Ariel Salassa

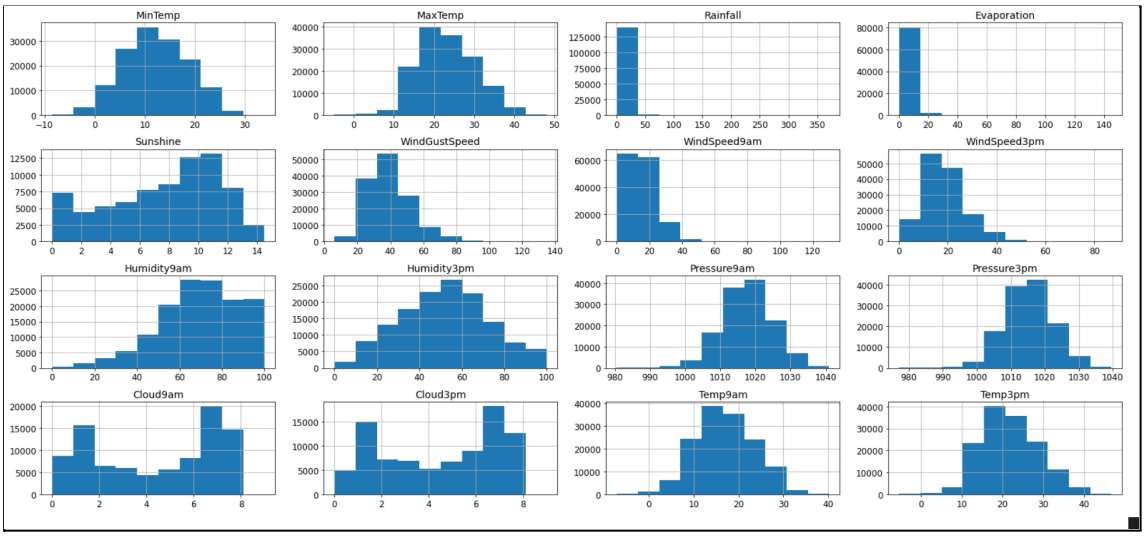
Trabajo Práctico Integrador ADD & Intro IA

1. Análisis exploratorio inicial

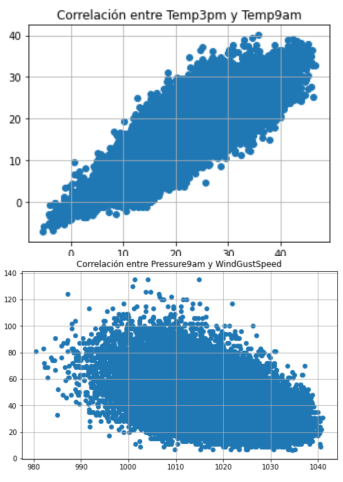
En esta etapa se analizó en dataset: detección de features y label. Se clasificaron los features y el label por tipo (categórico, numérico, compuesto).

Variables numéricas

Se realizaron histogramas para tener un pantallazo de su distribución.

