****

**TP Integrador**

**2021**

Análisis de Datos

Autor: Maximiliano Torti

Contenido

[Ejercicio 1 3](#_Toc69756038)

[a. 3](#_Toc69756039)

[b. 4](#_Toc69756040)

[c. 5](#_Toc69756041)

[Ejercicio 2 6](#_Toc69756042)

[a. 6](#_Toc69756043)

[b. 7](#_Toc69756044)

[c. 7](#_Toc69756045)

[d. 7](#_Toc69756046)

[Ejercicio 3 9](#_Toc69756047)

[a. 9](#_Toc69756048)

[b. 9](#_Toc69756049)

[c. 10](#_Toc69756050)

# Apartado 1

Se comenzó cargando el dataset con pandas y luego se utilizaron los métodos “head” y “describe2 y las propiedades “shape” y “dtypes” para realizar un primer resumen.

De lo anterior se identificó que tenemos un dataset con 23 columnas (features) y alrededor de 145 mil registros. Se observó también a priori que los features contienen diferentes tipos de datos y la presencia de valores faltantes.

PONER TODAS LAS COLUMNAS CON EXPLCIACION QUE SON?

Por el tipo de problema planteado, se identificó como variable de salida ‘RainTomorrow’ y las otras 22 variables son de entrada. También que el problema es de clasificación, donde la salida será 1 o 0 según se prediga que llueve mañana o no con datos del día de hoy.

Posteriormente se analizaron las variables de entrada según el tipo de dato:

* **Numéricas (16 en total)**

Utilizando el método describe y las funciones “distplot” y “boxplot” de la librería Seaborn se obtuvieron el dominio, valores máximos / mínimos, la distribución y el diagrama “Box-Whisker”. Adjuntamos tabla resumen y gráficas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Valor** | **Min** | **Max** | **Distribución** |
| MinTemp | Real | -8.5 | 33.9 | Aproximadamente gaussiana. Sesgo leve hacia la derecha |
| MaxTemp | Real | -4.8 | 48.1 | Aproximadamente gaussiana. Sesgo no tan leve hacia la derecha |
| Rainfall | Real | 0 | 371 | Irregular. Fuertemente puntiaguda en valores cercanos a 0. Con gran cantidad de valores extremos |
| Evaporation | Real | 0 | 145 | Puntiaguda en valores menores a 10. Con gran cantidad de valores extremos |
| Sunshine | Real | 0 | 14.5 | Irregular |
| WindGustSpeed | Real | 6 | 135 | Aproximadamente gaussiana con elevado nivel de ruido |
| WindSpeed9am | Real | 0 | 130 | Aproximadamente gaussiana con elevado nivel de ruido |
| WindSpeed3pm | Real | 0 | 87 | Aproximadamente gaussiana con elevado nivel de ruido |
| Humidity9am | Real | 0 | 100 | Gaussiana recortada con pico adicional en 100 |
| Humidity3pm | Real | 0 | 100 | Aproximadamente gaussiana |
| Pressure9am | Real | 980.5 | 1041 | Gaussiana |
| Pressure3pm | Real | 977.1 | 1039 | Gaussiana |
| Cloud9am | Entero | 0 | 9 | Aproximadamente gaussiana (a valores discretos) bimodal |
| Cloud3pm | Entero | 0 | 9 | Aproximadamente gaussiana (a valores discretos) bimodal |
| Temp9am | Real | -7.2 | 40.2 | Gaussiana |
| Temp3pm | Real | -5.4 | 46.7 | Gaussiana |

Gráfico

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

* **Categóricas (5 en total)**

Utilizando el método “describe” y “groupby” con “count” se obtuvieron el tipo de valor, la cardinalidad y la representación de cada categoría. La siguiente tabla resume los resultados.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Tipo** | **Informativa** | **Cardinalidad** | **Todas las categorías representadas** | **Comentarios** |
| Location | Nominal | Si | 49 | Si | Se puede codificar con one hot encoding (no es recomendable debido a la cardinalidad). También podríamos utilizar ordinal enconding o mean encoding |
| WindGustDir | Nominal | Si | 16 | Si | Podríamos codificar con one hot encoding o codificar numéricamente según el cuadrante de dirección del viento. |
| WindDir9am | Nominal | Si | 16 | Si | Podríamos codificar con one hot encoding o codificar numéricamente según el cuadrante de dirección del viento. |
| WindDir3pm | Nominal | Si | 16 | Si | Podríamos codificar con one hot encoding o codificar numéricamente según el cuadrante de dirección del viento. |
| RainToday | Nominal | Si | 2 | Si | Codificamos con integer encoding donde 0 o 1 representan “No” o “Yes” respectivamente. |

