CODER HOUSE

Proyecto Final Data Science

“Modelo de clasificación para predecir lluvias en forma temprana”

Primera entrega

Alumno: Foletto Mariano

Comisión: 90455

Profesor: Jorge Ruiz

Tutor: Luciano Lisachi

Problemática a resolver

Analizar datos históricos de lluvias, periodo 1/11/2007 hasta 25/6/2017, y lograr predecir en forma temprana si hay lluvia al día siguiente. Se puede suponer que se trabaja para un privado y/o estatal, que necesite prever las lluvias para la toma de decisiones.

Solución propuesta

Se propone realizar un modelo de clasificación binaria y/o regresión, realizando previamente un análisis de los datos persistidos en la base, tanto de la integridad de estos como de su poder predictivo para la variable objetivo a inferir. Se seleccionará aquel modelo que no solo brinde el mayor accuracy sobre el set de validación, sino que se prestará especial atención en el valor del recall buscando que la misma tienda a 1 (uno) o -lo que es lo mismo- que los falsos negativos tiendan a cero.

$$Recall = TP \,/\, (TP + FN)$$

$$Accuracy = TP + TN \,/\, (TP + TN + FN + FP)$$

Modelado

Data Acquisition

Una vez obtenido el acceso a la base, se volcó el contenido de esta a un archivo .csv el cuál se encuentra en github (Link). El usuario de los datos en cuestión nos brindó una descripción de las variables presentes en el mismo, como así también el dominio o rango esperado de sus variables numéricas, en caso de que haya algún tipo de error en la base de datos. En una primera etapa se le prestará especial atención a aquellas variables que contengan valores considerados absurdos e imposibles para un ser humano.

Descripción de variables y dominio

- **Date:** Fecha de la observación [1/11/2007 hasta 25/6/2017].
- **Location:** Nombre común de la ubicación de la estación meteorológica [49 location].
- **MinTemp:** Temperatura mínima en grados Celsius [rango de -8,5 a 33,9].
- **MaxTemp:** Temperatura máxima en grados Celsius [rango de -4,8 a 48,1].
- **Rainfall:** Cantidad de lluvia registrada en el día (mm) [rango de 0 a 371].
- **Evaporation:** Evaporación del tipo Clase A (mm) en las 24 horas hasta las 9am [rango de 0 a 145].
- **Sunshine:** Número de horas de sol brillante en el día [rango de 0 a 14,5].
- **WindGustDir:** Dirección de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche [E, ENE, ESE, N, NE, NNE, NNW, NW, S, SE, SSE, SSW, SW, W, WNW, WSW].
- **WindGustSpeed:** Velocidad (km/h) de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche [rango de 6 a 135].
- **WindDir9am:** Dirección del viento a las 9am [E, ENE, ESE, N, NE, NNE, NNW, NW, S, SE, SSE, SSW, SW, W, WNW, WSW].
- **WindDir3pm:** Dirección del viento a las 3pm [E, ENE, ESE, N, NE, NNE, NNW, NW, S, SE, SSE, SSW, SW, W, WNW, WSW].
- **WindSpeed9am:** Velocidad del viento (km/h) promedio en los 10 minutos previos a las 9am [rango de 0 a 130].
- **WindSpeed3pm:** Velocidad del viento (km/h) promedio en los 10 minutos previos a las 3pm [rango de 0 a 87].
- **Humidity9am:** Humedad (porcentaje) a las 9am [rango de 0 a 100].
- **Humidity3pm:** Humedad (porcentaje) a las 3pm [rango de 0 a 100].
- **Pressure9am:** Presión atmosférica (hPa) reducida al nivel del mar a las 9am [rango de 980,5 a 1041].
- **Pressure3pm:** Presión atmosférica (hPa) reducida al nivel del mar a las 3pm [rango de 977,1 a 1039,6].
- **Cloud9am:** Fracción del cielo cubierta por nubes a las 9am (en oktas: octavos) [rango de 0 a 9].
- **Cloud3pm:** Fracción del cielo cubierta por nubes a las 3pm (en oktas: octavos) [rango de 0 a 9].
- **Temp9am:** Temperatura (°C) a las 9am [rango de -7,2 a 40,2].
- **Temp3pm:** Temperatura (°C) a las 3pm [rango de -5,4 a 46,7].
- **RainToday:** Booleano: 1 si la precipitación (mm) en las 24 horas hasta las 9am supera 1mm, de lo contrario 0 [posee "Yes" y "No" según corresponda precipitación mayor a 1mm].
- **RainTomorrow:** Cantidad de lluvia del día siguiente (mm), usada para crear la variable de respuesta RainTomorrow [posee "Yes" y "No" según corresponda precipitación mayor a 1mm].