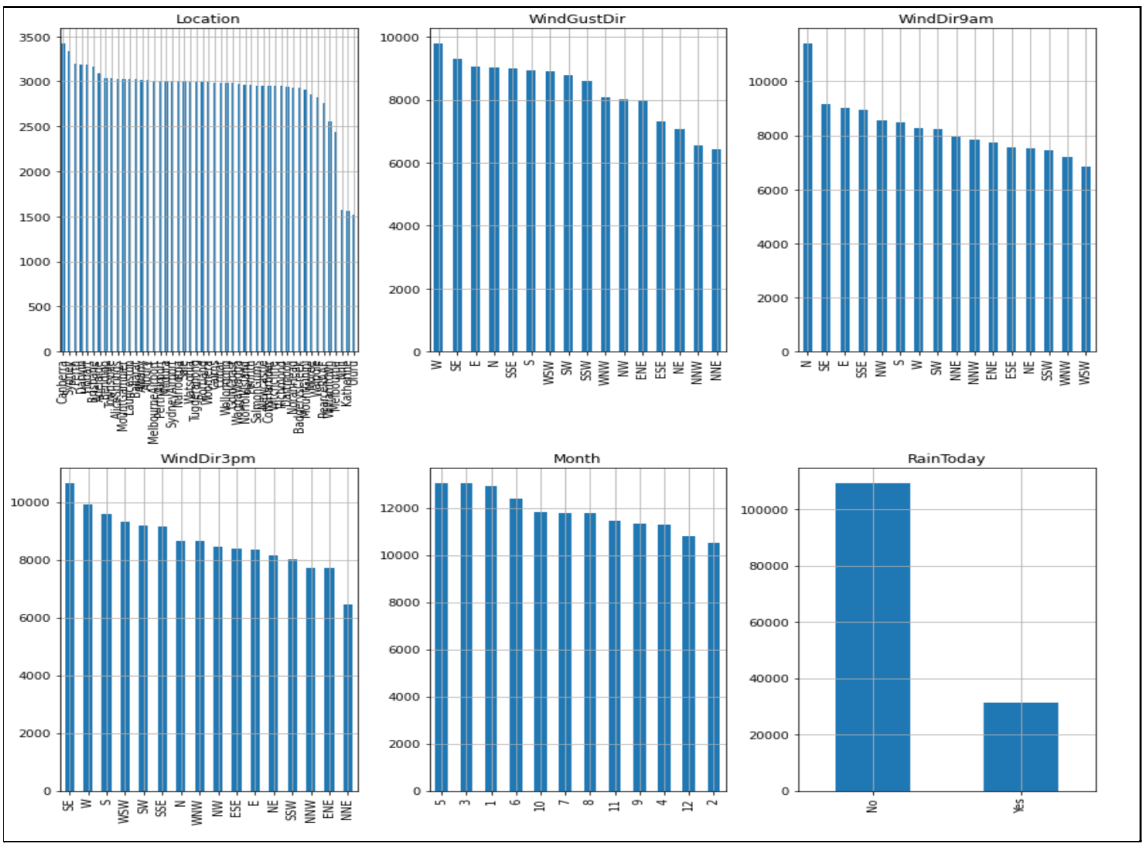


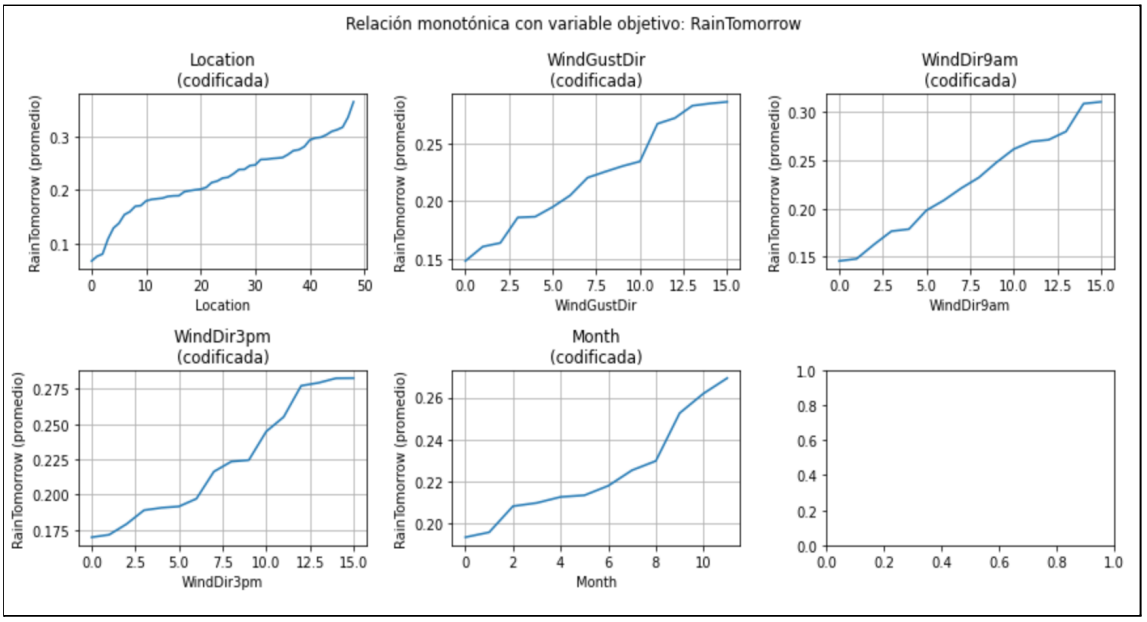
Se analizó en una primera instancia la correlación entre ellas con la matriz de correlación y con algunas nubes de puntos.