* **Otros**

La variable “date” es de tipo fecha, con formato YYYY-MM-DD. Con ese formato no es posible utilizarla para entrenar los modelos, es necesario codificarla. Se propone como método de codificación extraer el únicamente el mes y utilizarlo como variable numérica. Otro método sería convertir la fecha a día del año.

* **Variable de salida**

La variable de salida se analizó mediante los métodos “describe” y “groupby / count”. Al ser un problema de clasificación binaria, es conveniente codificar la salida con 0 y 1 representando “No” y “Yes” respectivamente.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Tipo** | **Cardinalidad** | **Balance** |
| RainTomorrow | Categórica nominal | 2 | Desbalanceada con mayor cantidad de “No” |

# Apartado 2

Se utilizó la función “train\_test\_split” de la librería “sklearn” para particionar el dataset en 70% entrenamiento, 15% para validación de las técnicas y métodos que utilicemos y un 15% para un testeo final de la performance.

# Apartado 3

A fin de conocer los valores faltantes para cada variable, se utilizó el método “isna().sum()”. ESTO SOBRE TRAIN O TODO?

*Date 0 0 %*

*Location 0 0 %*

*MinTemp 1485 1.02 %*

*MaxTemp 1261 0.87 %*

*Rainfall 3261 2.24 %*

*Evaporation 62790 43.17 %*

*Sunshine 69835 48.01 %*

*WindGustDir 10326 7.1 %*

*WindGustSpeed 10263 7.06 %*

*WindDir9am 10566 7.26 %*

*WindDir3pm 4228 2.91 %*

*WindSpeed9am 1767 1.21 %*

*WindSpeed3pm 3062 2.11 %*

*Humidity9am 2654 1.82 %*

*Humidity3pm 4507 3.1 %*

*Pressure9am 15065 10.36 %*

*Pressure3pm 15028 10.36 %*

*Cloud9am 55888 38.42 %*

*Cloud3pm 59358 40.81 %*

*Temp9am 1767 1.21 %*

*Temp3pm 3609 2.48 %*

*RainToday 3261 2.24 %*

*RainTomorrow 3267 2.25 %*

Posteriormente, con el método “head”, se inspeccionó cada columna para identificar los motivos de la ausencia de valores. También se utilizó el método “isna().sum()” combinando los “NA” de diferentes variables en simultáneo. Las conclusiones que se obtuvieron fueron:

* RainTomorrow. El motivo de la falta de valores es desconocido (MCAR). Dado que es la salida en el problema “supervisado” planteado, se decide eliminar estos registros.
* Evaporation, Sunshine, Pressure9am, Pressure3pm, Cloud9am y Cloud3pm. La gran cantidad de datos faltantes se debe a que en algunas ciudades no se realizan mediciones de una o varias de estas variables (MNAR). Dada la gran proporción, se propone utilizar imputación por valor arbitrario fuera de la distribución o por valor de fin de cola.
* WindGustDir, WindGustSpeed, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Temp9am, Temp3pm. En estos casos la mayor parte de valores faltantes se debe a que hay días al azar donde no se miden ciertas combinaciones como velocidad y dirección del viento, o temperatura y humedad (MAR). En su mayoría las proporciones de faltantes de estas variables es bajo, por lo que proponemos la imputación por mediana o moda para las numéricas y categoría faltante o frecuente para las categóricas.
* MinTemp, MaxTemp, WindGustDir9am, WindGustDir3pm, Rainfall, RainToday. El motivo de la falta de valores es totalmente desconocido y aleatorio (MCAR). Dado el bajo porcentaje de valores faltantes, se propone imputación con media, mediana o moda para las variables numéricas y categoría faltante o frecuente para las categóricas.

Nota: en un análisis mas elaborado se podría utilizar métodos multivariable como MICE para, por ejemplo, calcular el valor de Temp3pm a partir del resto de las temperaturas.