Estadística descriptiva

- El dataset obtenido de la base de datos posee 145.560 filas y 23 columnas, de las cuales 22 (veintidós) son variables independientes para analizar y 1 (una) corresponde a la variable target o dependiente.

- El dataset no ni filas repetidas.

- El dataset posee valores nulos, en el dataset figuran como "NA". Estos se tratarán posteriormente.

- De las 23 (veintitrés) variables independientes, 6 (seis) son categóricas (Date, Location, WindGustDir, WindGustDir, WindDir9am, WindDir3pm) y 16 (dieciséis) son numéricas (MinTemp, MaxTemp, Rainfall, Evaporation, Sunshine, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3m).
  La variable RainTomorrow (target) y RainToday ya se encuentran binarizada.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 145460 entries, 0 to 145459
Data columns (total 23 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Date           145460 non-null  object
 1   Location       145460 non-null  object
 2   MinTemp        143975 non-null  float64
 3   MaxTemp        144199 non-null  float64
 4   Rainfall       142199 non-null  float64
 5   Evaporation    82670 non-null   float64
 6   Sunshine       75625 non-null   float64
 7   WindGustDir    135134 non-null  object
 8   WindGustSpeed  135197 non-null  float64
 9   WindDir9am     134894 non-null  object
 10  WindDir3pm     141232 non-null  object
 11  WindSpeed9am   143693 non-null  float64
 12  WindSpeed3pm   142398 non-null  float64
 13  Humidity9am    142806 non-null  float64
 14  Humidity3pm    140953 non-null  float64
 15  Pressure9am    130395 non-null  float64
 16  Pressure3pm    130432 non-null  float64
 17  Cloud9am       89572 non-null   float64
 18  Cloud3pm       86102 non-null   float64
 19  Temp9am        143693 non-null  float64
 20  Temp3pm        141851 non-null  float64
 21  RainToday      142199 non-null  object
 22  RainTomorrow   142193 non-null  object
dtypes: float64(16), object(7)
memory usage: 25.5+ MB
```

Figura 1: Información del dataset.

	count	unique	top	freq
Date	145460	3436	2017-06-24	49
Location	145460	49	Canberra	3436
WindGustDir	135134	16	W	9915
WindDir9am	134894	16	N	11758
WindDir3pm	141232	16	SE	10838
RainToday	142199	2	No	110319
RainTomorrow	142193	2	No	110316

Figura 2: Estadística descriptiva básica de las variables categóricas.

	count	mean	std	min	25%	50%	75%	max
MinTemp	143975.0	12.194034	6.398495	-8.5	7.6	12.0	16.9	33.9
MaxTemp	144199.0	23.221348	7.119049	-4.8	17.9	22.6	28.2	48.1
Rainfall	142199.0	2.360918	8.478060	0.0	0.0	0.0	0.8	371.0
Evaporation	82670.0	5.468232	4.193704	0.0	2.6	4.8	7.4	145.0
Sunshine	75625.0	7.611178	3.785483	0.0	4.8	8.4	10.6	14.5
WindGustSpeed	135197.0	40.035230	13.607062	6.0	31.0	39.0	48.0	135.0
WindSpeed9am	143693.0	14.043426	8.915375	0.0	7.0	13.0	19.0	130.0
WindSpeed3pm	142398.0	18.662657	8.809800	0.0	13.0	19.0	24.0	87.0
Humidity9am	142806.0	68.880831	19.029164	0.0	57.0	70.0	83.0	100.0
Humidity3pm	140953.0	51.539116	20.795902	0.0	37.0	52.0	66.0	100.0
Pressure9am	130395.0	1017.649940	7.106530	980.5	1012.9	1017.6	1022.4	1041.0
Pressure3pm	130432.0	1015.255889	7.037414	977.1	1010.4	1015.2	1020.0	1039.6
Cloud9am	89572.0	4.447461	2.887159	0.0	1.0	5.0	7.0	9.0
Cloud3pm	86102.0	4.509930	2.720357	0.0	2.0	5.0	7.0	9.0
Temp9am	143693.0	16.990631	6.488753	-7.2	12.3	16.7	21.6	40.2
Temp3pm	141851.0	21.683390	6.936650	-5.4	16.6	21.1	26.4	46.7

Figura 3: Estadística descriptiva básica de las variables numéricas.

Data Wrangling y EDA

Tratamiento de valores nulos

En esta primera iteración no las descarto y le imputo valores en el código posterior (otra posibilidad a evaluar es descartando estas variables que tienen muchos faltantes: Sunshine, Evaporation, Cloud3pm y Cloud9am).