Variables categóricas

Se analizaron los distintos valores que pueden tomar las variables categóricas y se dio una primera visibilidad de la cardinalidad de cada uno de los posibles valores. Además se analizó si había una relación monotónica entre los distintos valores de cada categoría y la variable objetivo (excluyendo las variables categóricas de sólo dos categorías)





Variables compuestas

La única variable compuesta es la fecha, para la cual se estudió el rango.

Variables de salida

Se corresponde con la columna *'RainTomorrow'* la cual es categórica binaria (con valores *Yes/No)* y está desbalanceada. Aproximadamente el 76% de los datos toman valor `*No*` y el 24% restante toma el valor `*Yes*`.

La otra variable se corresponde con la columna *'RainfallTomorrow'* la cual es numérica.

2. Limpieza inicial

En primer lugar, se borraron las filas que no tenían label como así también las que no tenían definido `*RainToday*`. En segundo lugar, se analizó la completitud de datos por tipo de variable. Dentro de las variables numéricas los resultados fueron los siguientes:

	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustSpeed	WindSpeed9am	WindSpeed3pm	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm
0	637.00000	322.000000	1406.000000	60843.00000	67816.000000	9270.000000	1348.00000	2630.000000	1774.000000	3610.000000	14014.000000	13981.000000	53657.000000	57094.000000	904.000000	2726.000000
1	0.00448	0.002265	0.009888	0.42789	0.476929	0.065193	0.00948	0.018496	0.012476	0.025388	0.098556	0.098324	0.377353	0.401525	0.006358	0.019171

Dado que hay datos faltantes en la mayoría de las columnas se pueden tomar dos caminos: descartar las filas con datos faltantes o descartar las columnas con gran cantidad de datos faltantes (*Evaporation, Sunshine, Cloud9am, Cloud3pm*) e implementar técnicas de imputación para el resto. Para enriquecer el análisis se tomó el segundo enfoque.

3. Esquema de validación de resultados

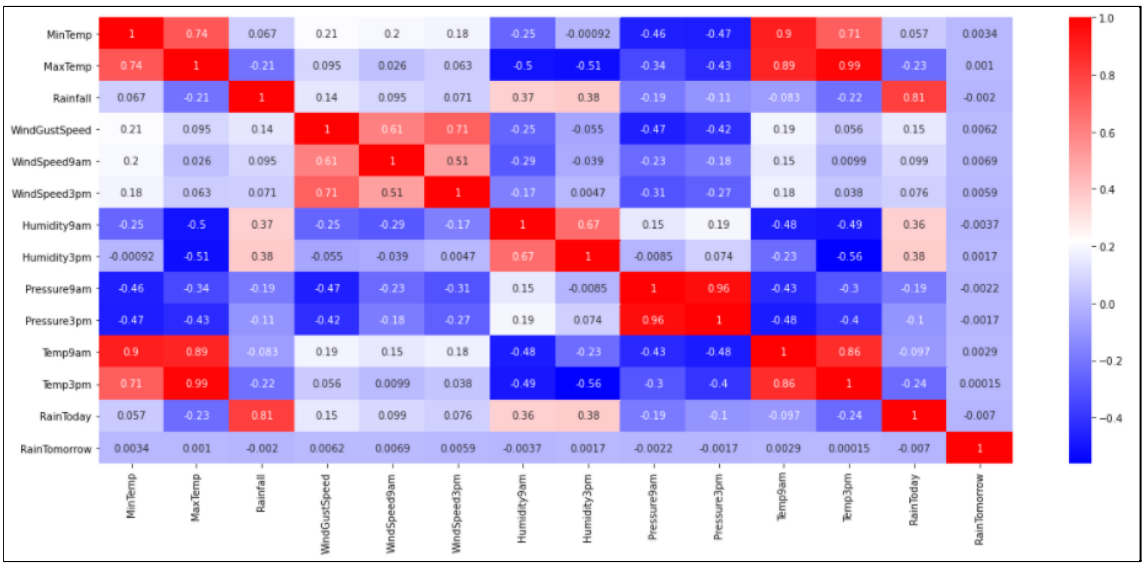
El dataset se particiona en 80% de datos para el entrenamiento y 20% para el test.

