Análisis de Datos
Introducción a Inteligencia Artificial

Trabajo Final Integrador

Pablo Brillanti

Índice

Código

Todo el código en este trabajo práctico se implementó en “Jupyter notebook” y se deja el Link a continuación.

https://github.com/Pbrillan/CEIA/tree/main/I%20A/TP%20Integrador

Análisis exploratorio inicial

Primero se analizaron cantidad de columnas y algunos datos estadísticos del dataset

	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindDir9am	...	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Cloud9am	Cloud3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	...	71.0	22.0	1007.7	1007.1	8.0	NaN	16.9	21.8	No	No
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	NNW	...	44.0	25.0	1010.6	1007.8	NaN	NaN	17.2	24.3	No	No
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	W	...	38.0	30.0	1007.6	1008.7	NaN	2.0	21.0	23.2	No	No
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	SE	...	45.0	16.0	1017.6	1012.8	NaN	NaN	18.1	26.5	No	No
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	ENE	...	82.0	33.0	1010.8	1006.0	7.0	8.0	17.8	29.7	No	No

5 rows × 23 columns

```
<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
```

	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

Se analizaron los tipos de datos de cada una de las columnas

```
Combinada:
Date              object

Categoricas:
Location          object
WindGustDir       object
WindDir9am        object
WindDir3pm        object
RainToday         object
RainTomorrow      object

Numericas:
MinTemp          float64
MaxTemp          float64
Rainfall         float64
Evaporation      float64
Sunshine         float64
WindGustSpeed    float64
WindSpeed9am     float64
WindSpeed3pm     float64
Humidity9am      float64
Humidity3pm      float64
Pressure9am      float64
Pressure3pm      float64
Cloud9am         float64
Cloud3pm         float64
Temp9am          float64
Temp3pm          float64
```

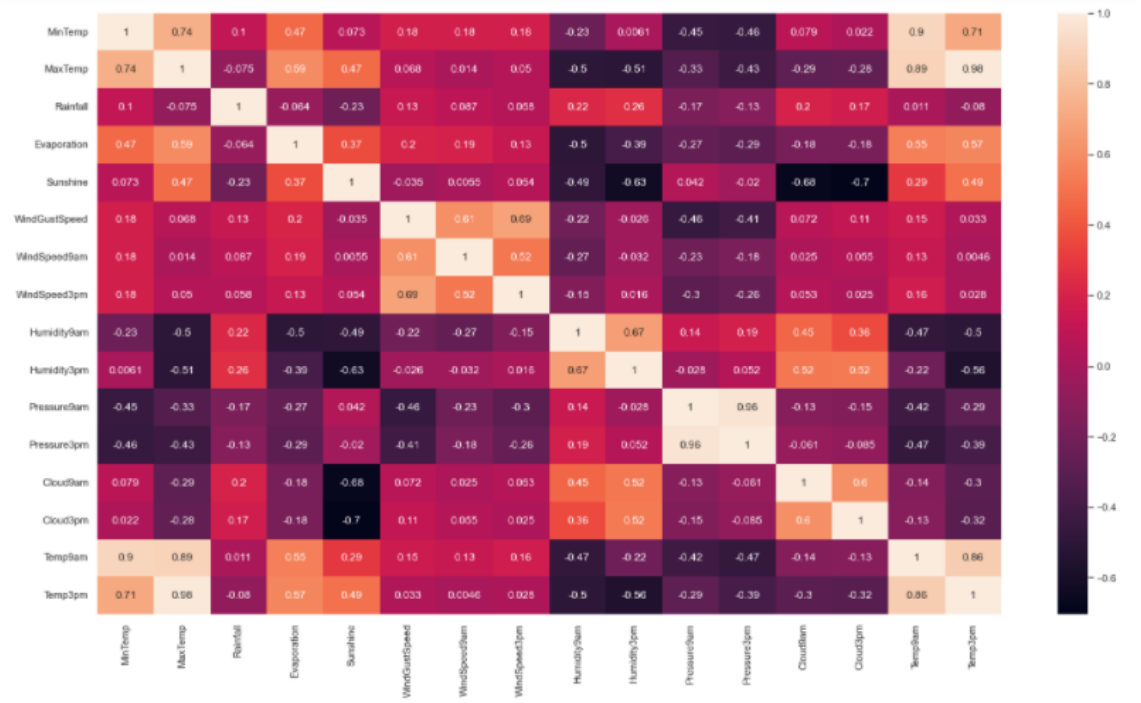
Posteriormente se analizaron exhaustivamente las variables de entrada.

