

Análisis de Datos

Introducción a Inteligencia Artificial

Trabajo Final Integrador

Pablo Brillanti

Índice

[Código 3](#_Toc84885319)

[Análisis exploratorio inicial 3](#_Toc84885320)

[Variables numéricas 5](#_Toc84885321)

[Variables Compuestas: 7](#_Toc84885322)

[Variables Categóricas: 7](#_Toc84885323)

[Variables de Salida 8](#_Toc84885324)

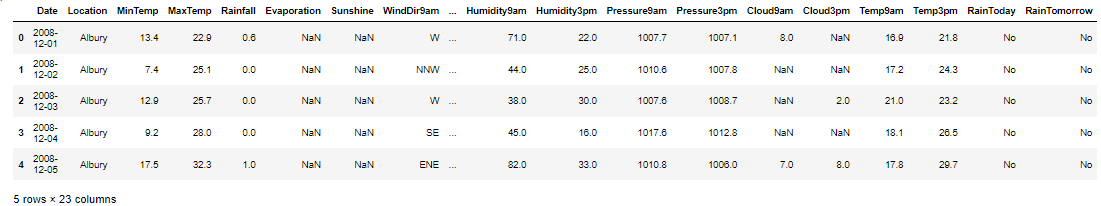
[Esquema de Validación de los resultados 8](#_Toc84885325)

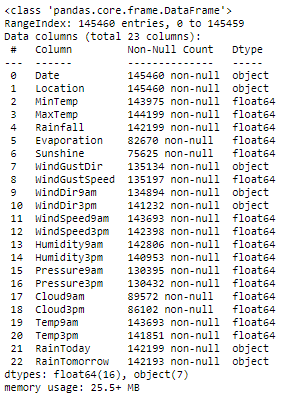
# Código

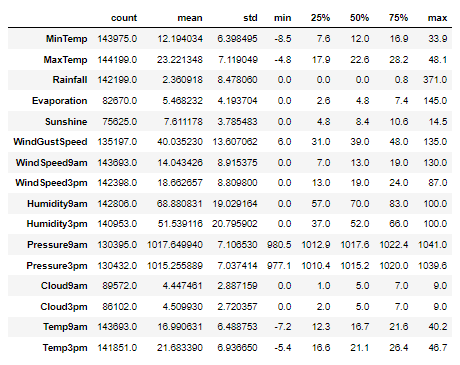
Todo el código en este trabajo práctico se implementó en “Jupyter notebook” y se deja el Link a continuación.

# Análisis exploratorio inicial

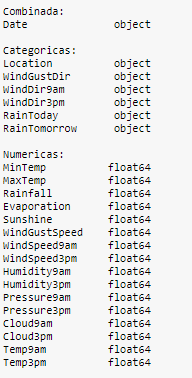
Primero se analizaron cantidad de columnas y algunos datos estadísticos del dataset







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



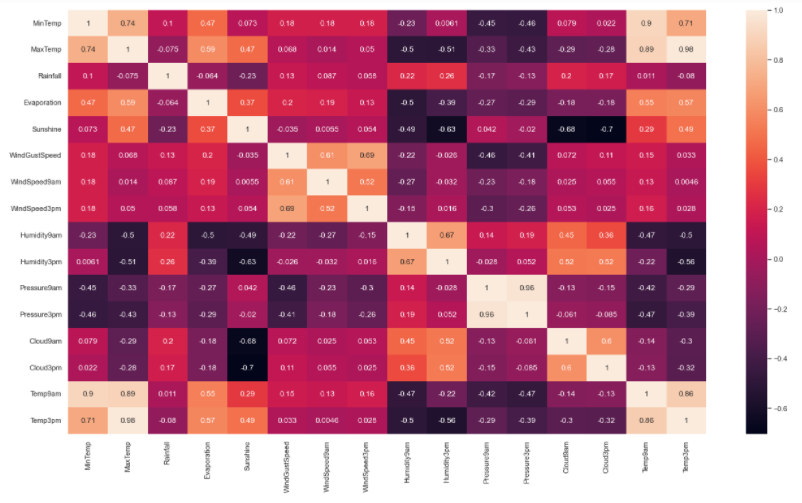
Posteriormente se analizaron exhaustivamente las variables de entrada.