4. Preparación de los datos e ingeniería de features

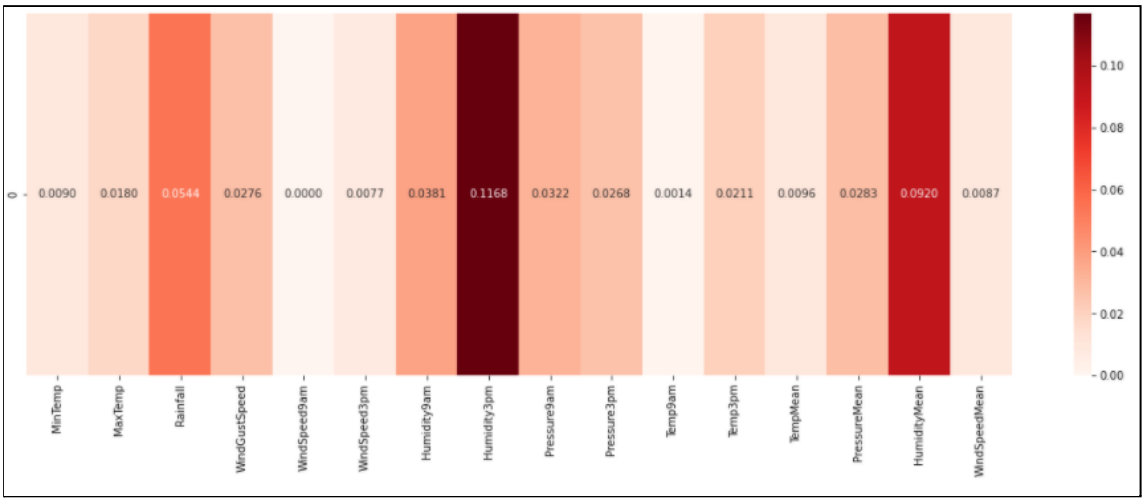
Codificación de la variables de dos clases: Tanto *RainTomorrow* como *RainToday* se macaron de *Yes/No* a *1/0* respectivamente.
Codificación de la fecha: Se extrajo el mes y el resto se descartó. Sobre el mes se realizará one hot encoding.
Codificación de la dirección del viento: Se transformaron las siglas de la dirección en valores angulares sexagesimales tomando como referencia 0º al este. Sobre el ángulo se hizo descomposición en seno y coseno.
Codificación de la ubicación: Se utilizó la API Nominatim de paquete geopy que devuelve la coordenadas de Latitud y Longitud según el nombre de la ciudad.
Imputación de datos: Teniendo una representación numérica de la mayor parte de las variables categóricas se optó por utilizar MICE para la imputación de los datos.

5. Selección de features

Para la selección de features se realizó nuevamente una matriz de correlación con las variables restantes, luego de realizar una normalización *yeo-johnson.*



Se observa que hay una alta correlación entre las variables relacionadas con temperaturas, viento y presión.
Para decidir qué variables usar y qué variables descartar se hizo un análisis de información mutua y se agregaron features “promedio” para determinar si aportan más información.