```
Sunshine         48.01
Evaporation      43.17
Cloud3pm         40.81
Cloud9am         38.42
Pressure9am      10.36
Pressure3pm      10.33
WindDir9am        7.26
WindGustDir       7.10
WindGustSpeed     7.06
Humidity3pm       3.10
WindDir3pm        2.91
Temp3pm           2.48
RainTomorrow      2.25
Rainfall          2.24
RainToday         2.24
WindSpeed3pm      2.11
Humidity9am       1.82
WindSpeed9am      1.21
Temp9am           1.21
MinTemp           1.02
MaxTemp           0.87
Date              0.00
Location          0.00
dtype: float64
```

Figura 4: Análisis del porcentaje de faltante en cada columna

Transformación de la información: primero se pasó la columna "date" a formato fecha. Se agrego la columna mes para agrupar por posible estacionalidad y usarla para la imputación de valores numéricos.

Estrategia de imputación:

- ✓ Numéricas por mediana de grupo
- ✓ Categóricas por moda de grupo (excepto direcciones)
- ✓ Relleno residual (si aún quedó algo, por ejemplo, direcciones del viento)

Variables numéricas

Se dispuso a graficar la distribución y el boxplot de cada una, habiendo filtrado previamente los valores absurdos o irreales encontrados mediante el análisis descriptivo. Se adjuntan gráficos de aquellas que posean outliers su posterior interpretación.

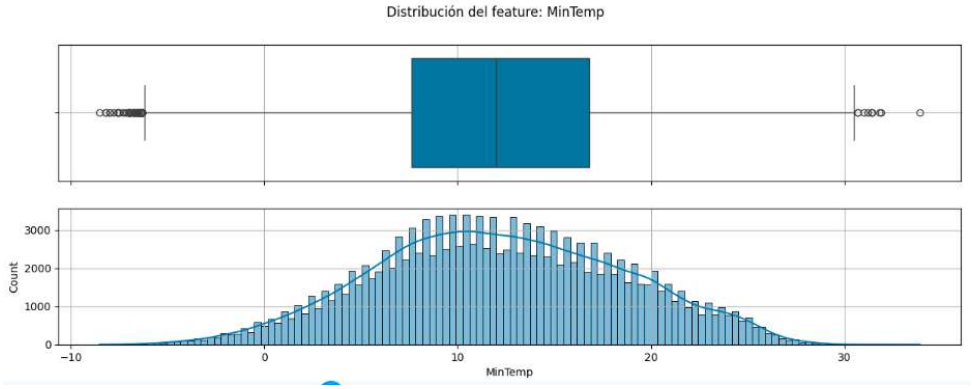

Figura 5

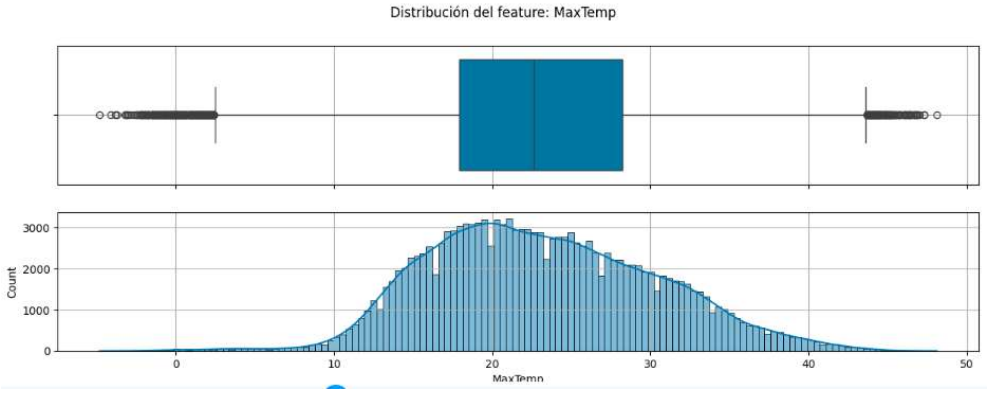

Figura 6

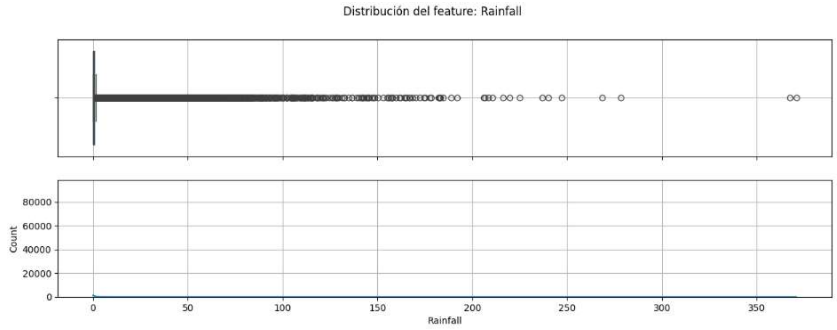

Figura 7

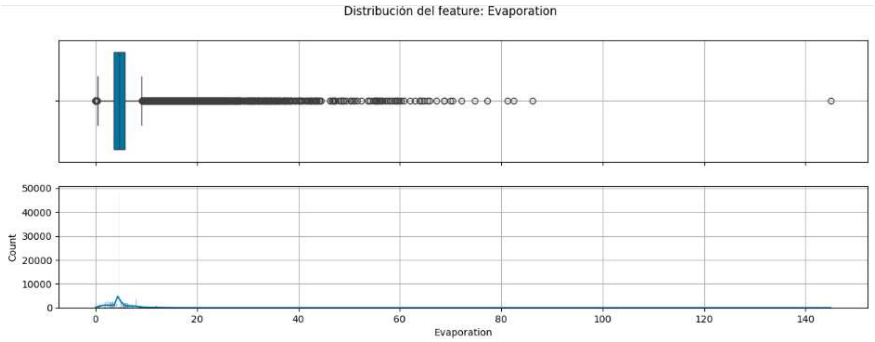

Figura 8

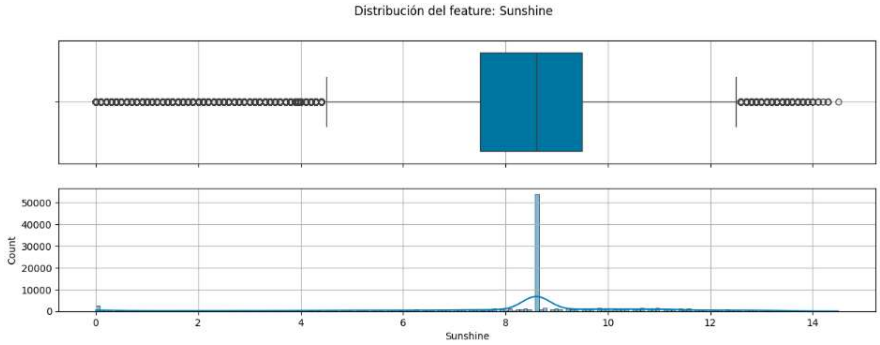

Figura 9

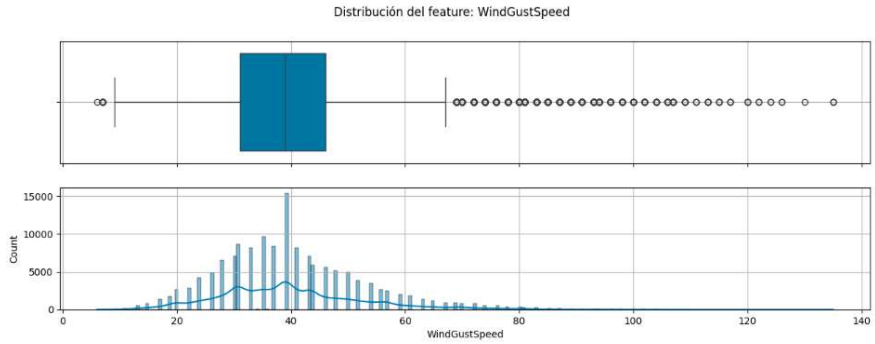

Figura 10

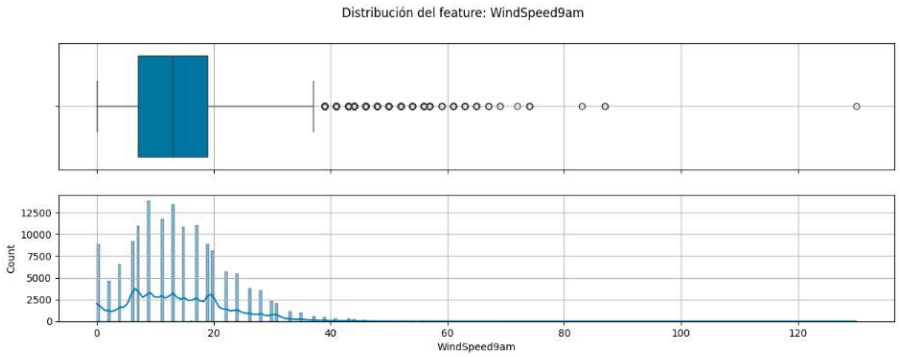

Figura 11

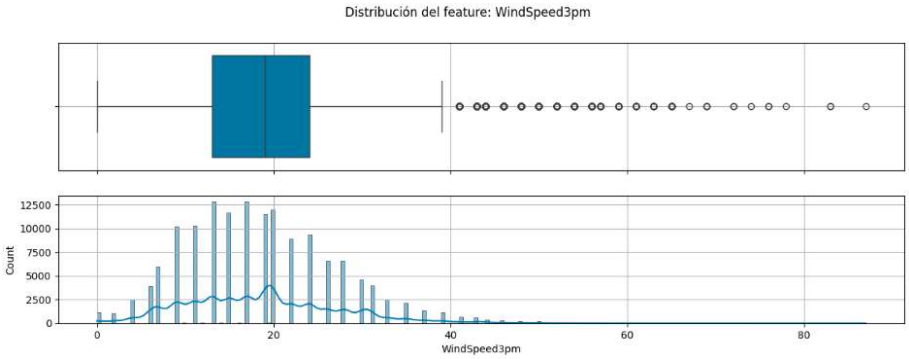

Figura 12

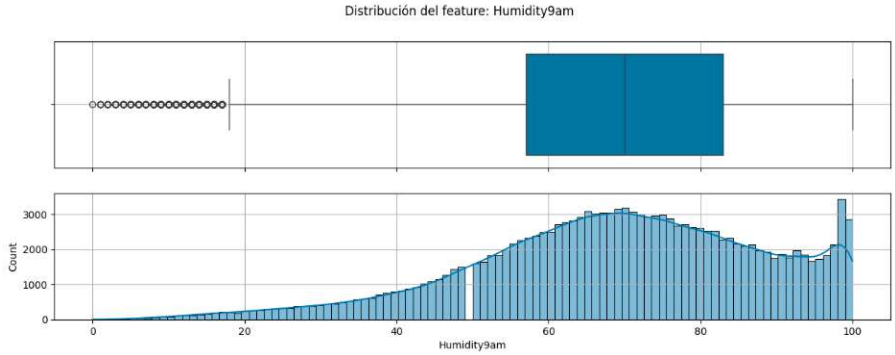

Figura 13

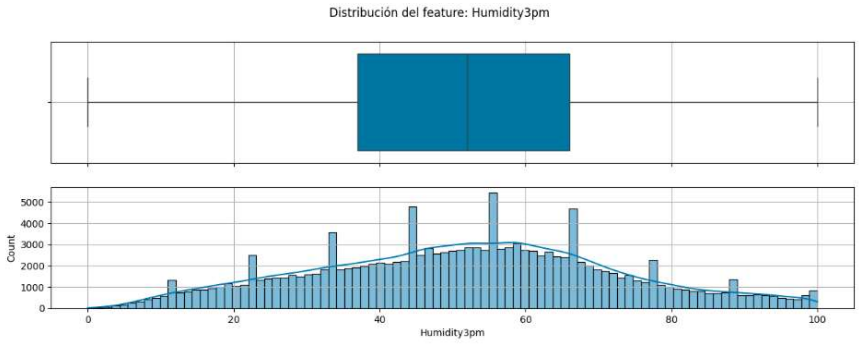

Figura 14

Correlación sin encodear (variables numéricas)

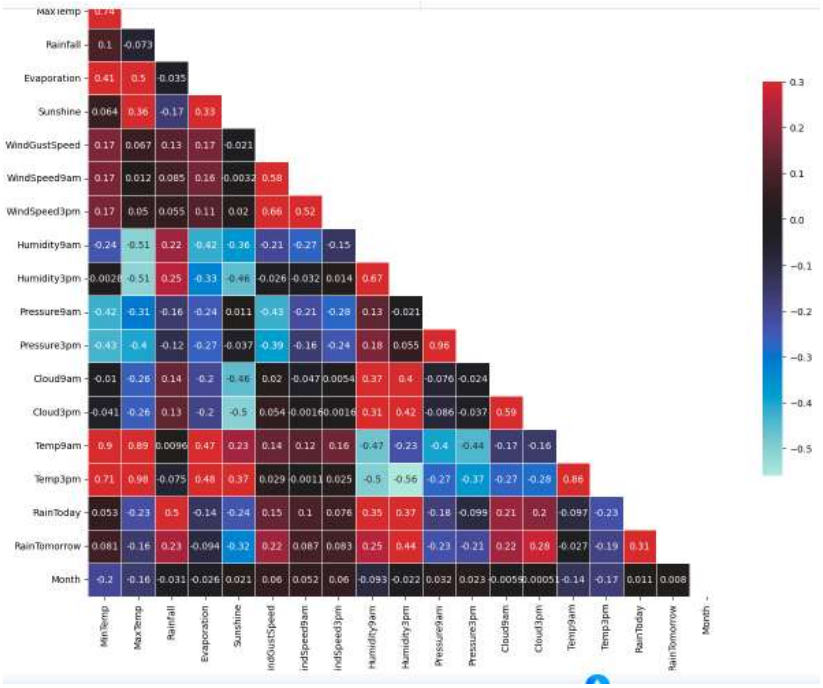

Figura 15: Correlación de Pearson utilizando solo variables numéricas.

Puede observarse que -en módulo- las correlaciones de mayor valor se dan entre la variable target y Rainfall, WindGustSpeed, Humidity9am, Humidity3pm, Cloud9am, Cloud3pm y RainToday.

Feature Engineering