Variables numéricas:

Se realizaron histogramas para estudiar la distribución de cada una de ellas, acá detectamos que en las variables de nubosidad se observa una distribución bimodal.

También se graficaron boxplot para cada una de las variables de entrada.

Graficamos la matriz de correlación donde vemos que las temperaturas están fuertemente correlacionadas entre sí, se aprecia además que existe correlación entre las velocidades de viento, las presiones, las humedades. Una relación inversa entre la nubosidad y la luz de sol.



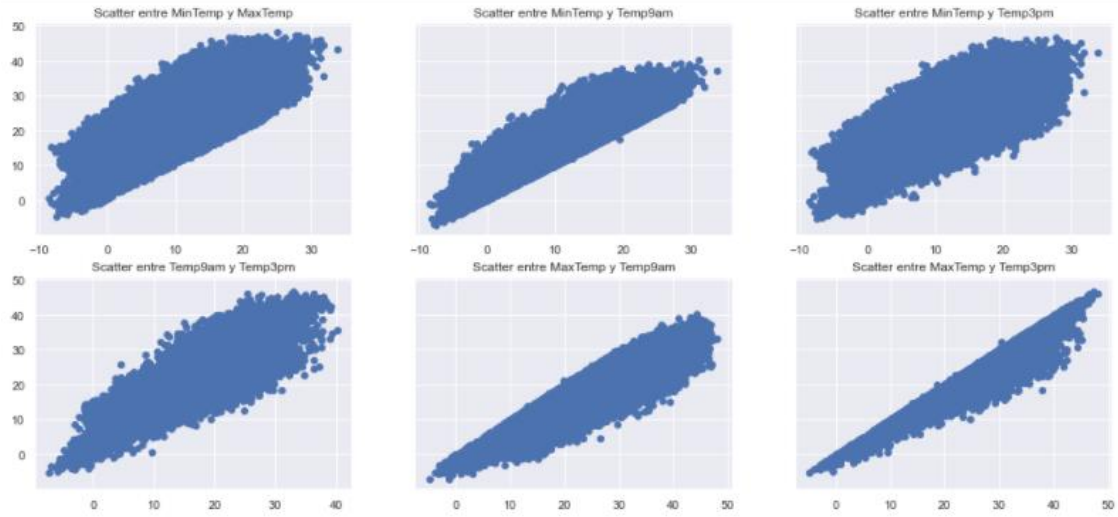
El tratamiento de las variables correlacionadas se realizó después de la eliminación por valores faltantes.

Vemos que tienen alta correlación las 4 temperaturas entre si (MinTemp, MaxTemp, Temp9am, Temp3pm)

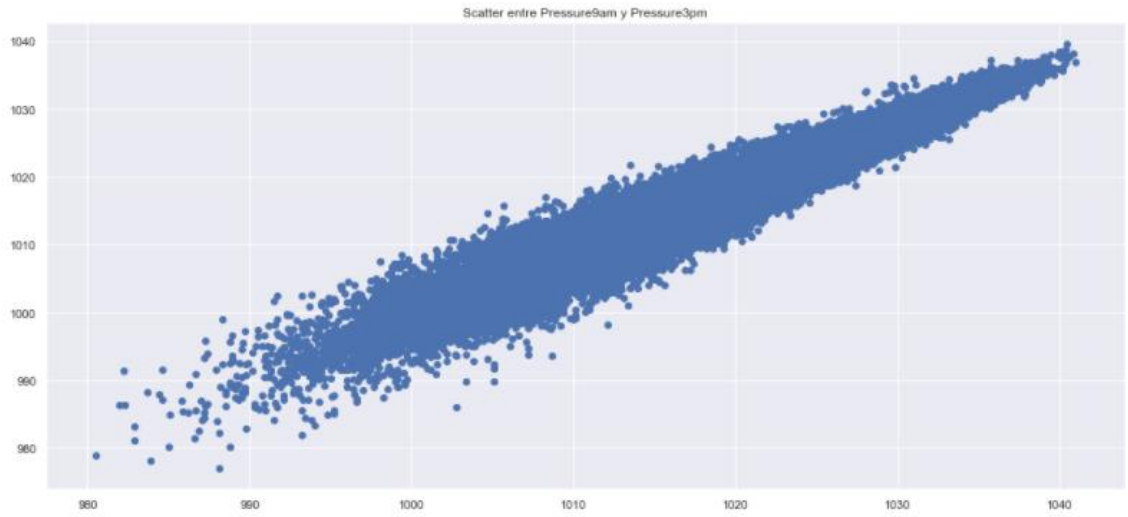
La velocidad del viento (WindGustSpeed, WindSpeed9am,

