

Análisis de Datos

Introducción a Inteligencia Artificial

Trabajo Final Integrador

Pablo Brillanti

Índice

[Código 3](#_Toc85211570)

[Análisis exploratorio inicial 3](#_Toc85211571)

[Variables numéricas: 5](#_Toc85211572)

[Variables Compuestas: 7](#_Toc85211573)

[Variables Categóricas: 7](#_Toc85211574)

[Variables de Salida: 8](#_Toc85211575)

[Esquema de Validación de los resultados 8](#_Toc85211576)

[Limpieza y preparación de datos / ingeniería de features 9](#_Toc85211577)

[Modelos y Análisis de resultados 9](#_Toc85211578)

[Conclusiones 10](#_Toc85211579)

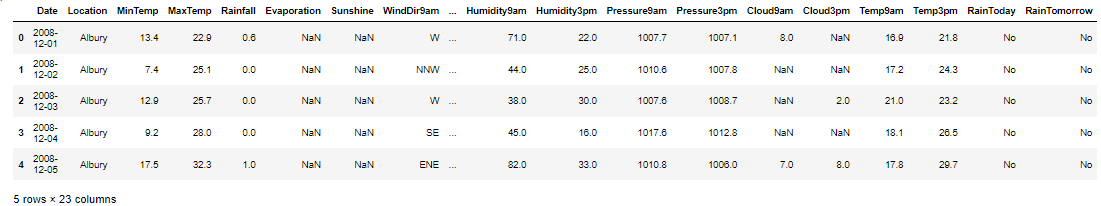
# Código

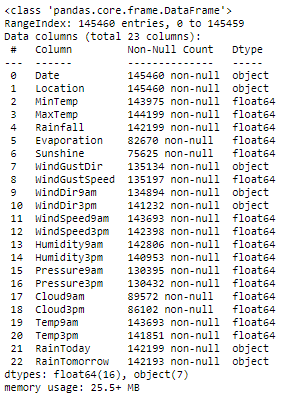
Todo el código en este trabajo práctico se implementó en “Jupyter notebook” y se deja el Link a continuación.

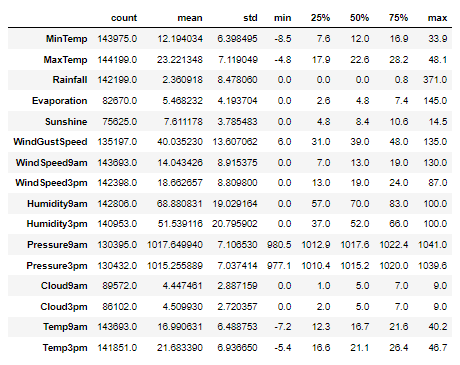
<https://github.com/Pbrillan/CEIA/tree/main/I%20A/TP%20Integrador>

# Análisis exploratorio inicial

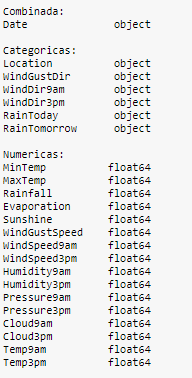
Primero se analizaron cantidad de columnas y algunos datos estadísticos del dataset







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



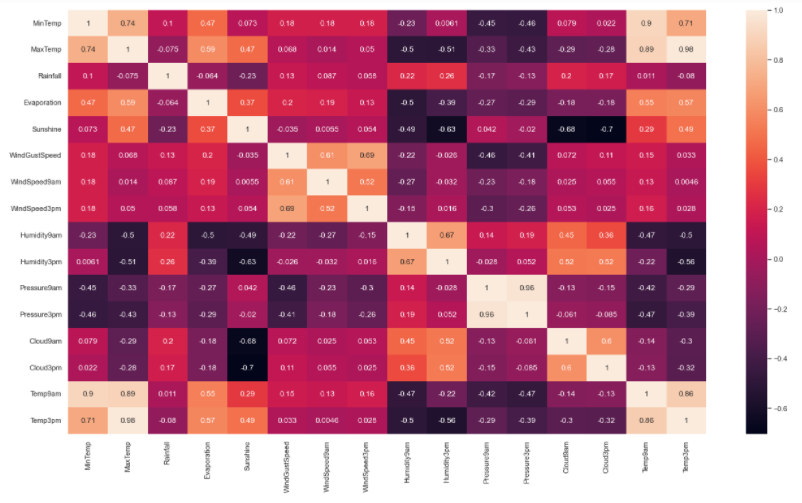
Posteriormente se analizaron exhaustivamente las variables de entrada.