En una primera iteración, se dispuso a encodear las variables categóricas ordinales en enteros representativos y las nominales se binarizaron en 0 y 1, quedando el dataset de la siguiente manera:

	Location	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	Humidity9am	Humidity3pm	...	Cloud9am	Cloud3pm	T
0	62	0.6	4.6	8.6	270.0	44.0	270.0	292.5	71.0	22.0	...	8.0	5.0	
1	62	0.0	4.6	8.6	292.5	44.0	337.5	247.5	44.0	25.0	...	7.0	5.0	
2	62	0.0	4.6	8.6	247.5	46.0	270.0	247.5	38.0	30.0	...	7.0	2.0	
3	62	0.0	4.6	8.6	45.0	24.0	135.0	90.0	45.0	16.0	...	7.0	5.0	
4	62	1.0	4.6	8.6	270.0	41.0	67.5	315.0	82.0	33.0	...	7.0	8.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
145455	23	0.0	4.6	8.6	90.0	31.0	135.0	67.5	51.0	24.0	...	6.0	3.5	
145456	23	0.0	4.6	8.6	337.5	22.0	135.0	0.0	56.0	21.0	...	6.0	3.5	
145457	23	0.0	4.6	8.6	0.0	37.0	135.0	292.5	53.0	24.0	...	6.0	3.5	
145458	23	0.0	4.6	8.6	135.0	28.0	157.5	0.0	51.0	24.0	...	3.0	2.0	
145459	23	0.0	4.6	8.6	90.0	34.0	112.5	112.5	62.0	36.0	...	8.0	8.0	

145460 rows × 21 columns

Figura 16: Dataset a utilizar en una primera iteración con variables categóricas encodeadas.

