Análisis de datos e Introducción Inteligencia Artificial CEIA

TP 1
Integrador

Alumno: Lic. Pedro Perez
Docentes: Ing. Magdalena Bouza, Ing. Lautaro Delgado.

Introducción:

En este trabajo practico se desarrolló un modelo de aprendizaje supervisado a partir de un conjunto de datos meteorológicos de diferentes localidades de Australia. dado que la característica del problema era predecir si lloverá o no al día siguiente, se utilizó un modelo de clasificación binaria para predecir la presencia de lluvia para un determinado día. Se exploraron diferentes modelos de clasificación basados en regresión logística y árboles de decisión. La métrica utilizada para determinar el mejor modelo fue el área bajo la curva ROC (Receiver Operating Characteristic). Los modelos con mejor performance fueron Random Forest con 0.77 y Regresión Logística con 0.76.

Set de datos:

Se trabajó con un dataset obtenido de la plataforma Kaggle con datos de distintas estaciones meteorológicas de Australia. El objetivo es predecir si lloverá o no al día siguiente (variable RainTomorrow), en función datos meteorológicos del día actual.

El dataset weatherAUS es un dataset que contiene más de 140 mil observaciones climatológicas de estaciones meteorológicas en el territorio australiano, el dataset contiene las siguientes variables:

- Date: Fecha de la observación.
- Location: Nombre de la estación meteorológica.
- MinTemp: Temperatura mínima en grados Celsius.
- MaxTemp: Temperatura máxima en grados Celsius.
- RainFall: Cantidad de lluvia en mm.
- Evaporation: Evaporación de agua durante el día en mm.
- Sunshine: Número de horas de luz solar durante el día.
- WindGustDir: Dirección de la ráfaga de viento más fuerte durante el día.
- WindGustSpeed: Velocidad en (km/h) de la ráfaga de viento más fuerte durante el día.
- WindDir9am: Dirección del viento a las 9 am.
- WindDir3pm: Dirección del viento a las 3 pm.
- WindSpeed9am: Velocidad en (km/h) del viento a las 9 am.
- WindSpeed3pm: Velocidad en (km/h) del viento a las 3 pm.
- Humidity9am: Humedad relativa en porcentaje a las 9 am.
- Humidity3pm: Humedad relativa en porcentaje a las 3 pm.
- Pressure9am: Presión atmosférica (mmHg) a las 9 am.
- Pressure3pm: Presión atmosférica (mmHg) a las 3 pm.
- Cloud9am: Nivel de nubosidad (escala Octa) a las 9 am.
- Cloud3pm: Nivel de nubosidad (escala Octa) a las 3 pm.
- Temp9am: Temperatura medida a las 9 am en grados Celsius.
- Temp3pm: Temperatura medida a las 3 pm en grados Celsius.
- RainToday: 1 si el día de la medición llovía y 0 en otro caso.
- RainTomorrow: 1 si el día posterior a la medición llovía y 0 en otro caso.

1. Análisis exploratorio inicial:

- El dataset cuenta con 145460 registros y 22 variables de entrada.
- De las 22 variables de entrada existen 14 variables numéricas y 8 categóricas.
- La variable RainTomorrow es la variable de salida.
- Todas las columnas salvo Date y Location tienen valores faltantes.
- La variable de salida RainTomorrow tiene 3267 valores faltantes.

Variables numéricas de entrada:

- MinTemp
- MaxTemp
- Temp9am
- Temp3pm
- Pressure9am
- Pressure3pm
- Rainfall
- Evaporation
- Sunshine
- WindGustSpeed
- WindSpeed9am
- WindSpeed3pm
- Humidity9am
- Humidity3pm

Distribución de las variables numéricas de entrada:

Se realizaron gráficos de histogramas y QQ plot para visualizar las distribuciones de las variables de numéricas de entrada y que tanto se acercan a distribuciones normales:

- Las variables MinTemp, MaxTemp, Temp9am, Temp3pm, Pressure9am y Pressure3pm poseen una asimetría baja.
- El gráfico Q-Q plot muestra que las variables, MinTemp, MaxTemp, Temp9am, Temp3pm, Pressure9am y Pressure3pm podrían asumirse con distribución normal, sin embargo, se observa que las colas se separan bastante. Es probable que existan outliers en ambos lados de las colas.
- Las variable Rainfall, Evaporation, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Sunshine, Humidity9am, Humidity3pm no muestran una distribución normal.
- Se observa que la variable Rainfall es fuertemente centrada en cero, esto se debe a que el dataset está desbalanceado hacia días sin lluvia, RainToday="No".

