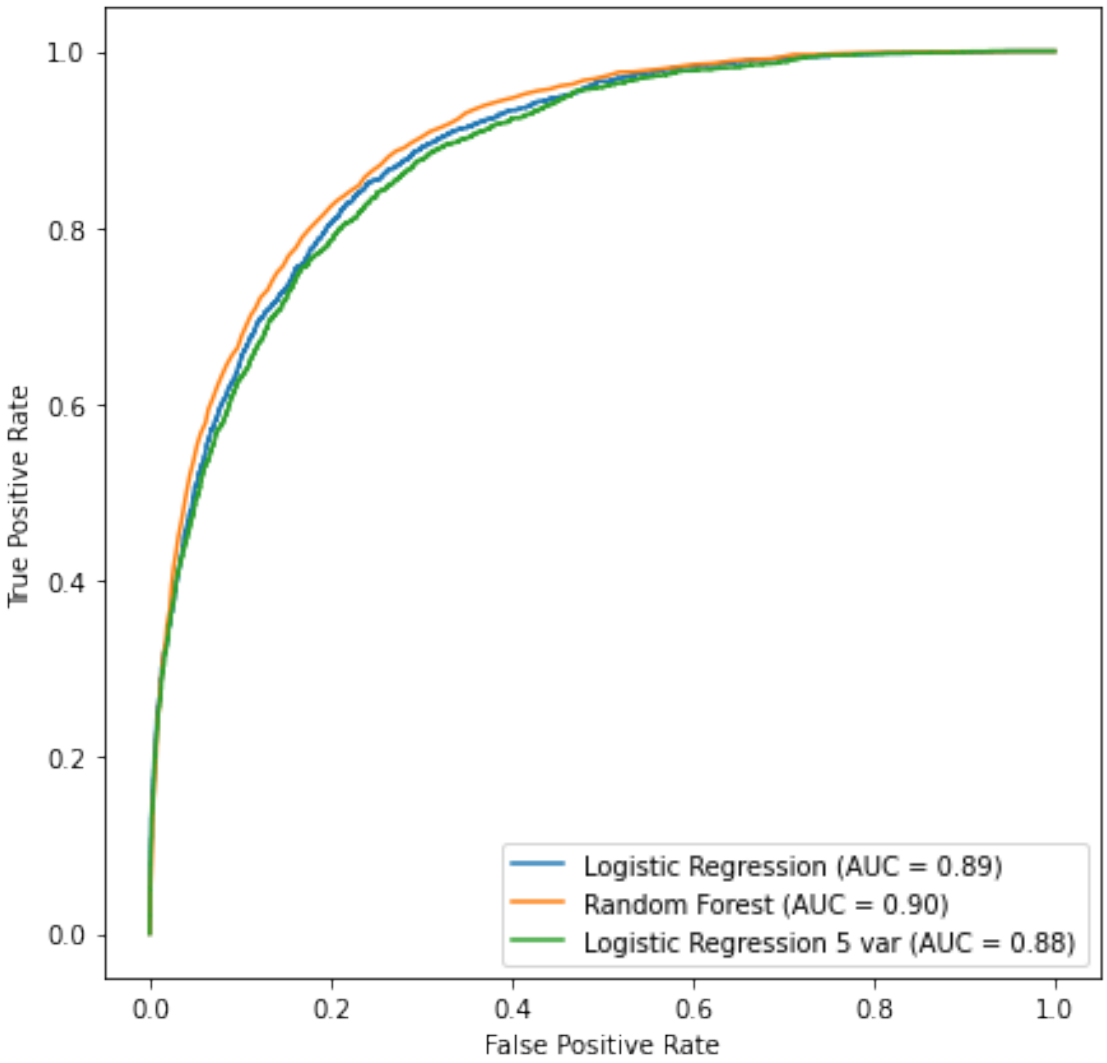
- Sunshine
- Humidity3pm
- Pressure3pm
- Cloud3pm
- WindGustSpeed

Scores CV set de entrenamiento

	Logistic	RandomForest	Logistic 5
Accuracy	0.801	0.811	0.793
Recall	0.791	0.815	0.789
Precision	0.807	0.807	0.796
F1	0.799	0.813	0.792
AUC	0.885	0.894	0.878

Scores set de evaluación

	Logistic	RandomForest	Logistic 5
Accuracy	0.800	0.805	0.793
Recall	0.805	0.824	0.795
Precision	0.534	0.541	0.524
F1	0.642	0.653	0.632
AUC	0.802	0.812	0.794



- Se oberva una pérdida de desempeño muy pequeña respecto de la simplificación del modelo realizada.
- Siguiendo el principio de parsimonia, se gana muchísima interpretación con este modelo.

10.1 Interpretación

	Coeficiente
Sunshine	-0.602
Humidity3pm	1.091
Pressure3pm	-0.557
Cloud3pm	0.299
WindGustSpeed	0.558

- Podemos concluir que un aumento de '*Sunshine*' y '*Pressure3pm*' tienen un efecto negativo

sobre la probabilidad de lluvia al día siguiente.
- En cambio, si '*Humidity3pm*', '*WindGustSpeed*' y '*Cloud3pm*' aumentan, también lo hace probabilidad de lluvia, en el respectivo orden de importancia.

11 Conclusiones

De acuerdo al análisis realizado, y los resultados de cada uno de los modelos abordados, se decide que puede crearse un modelo con las variables disponibles a fines de predecir si lloverá al día siguiente. Las métricas obtenidas para los últimos modelos ensayados son buenas, permitiendo predecir la clase de interés con un alto índice de éxito.

El modelo *Random Forest* de la sección **Variables Condificadas** logra el mejor *ROC*, con un índice de **0.81**. Cabe resaltar que tanto el *Recall* como la *Precision*, adoptan valores de un orden similar al comparar los mejores modelos (RF, RL, LDA). Por lo tanto, a la hora de discriminar si se prefieren más errores en un falso positivo para lluvia, o, viceversa, optar por un modelo en específico no acarrearía ventajas significativas.

Además, se han podido experimentar muchas variantes de conjunto de datos, siendo la más efectiva, en este caso, la de realizar un subsampleo de la clase mayoritaria, descartando los valores nulos.

Se consiguió también construir un modelo simplificado con las cinco variables más importantes, logrando resultados comparables a los modelos completos.