El dataset final a utilizar en la primera iteración cuenta con 113.516 filas correspondientes a días sin lluvias en el día de mañana y 31.944 filas correspondientes a días que si hubo lluvia en día posterior. Se dispuso a balancear las clases en el dataset, de modo tal de contar con las 31.944 filas de días de lluvias al día posterior, y otras 31.944 filas elegidas al azar entre las 113.516 de días sin lluvias. El dataset a utilizar para entrenar los modelos posee finalmente 63.888 filas y 21 columnas:

	Location	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	WindDir3pm	Humidity9am	Humidity3pm	...	Cloud9am	Cloud3pm
37921	60	0.0	8.8	13.2	67.5	39.0	67.5	67.5	50.0	21.0	...	5.0	2.0
37871	60	0.0	5.0	11.7	292.5	54.0	67.5	315.0	49.0	15.0	...	0.0	1.0
52220	82	0.0	4.6	8.6	135.0	24.0	157.5	247.5	26.0	54.0	...	6.0	5.0
40081	78	0.0	1.8	7.8	292.5	31.0	292.5	315.0	73.0	44.0	...	1.0	3.0
43143	71	0.0	4.6	8.6	180.0	43.0	202.5	157.5	59.0	57.0	...	8.0	6.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
94184	52	0.0	3.6	10.3	45.0	35.0	157.5	67.5	52.0	54.0	...	0.0	1.0
66098	73	0.0	2.2	5.0	0.0	61.0	0.0	0.0	83.0	69.0	...	4.0	7.0
4171	65	0.2	4.6	8.6	202.5	37.0	225.0	225.0	69.0	82.0	...	6.0	5.0
74227	48	0.4	4.6	8.6	270.0	46.0	22.5	292.5	88.0	90.0	...	6.0	5.0