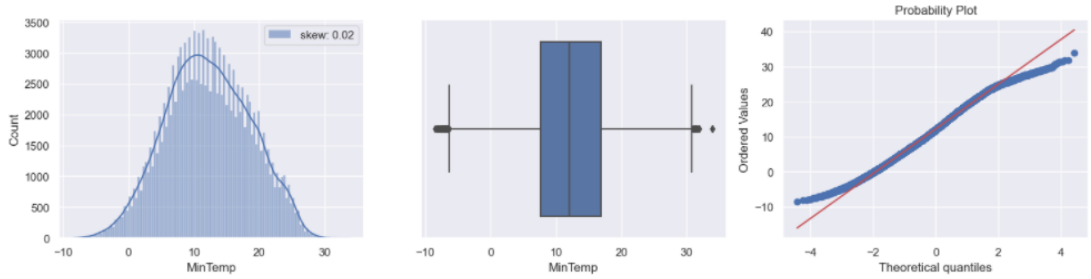

Figura 1 Histograma, Box plot y QQ plot para las variable MinTemp.

Se aplicaron diferentes métodos de transformación de variables para que las mismas siguieran una distribución normal. Se observó el efecto de las transformaciones, por un lado aquella que utiliza información de cuantil y por otro lado el método "yeo-johnson":

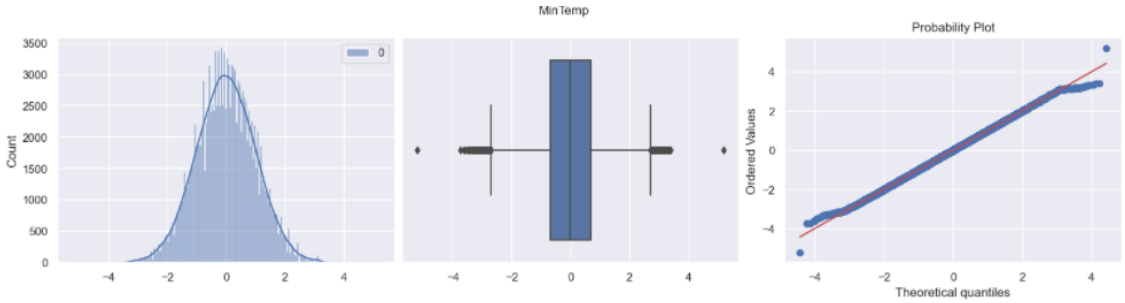

Figura 2 variables transformadas información del cuantil.

Análisis de outliers:

Se probaron tres técnicas de detección de outliers: z-score, outliers rango intercuartil, forest. Se observó presencia de outliers en casi todas las variables, y que el análisis realizado por el rango intercuartil tiende a penalizar más a los outliers.

	feature	outliers_livianos	outliers_pesados	outliers_zscore	outliers_forest
0	MinTemp	54.0	0.0	19.0	41.0
1	MaxTemp	489.0	0.0	331.0	130.0
2	Temp9am	262.0	0.0	149.0	101.0
3	Temp3pm	764.0	0.0	397.0	94.0
4	Pressure9am	1191.0	8.0	503.0	94.0
5	Pressure3pm	919.0	7.0	426.0	180.0
6	Rainfall	25578.0	20762.0	2456.0	329.0
7	Evaporation	1995.0	481.0	870.0	500.0
8	Sunshine	0.0	0.0	0.0	22.0
9	WindGustSpeed	3092.0	153.0	1368.0	34.0
10	WindSpeed9am	1817.0	114.0	1362.0	16.0
11	WindSpeed3pm	2523.0	82.0	958.0	32.0
12	Humidity9am	1425.0	0.0	472.0	1.0
13	Humidity3pm	0.0	0.0	0.0	4.0

Tabla 1 Detección de outliers variables numéricas por diferentes métodos.

Variables categóricas de entrada:

- Date
- Location
- WindGustDir
- WindDir9am
- WindDir3pm
- Cloud9am
- Cloud3pm
- RainToday

Observaciones:

- Las variables Cloud9am y Cloud3pm son ordinales.
- La variable Location posee 49 categorías e identifican nombres puntos de locación en la cual se realizaron las mediciones.
- Las variables WindGustDir, WindDir3pm y WindDir9am poseen 17 categorías las cuales son iguales las cuales son códigos que describen puntos cardinales de dirección del viento.
- Las variables Cloud9am y Cloud3pm poseen 10 categorías.
- La variable RainToday y RainTomorrow poseen dos categorías (Yes, No).
- La variable RainTomorrow es la variable de salida u objetivo.