Con ambos gráficos el grupo de features numéricas reducido que se obtuvo fue el siguiente: *'Rainfall', 'WindGustSpeed', 'Humidity9am', 'Humidity3pm', 'Pressure9am', 'Pressure3pm', 'Temp3pm', 'WindSpeedMean'*
Junto con los datos categóricos transformados datos numéricos los sets de variables numéricas conformadas por:
1- Set numérico reducido*: 'Humidity3pm', 'Humidity9am', 'Lat', 'Lon', 'Pressure3pm', 'Pressure9am', 'Rainfall', 'Temp3pm', 'WindDir3pm_x', 'WindDir3pm_y', 'WindDir9am_x', 'WindDir9am_y', 'WindGustDir_x', 'WindGustDir_y', 'WindGustSpeed', 'WindSpeedMean'.*
2- Set numérico completo sin modificar*: 'Humidity3pm', 'Humidity9am', 'Lat', 'Lon', 'MaxTemp', 'MinTemp', 'Pressure3pm', 'Pressure9am', 'Rainfall', 'Temp3pm', 'Temp9am', 'WindDir3pm_x', 'WindDir3pm_y', 'WindDir9am_x', 'WindDir9am_y', 'WindGustDir_x', 'WindGustDir_y', 'WindGustSpeed', 'WindSpeed3pm', 'WindSpeed9am'.*
3- Set numérico completo, agregando promedios*: 'Humidity3pm', 'Humidity9am', 'HumidityMean', 'Lat', 'Lon', 'MaxTemp', 'MinTemp', 'Pressure3pm', 'Pressure9am', 'PressureMean', 'Rainfall', 'Temp3pm', 'Temp9am', 'TempMean', 'WindDir3pm_x', 'WindDir3pm_y', 'WindDir9am_x', 'WindDir9am_y', 'WindGustDir_x', 'WindGustDir_y', 'WindGustSpeed', 'WindSpeed3pm', 'WindSpeed9am', 'WindSpeedMean'.*
A todos estos sets se le agrega el feature *Month* codificado con *one hot encoder* y el feature *RainToday* ya codificado como se especificó anteriormente.

6. Entrenamiento de modelos

Para los entrenamientos se exploraron múltiples soluciones con cada set de features (fs):

1. Regresión logística (LR).
2. Regresión logística con pesos de clase (LR + cw) para mitigar el desbalance (como no dio mejores resultados que la regresión logística sólo se probó con el fs1).
3. Red neuronal (NN) de 5 capas ocultas (32,16,8,8,4).

A continuación se muestra una tabla con el resumen de los resultados aplicados al set de test:

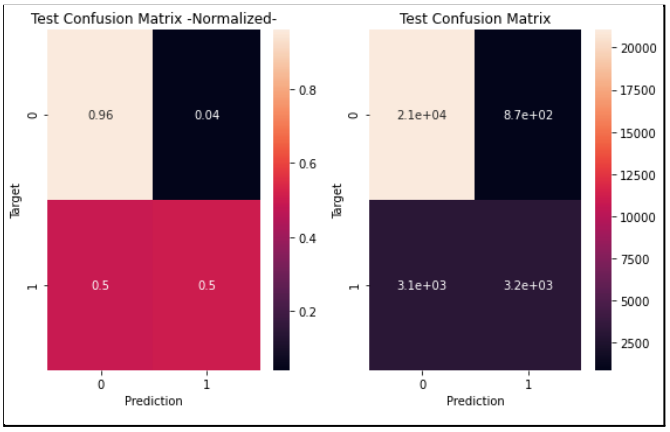
Modelo	Feature set 1			Feature set 2			Feature set 3		
	Acc	AUC	PR AUC	Acc	AUC	PR AUC	Acc	AUC	PR AUC
LR	84.53	0.86	0.68	84.64	0.86	0.69	84.56	0.86	0.68

LR + cw	78.6	0.86	0.68	-	-	-	-	-	-
NN	85.93	0.89	0.74	85.98	0.89	0.74	86.08	0.89	0.74

7. Evaluación de resultados

De la comparación de la tabla anterior surge como resultado que las redes neuronales funcionan mejor que la regresión logística clásica a la hora de clasificar. Por otro lado, no se ve una mejora o un degradamiento considerable usando uno u otro features set en cada uno de los modelos.