99342	69	0.0	8.7	12.6	22.5	30.0	45.0	67.5	34.0	18.0	...	6.0	5.0

63888 rows × 21 columns

Figura 17: Dataset balanceado listo para utilizar en el modelado.

Correlación con variables encodeadas

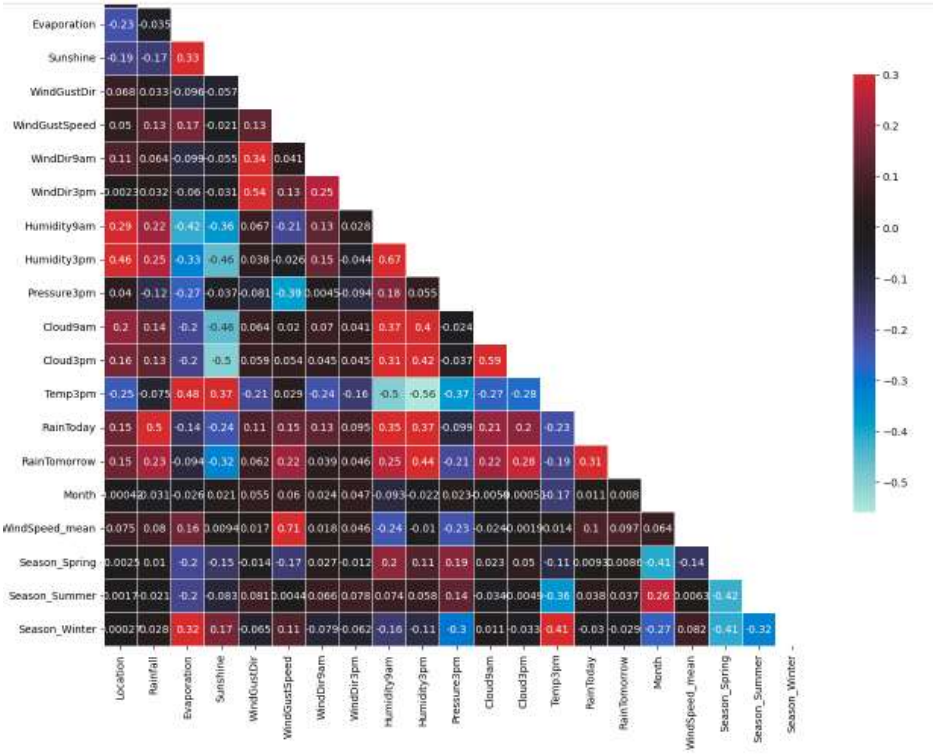

Figura 18: Correlación de Pearson utilizando todas las variables.

Baseline

Como primera iteración se entrenarán dos modelos: un Árbol de Decisión clásico y un Random Forest, ambos sin optimizar hiperparámetros ni aplicar validación cruzada. Se generarán tres (3) conjuntos de datos: uno para entrenamiento (train), otro para validación (val) y otro para testeo (test) del mejor modelo. El conjunto de prueba se reservará y no será utilizado hasta haber seleccionado el modelo más adecuado mediante el conjunto de validación.

Test - Train - Validation Split

Se destinará un 80% del dataset para entrenamiento y un 20% para el testeo. El 10% del set de entrenamiento será destinado para el set de validación.

Métricas

El modelo que mejor performó fue el Random Forest con su configuración estándar, obteniéndose un accuracy de 0.78 y un recall de 0.74 para la clase 1.