## 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.



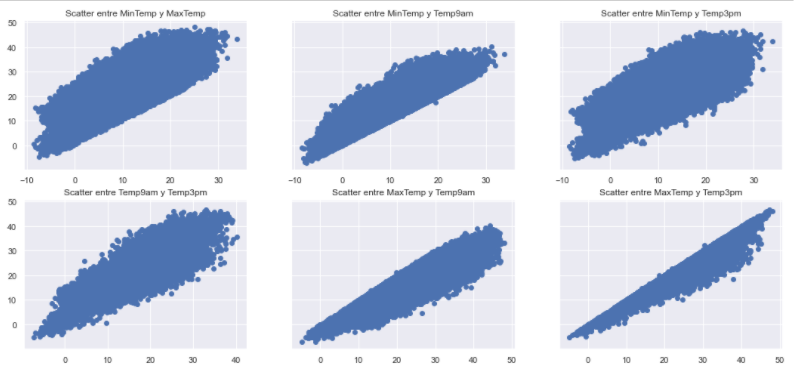
El tratamiento de las variables correlacionadas se realizó después de la eliminación por valores faltantes.

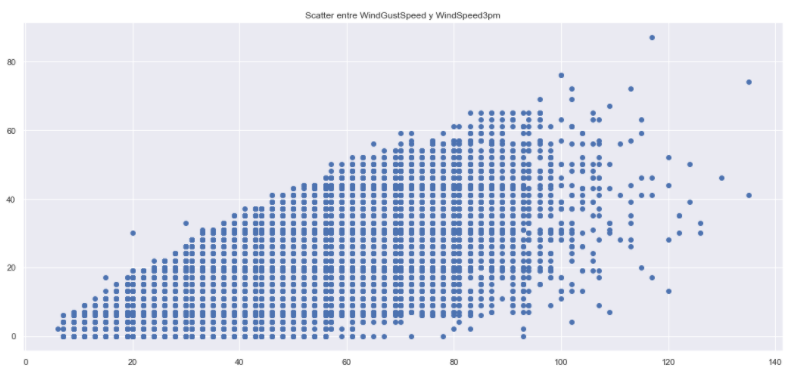
Vemos que tienen alta correlación las 4 temperaturas entre si (MinTemp, MaxTemp, Temp9am, Temp3pm)

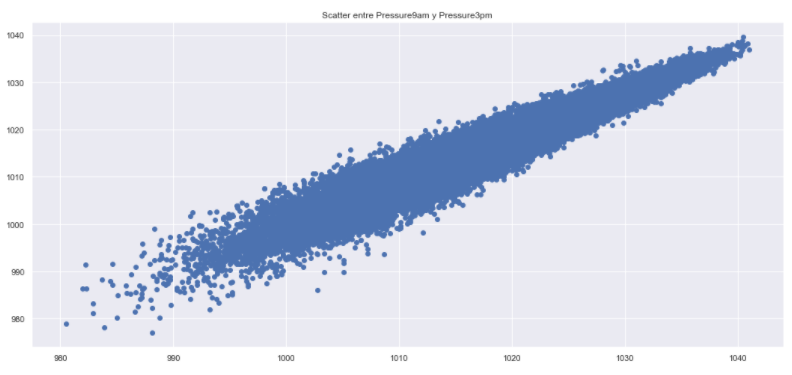
La velocidad del viento (WindGustSpeed, WindSpeed9am,

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.

## 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.

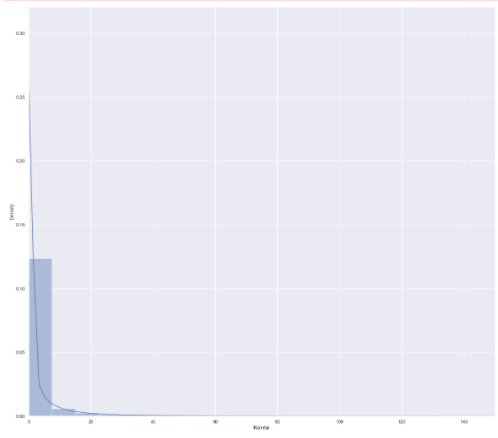
Para codificarlas acudimos a poner una referencia angular y transformar el punto cardinal en un ángulo de dirección de viento. De esta manera se codifico cada una de estas variables categóricas en una variable numérica.

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 30% y test 70%) y hacemos una copia para cada problema.

Se eliminan las muestras con salida NaN

# 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.

Imputación de valores faltantes

Entrenamiento de modelos

Análisis de resultados

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