A continuación se detallan los resultados del modelo de red con el conjunto de features 1.




```
Test:
              precision    recall  f1-score   support

           0       0.87      0.96      0.91     21918
           1       0.78      0.50      0.61      6240

    accuracy                           0.86     28158
   macro avg       0.83      0.73      0.76     28158
weighted avg       0.85      0.86      0.85     28158
```

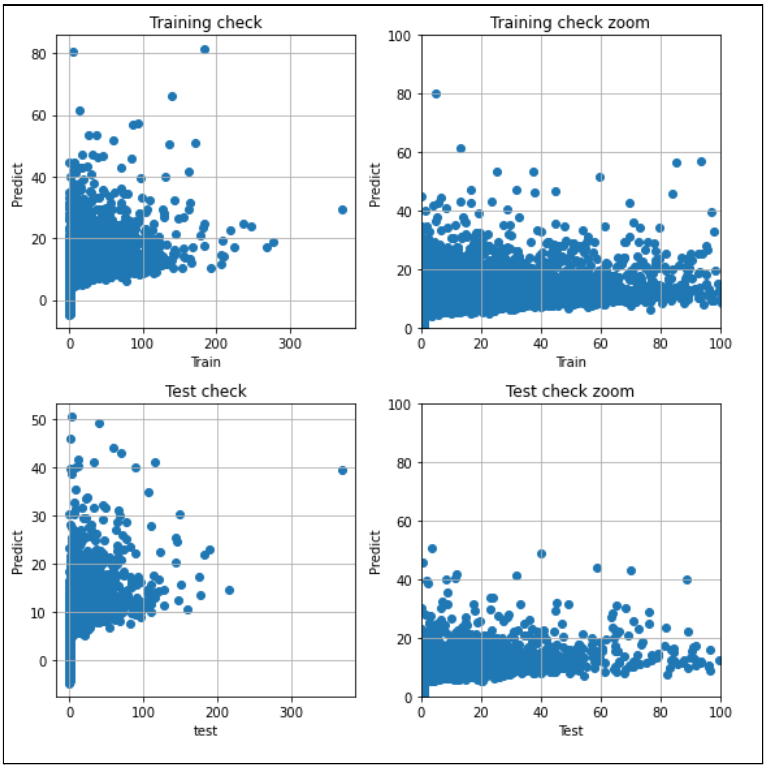
8. Predicción de RainfallTomorrow

Para la predicción de precipitaciones del día posterior se utilizaron prácticamente los mismos set de features. La única diferencia es que se agregó como feature adicional la columna de *RainTomorrow.*

Se entrenó sólamente un Regresor lineal (LR) y los resultados obtenidos fueron los siguientes:

Modelo	Feature set 1			Feature set 2			Feature set 3		
	MAE	MSE	R2 SQR	MAE	MSE	R2 SQR	MAE	MSE	R2 SQR
LR	2.62	46.81	0.32	2.65	50.25	0.31	2.64	50.25	0.31

En este caso se observa una ligera mejora del feature set 1 que es justamente el que tenía una menor cantidad de features correlacionadas.



9. Conclusiones

Se logró aplicar una buena parte de los conceptos aprendidos durante la cursada y se obtuvieron resultados bastante aceptables.

En cada una de las etapas del proyecto se podría haber hecho un análisis más exhaustivo, que no se realizaron por cuestiones de tiempo. Por ejemplo, en el caso de análisis de datos, se podría haber explorado la cantidad de muestras de días lluviosos por cada una de las ubicaciones, se podría haber analizado qué tan cerca de una costa está cada una de las ubicaciones, cuál es el promedio de precipitaciones en cada ubicación, el histórico de temperaturas a lo largo del tiempo (en general y dividido por ubicaciones), etc. En el caso de ingeniería de features se podrían haber explorado métodos de imputación estadísticos univariados, otros métodos de imputación multivariado (como KNN), se podrían haber codificado las variables categóricas con métodos como BinaryEncoding, se podrían haber sintetizado features con el acumulado de días sin llover para cada ubicación, etc.

En cuanto a modelos, se podría haber probado también con algún tipo de árbol para comparar resultados.

En cuanto a performance de modelos y selección de features se concluyó que, con el preprocesamiento realizado, prácticamente no había diferencias en los modelos de clasificación y había una leve mejora en los modelos de clasificación cuando se elegía el set de features con features poco correlacionadas.