```
% de aciertos sobre el set de test: 0.7834559398966975

              precision    recall  f1-score   support

           0       0.76      0.82      0.79      6332
           1       0.81      0.74      0.78      6446

    accuracy                           0.78     12778
   macro avg       0.79      0.78      0.78     12778
weighted avg       0.79      0.78      0.78     12778
```

Figura 19: Classification report del modelo.

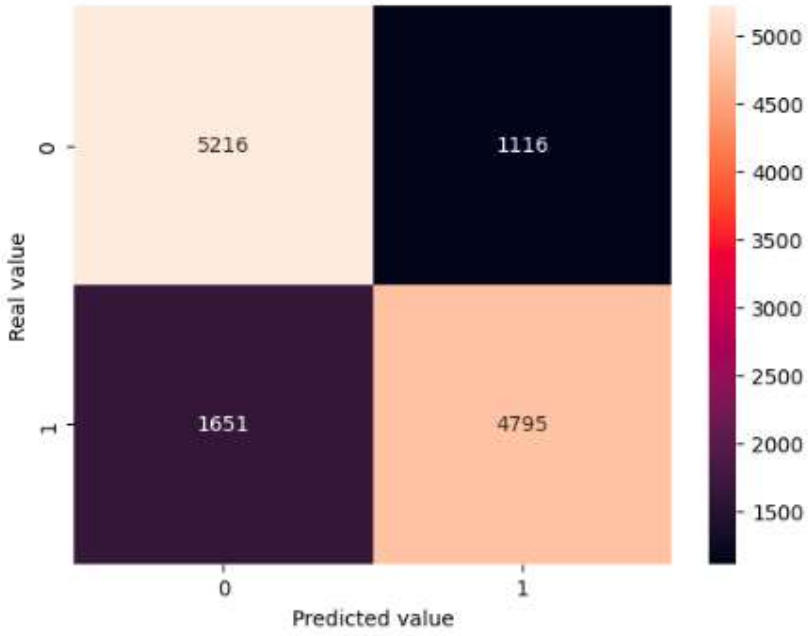

Figura 20: Matriz de confusión obtenida.

Puede observarse que los falsos positivos (FP), es decir, los valores que fueron predichos como positivos pero que en realidad no lo son, son menores a los falsos negativos (FN) que son los valores que fueron predichos como negativos, pero en realidad son positivos. Si bien esto no condice exactamente con lo buscado para la clase 1, será mejorada en las próximas iteraciones.

Conclusiones primera iteración

El presente trabajo fue realizado bajo el concepto de estudio de factibilidad, realizando una única iteración sobre las variables elegidas, los motivos y la escueta selección de modelos. Aun así, se concluye que la clasificación es factible de ser realizada, habiendo alcanzado valores de métricas altos, susceptibles de ser utilizados.