WindSpeed3pm)

Y las dos presiones (Pressure9am, Pressure3pm)


Scatter entre MinTemp y MaxTemp
Scatter entre MinTemp y Temp9am
Scatter entre MinTemp y Temp3pm
Scatter entre Temp9am y Temp3pm
Scatter entre MaxTemp y Temp9am
Scatter entre MaxTemp y Temp3pm


Scatter entre WindGustSpeed y WindSpeed3pm


Scatter entre Pressure9am y Pressure3pm

Variables Compuestas:

La Variable de entrada Date tiene tipo de dato de Fecha. Para que pueda ser interpretada se codificó en 3 variables numéricas (Día, Mes y Año). Se observó que no hay información en las variables día y año. La variable mes puede tener alguna inferencia en la probabilidad de que llueva y el volumen de lluvia. Además, es esperable un comportamiento cíclico con periodo anual según la estación del año.

Por este motivo solo se conserva la variable Mes codificada de forma cíclica.

Variables Categóricas:

```
Location: 49
WindGustDir: 17
WindDir9am: 17
WindDir3pm: 17
RainToday: 3
RainTomorrow: 3
```

La variable Location se trata de una variable con muy alta cardinalidad. Para codificarla acudimos a las coordenadas geográficas de cada localidad aprovechando la información oculta en la relación espacial entre las mismas. Esto se implementó mediante la carga de un archivo que vincula cada localidad con sus coordenadas.

De esta manera se codifico esta variable categórica como dos variables numéricas.

Las variables de dirección de viento se pueden agrupar en la codificación ya que las tres tienen las mismas clases. Estas son variables con 16 categorías más los datos faltantes.

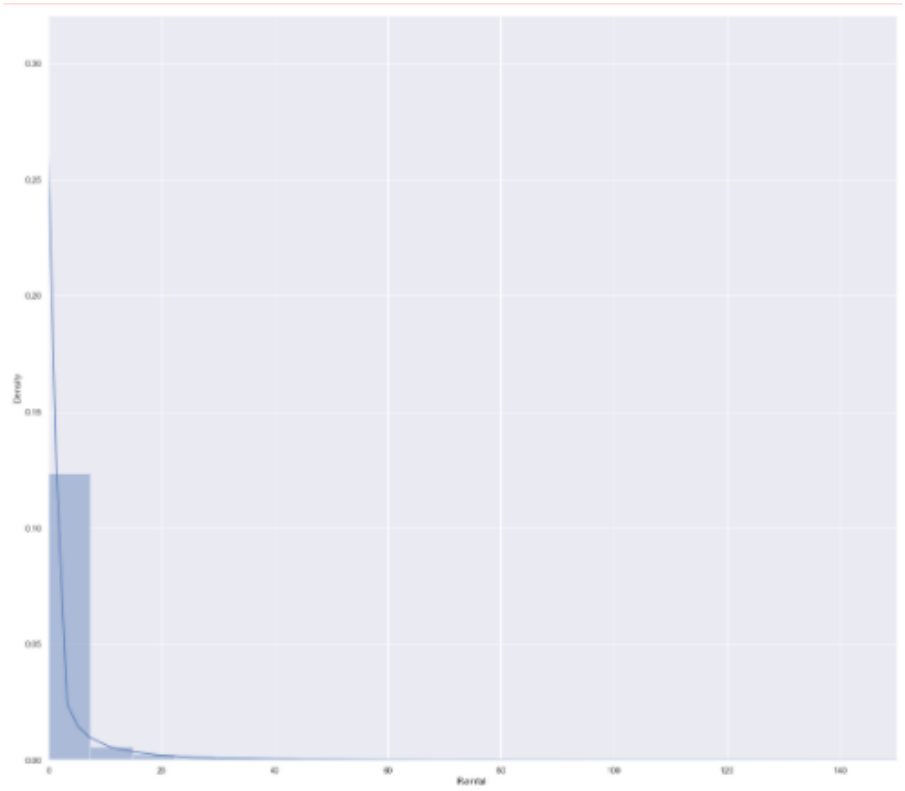
Para codificarlas acudimos a poner una referencia angular y transformar el punto cardinal en un ángulo de dirección de viento. De esta manera se codifico cada una de estas variables categóricas en dos variables numéricas.