## 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.



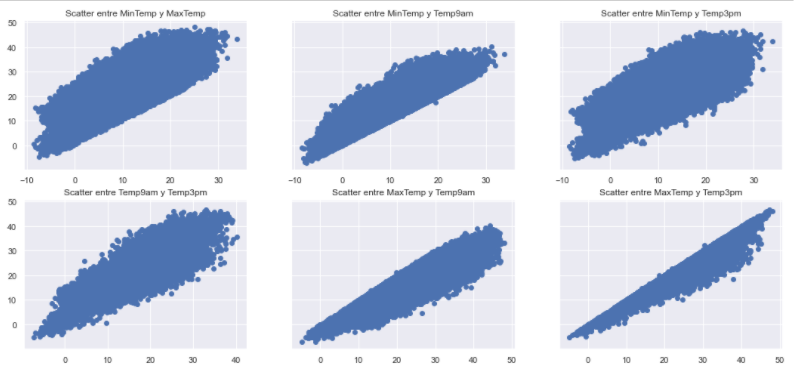
El tratamiento de las variables correlacionadas se realizó después de la eliminación por valores faltantes.

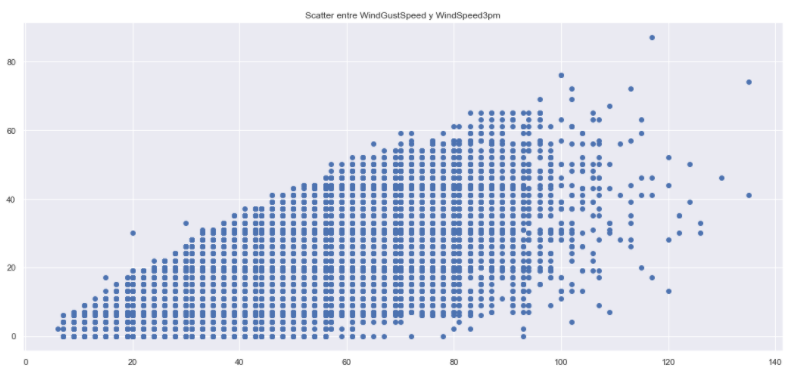
Vemos que tienen alta correlación las 4 temperaturas entre si (MinTemp, MaxTemp, Temp9am, Temp3pm)

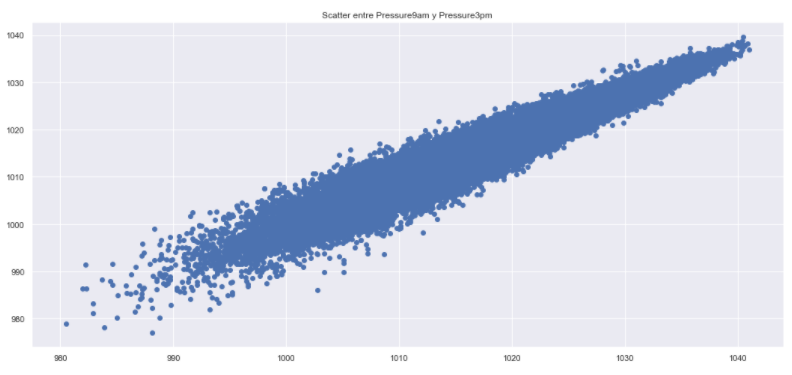
La velocidad del viento (WindGustSpeed, WindSpeed9am,

WindSpeed3pm)

Y las dos presiones (Pressure9am, Pressure3pm)







## 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.

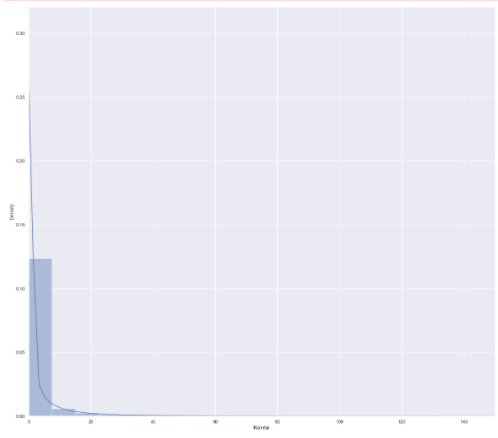
Para codificarlas acudimos a poner una referencia angular y transformar el punto cardinal en un ángulo de dirección de viento. De esta manera se codifico cada una de estas variables categóricas en dos variables numéricas.

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.