Correlación entre variables:

Se analizaron las correlaciones entre las variables. Se utilizando gráficos observar correlaciones y se utilizó el método Spearman ya que evalúa la relación monoatómica entre dos variables. Se obtuvieron los siguientes resultados:

- Existe una fuerte correlación entre las variables MinTemp, MaxTemp, Temp9am, Temp3pm.
- Existe una fuerte correlación entre las variables WindGustSpeed, WindSpeed3pm y WindSpeed9am.
- Existe una fuerte correlación entre las variables Sunshine y las variables Humidity3pm y Humidity9am.
- Existe una fuerte correlación entre las variables Sunshine y las variables Cloud9am y Cloud3pm.
- Existe una fuerte correlación entre las variables Cloud9am y Cloud3pm.
- Existe una fuerte correlación entre las variables Pressure9am y Pressure3pm.
- Existe una fuerte correlación entre las variables Rainfall, Sunshine, Humidity, Pressure, Cloud y RainToday con la variable de Salida.

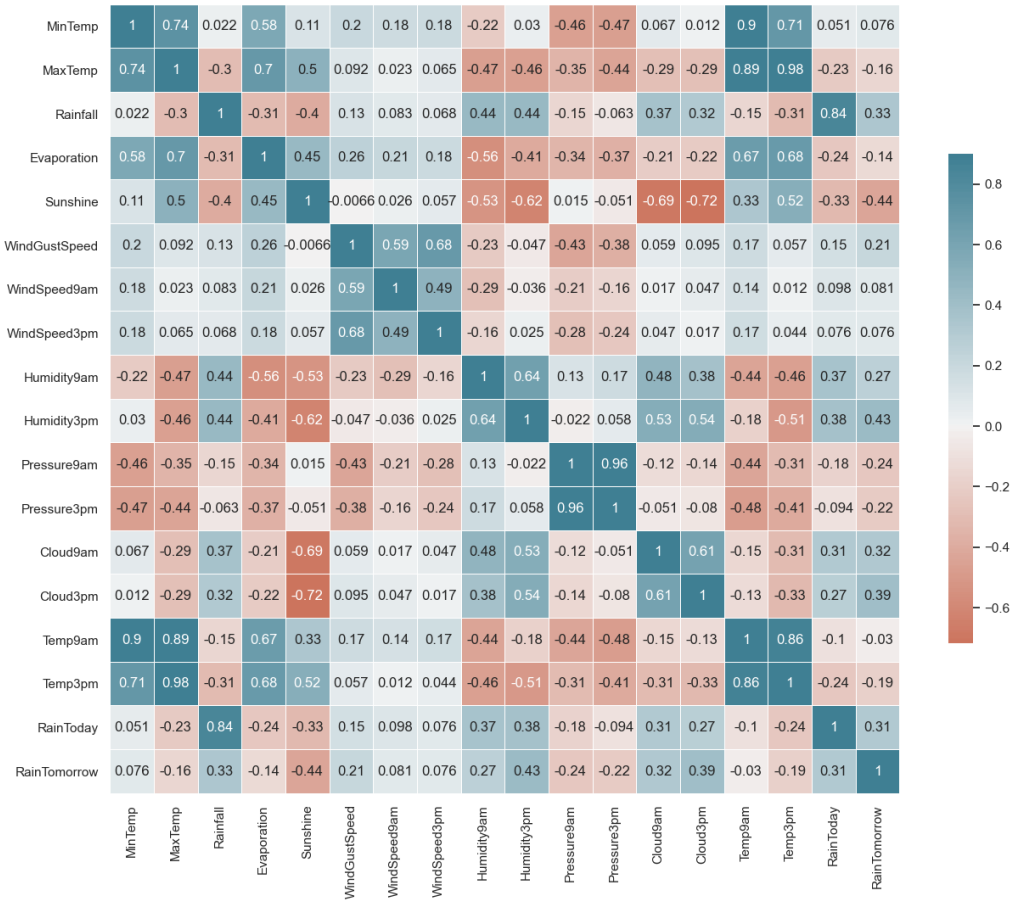

Figura 3 Matriz de correlación entre variables.