Las variables RainToday y RainTomorrow son variables binarias por este motivo se las codifico a cada una de ellas con una variable numérica binaria con el método de Label Encoder.

Variables de Salida:

Para el primer problema tenemos como variable de salida RainTomorrow. Esta tiene el 2,24% de valores faltantes y el dataset está fuertemente balanceado para el lado de los días que no llueve.

Para el segundo problema tenemos como variable de salida Rainfall. Esta tiene el 2,24% de valores faltantes.



Esquema de Validación de los resultados

Dividimos el Dataset dos partes (train 70% y test 30%)

Se eliminan las muestras con salida NaN para cada uno de los problemas

Limpieza y preparación de datos / ingeniería de features

En primer lugar, se codificaron las variables categóricas y se eliminaron columnas muy correlacionadas y con más de 38% de valores faltantes según se indicó en secciones anteriores.

Para la predicción de RainTomorrow se probaron 3 métodos de imputación.

1- Por media/mediana
2- Por Vecinos cercanos
3- Por MICE

Y para cada uno de estos métodos probamos los modelos de Regresión logística y de Random Forest.

Modelos y Análisis de resultados

Los Resultados los evaluamos con F1 Score para independizarnos del cesgo del dataset desbalanceado.

Random Forest con imputación por media/mediana f1_Score: 84.63 %

Logistic Regresion con imputación por media/mediana f1_Score: 68.13 %

Random Forest con imputación por Vecinos cercanos f1_Score: 84.63 %

Logistic Regresion con imputación por Vecinos cercanos f1_Score: 82.52 %

Random Forest con imputación por MICE f1_Score: 84.52 %

Logistic Regresion con imputación por MICE f1_Score: 82.49 %

Posteriormente al dataset imputado por MICE aplicamos PCA para explicar el 90% de la varianza. Se redujo el tamaño del dataset a 10 columnas y se entrenó Logistic Regresion obteniendo f1_Score: 78.95 %

Para estimar la columna Rainfall para el día siguiente se implementó una búsqueda para armar el dataset y luego se entrenó un modelo de regresión lineal

Obteniendo un Acurancy de 30%

Conclusiones

Se visualizaron los datos para tener una vista general de todo el dataset y se probaron métodos de codificación para variables categóricas.

Con el fin de reducir la cardinalidad la variable Location se codifico con sus coordenadas geográficas, para un mejor análisis se podrían haber agrupado estas coordenadas en regiones y de esta forma sacar más provecho a la información espacial aportada.

Destacamos que en el desarrollo del TP se eliminaron las columnas y se realizaron las codificaciones antes de realizar el Split del Dataset, no es lo adecuado, pero no se modificó por razones de tiempo y que no va a afectar a los resultados ya que solo se operó a nivel columnas.

Se implementaron 3 métodos de imputación de variable notando que el método univariado es mucho más fácil de implementar y los métodos multivariados obtuvieron bastante mejor resultado a la hora de entrenar los modelos.

Como posible mejora seria analizar más exhaustivamente el Missing at random de las variables a imputar.

A su vez también hay que destacar que el método univariado no se aplicó según las recomendaciones ya que se aplicó a modo de prueba a variables con más del 5% de faltantes.

Se usó F1 Score para tener una métrica más realista, ya que los datos en la columna target estaban muy desbalanceados. Sería interesante probar técnicas de balanceo del dataset para analizar los resultados

Para la implementación de los modelos hubiese sido mejor empezar la programación de forma organizada, legible y estructurada para al final disponer de un código que se pueda reutilizar y sea fácil de leer. Además, hubiese sido bueno iterar con los hiperparámetros para un ajuste mas fino.

Para la predicción de Rainfall al día siguiente se obtuvo un acuracy bastante bajo probablemente puede mejorarse validando con RainTomorrow es decir que para cuando predijimos que mañana llueve validar el valor de RainfallT y para el caso contrario RainfallT sea cero.