Utilizamos el test de Kolmogorov-Smirnov lograr determinar el grado de "separación" de las distribuciones de las variables dado RainTomorrow="Yes" y la distribución de las variables dado RainTomorrow="No".

	feature	statistic_z	pvalue
3	Temp9am	0.041036	1.268967e-36
1	MinTemp	0.069684	7.264208e-105
8	Evaporation	0.073726	2.389429e-117
11	WindSpeed9am	0.078010	2.393171e-131
12	WindSpeed3pm	0.083712	3.225377e-151
2	MaxTemp	0.162305	0.000000e+00
4	Temp3pm	0.188173	0.000000e+00
6	Pressure3pm	0.194916	0.000000e+00
5	Pressure9am	0.216400	0.000000e+00
10	WindGustSpeed	0.220457	0.000000e+00
9	Sunshine	0.250714	0.000000e+00
13	Humidity9am	0.283475	0.000000e+00
7	Rainfall	0.348507	0.000000e+00
0	dif_temp	0.367228	0.000000e+00
14	Humidity3pm	0.443290	0.000000e+00

Tabla 2 Kolmogorov-Smirnov correlación con la variable de salida.

Análisis de valores faltantes:

Se observa un gran número de faltantes en las columnas Evaporation, Sunshine, Cloud9am, Cloud3pm.

Se observa que los faltantes coinciden en esas columnas, es probable que sean Missing not at Random ya que estos datos no hayan estado disponibles para determinadas localidades.

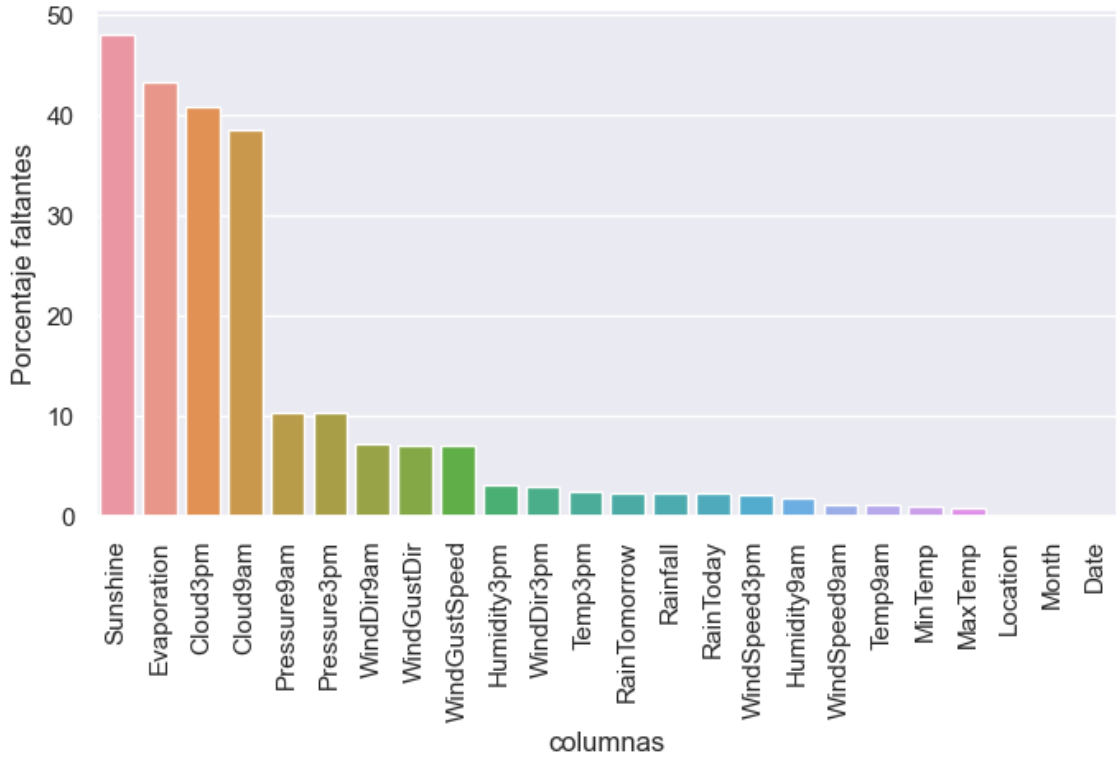

Figura 4 Porcentaje de faltantes por variables.

- Las columnas “Sunshine, “Evaporation”, "Cloud3pm" y “Cloud9am” son las variables que poseen mayor cantidad de faltantes.
- El resto de las columnas posee un número de faltantes por debajo al 10%.
- Las columnas "Dates" y "Location" no poseen valores faltantes.

Variable de salida:

La variable de salida RainTomorrow que identifica la presencia de lluvia o no al día siguiente se observó que se encontraba desbalanceada:

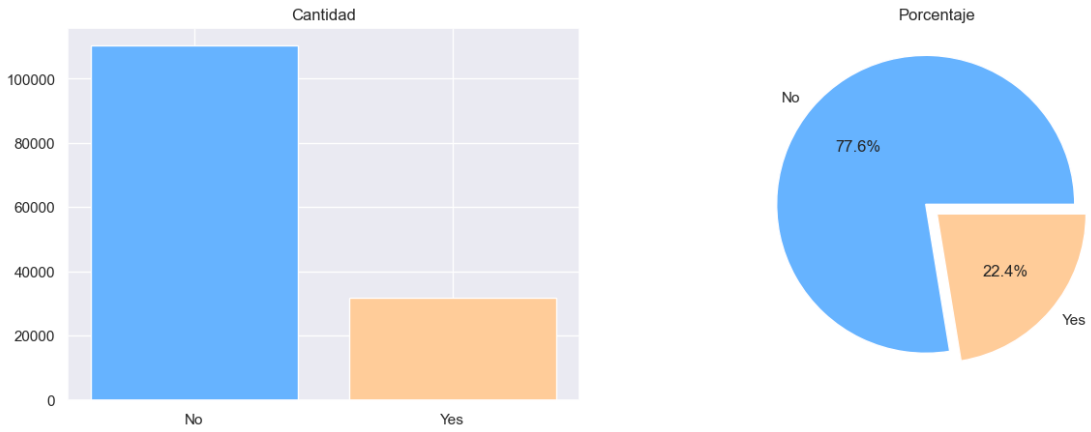

Figura 4 Relación de clases variable de salida.

2. Limpieza y preparación de datos:

Imputación de faltantes:

Se realizó una imputación manual de faltantes por medición de las siguientes variables:

- Cloud9am y Cloud3pm.
- Si RainToday='Yes' se imputo el faltante de Cloud3pm con el valor medio (6).
- Si RainToday=No se imputo el faltante de Cloud3pm con el valor medio (4).
- Se utilizaron las variables WindDir9am y WindDir3pm para imputar los faltantes de la variable WindGustDir.
- Se utilizaron las variables Temp9am y Temp3pm para imputar los faltantes de las variables MinTemp y MaxTemp respectivamente.
- Se utilizaron las variables WindSpeed3pm y WindSpeed3pm para imputar los faltantes de la variable WindGustSpeed.
- Se utilizaron las variables WindDir3pm y WindDir9am para imputar los faltantes de la variable WindGustDir.

Luego se realizó una imputación estadística. El método que arrojo mejores resultados para la imputación de valores faltantes fue el método MICE.

Ingeniería de features:

- Se creó una nueva variable DifTem para reflejar la diferencia de temperatura máxima y mínima. Se eliminaron las variables Temp9am y Temp3pm por tener alta correlación con las variables MinTemp y MaxTemp respectivamente.
- Se eliminó la variable Evaporation por tener muchos valores faltantes (43%) y por estar altamente correlacionada con las variables relacionadas a temperatura y humedad.
- Se eliminó la variable Sunshine por tener muchos valores faltantes (48%) y por estar altamente correlacionada con las variables de nubosidad.

- Se eliminaron las variables WindDir3pm y WindDir9am por tener alta correlación entre ellas y la variable WindGustDir. Se mantuvo la variable WindGustDir por tener mayor correlación con la variable de salida, la misma fue convertida a una escala cíclica.
- Se eliminaron las variables WindSpeed3pm y WindSpeed9am por tener poca correlación con la variable de salida y mucha correlación entre ellas.
- La variable Location fue convertida a coordenadas de latitud y longitud buscando una relación espacial.
- Las variables Pressure9am y Pressure3pm se promediaron y se unificaron en una sola variable por tener alta correlación entre ellas.

Sampling:

Se aplicó la técnica SMOTE para entrenar el modelo y tratar de evitar el desbalanceo de las clases, Se observó que hubo una mejora en las métricas del set de datos de prueba.

PCA:

Se aplicó PCA para evaluar si disminuyendo la dimensionalidad del modelo ayudaba a mejorar la performance de los modelos. Se encontró que el uso de PCA no ayudo en mejorar la performance del modelo debido a que el mismo fue entrenado con un bajo número de variable.

Conclusiones evaluación de modelos:

Los modelos con los cuales se obtuvieron mejores resultados fueron Random Forest con un score f1 de 0.64 de la clase positiva y regresión logística con un score f1 de la clase positiva.

Se observó que todos los modelos tendieron a clasificar muy bajo la clase positiva y en general bastante bien la clase negativa esto producto del problema del desbalanceo de las clases.