1. Analisis exploratorio inicial

Iniciamos cargando el dataset con pandas y luego utilizamos los métodos “head” y “describe2 y las propiedades “shape” y “dtypes” para realizar un primer resume.

De lo anterior se identifica que tenemos un dataset con 23 columnas y alrededor de 145 mil registros. Se observa también a priori que las columnas tienen valores faltantes y que tenemos variables de con diferentes tipos de datos.

PONER TODAS LAS COLUMNAS CON EXPLCIACION QUE SON

Por el tipo de problema planteado, la variable de salida es ‘RainTomorrow’ y el resto son variables de entrada. Identificamos también que el problema es de clasificación, donde la salida será 1 o 0 según se prediga que llueve mañana o no según datos del día de hoy.

Hacemos un análisis de las variables de entrada separando por tipo de dato:

* Numéricas:

Cantidad total = 16

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Tipo de valor** | **Valor Min** | **Valor 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 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 (valores discretos) bimodal |
| Cloud3pm | Entero | 0 | 9 | Aproximadamente gaussiana (valores discretos) bimodal |
| Temp9am | Real | -7.2 | 40.2 | Gaussiana |
| Temp3pm | Real | -5.4 | 46.7 | Gaussiana |

PONE GRAFICA DISTRIBUCIONES

* Categóricas:

Cantidad total = 5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Tipo** | **Informativa** | **Cardinalidad** | **Todas las categorías representadas** | **Comentarios** |
| Ciudad | 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:

Cantidad total = 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Tipo** | **Informativa** | **Cardinalidad** | **Comentarios** |
| Date | Fecha | Si | 3436 | Filtraremos el campo utilizando únicamente el mes. También se podría utilizar el día del año. |

Variables de Salida:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Tipo** | **Cardinalidad** | **Balance** | **Comentarios** |
| RainTomorrow | Categórica nominal | 2 | Desbalanceada con mayor cantidad de “No” | Al ser un problema de clasificación binaria, es conveniente codificar la salida con integer encoding donde 0 y 1 representan “No” y “Yes” respectivamente. |

1. Esquema de validación de resultados.

Se particiona el dataset en entrenamiento, validación y testeo. Los dataset de entrenamiento y validación se utilizarán para entrenar y validar los métodos aplicados sobre los datos y el dataset de testeo se utilizará sobre el modelo final para evaluar la performance obtenida.

La partición se realiza 70% entrenamiento, 15% validación y 15% testeo.

1. Limpieza y preparación de datos / ingeniería de features

Análisis de datos faltantes:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Cantidad NaN | Porcentaje NaN | Clase  NaN | Motivo | Método imputación propuesto |
| Date | 0 | 0 | - | - | - |
| Location | 0 | 0 | - | - | - |
| MinTemp | 1485 | 1.02 | MCAR | Desconocido | Por mediana / moda. MICE Con valores 9am y 3pm |
| MaxTemp | 1261 | 0.87 | MCAR | Desconocido | Por mediana / moda. MICE Con valores 9am y 3pm |
| Rainfall | 3261 | 2.24 | MCAR | Desconocido | Por mediana / moda. MICE. Valores aleatorios |
| Evaporation | 62790 | 43.17 | MCAR MNAR | Hay ciudades donde no se mide la evaporación | Valor arbitrario o de fin de cola |
| Sunshine | 69835 | 48.01 | MCAR MNAR | Hay ciudades donde no se mide las horas solares | Valor arbitrario o de fin de cola |
| WindGustDir | 10326 | 7.1 | MCAR  MAR | En ciertos días no se midió ni la dirección ni la velocidad del viento. | Categoría faltante o categoría frecuente. MICE con valores de 9am y 3pm. |
| WindGustSpeed | 10263 | 7.06 | MCAR  MAR | En ciertos días no se midió ni la dirección ni la velocidad del viento. | Por mediana / moda, aunque el porcentaje es mayor a 5%. MICE con valores de 3am y 9pm |
| WindGustDir9am | 10566 | 7.26 | MCAR | Desconocido | Categoría faltante o frecuente |
| WindGustDir3pm | 4228 | 2.91 | MCAR | Desconocido | Categoría faltante o frecuente |
| WindSpeed9am | 1767 | 1.21 | MCAR  MAR | En ciertos días no se midió ni la dirección ni la velocidad del viento a las 9 am. | Imputación por mediana o moda. Muestreo aleatorio |
| WindSpeed3pm | 3062 | 2.11 | MCAR  MAR | En ciertos días no se midió ni la dirección ni la velocidad del viento a las 3pm. | Imputación por mediana o moda. Muestreo aleatorio |
| Humidity9am | 2654 | 1.82 | MCAR | Desconocido | Imputación por mediana o moda. Muestreo aleatorio |
| Humidity3pm | 4507 | 3.1 | MCAR | Desconocido | Imputación por mediana o moda. Muestreo aleatorio |
| Pressure9am | 15065 | 10.36 | MCAR MNAR | Hay ciudades donde no se mide la presión | Valor arbitrario o de fin de cola |
| Pressure3pm | 15028 | 10.33 | MCAR MNAR | Hay ciudades donde no se mide la presión | Valor arbitrario o de fin de cola |
| Cloud9am | 55888 | 38.42 | MCAR MNAR | Hay ciudades donde no se mide la nubosidad | Valor arbitrario o de fin de cola |
| Cloud3pm | 59358 | 40.81 | MCAR MNAR | Hay ciudades donde no se mide la nubosidad | Valor arbitrario o de fin de cola |
| Temp9am | 1767 | 1.21 | MCAR  MAR | Hay días donde no se mide la temperatura ni la humedad | Imputación por mediana o moda. Muestreo aleatorio |
| Temp3pm | 3609 | 2.48 | MCAR  MAR | Hay días donde no se mide la temperatura ni la humedad | Imputación por mediana o moda. Muestreo aleatorio |
| RainToday | 3261 | 2.24 | MNAR | En ese día no se tiene registro de Rainfall | ¿???? |
| *Todas las variables de entrada en simultáneo* | 182 | - | MCAR | Desconocido | Eliminar estos registros |
| RainTomorrow | 3267 | 2.24 | MCAR | Desconocido | Eliminar estos registros |

Adicional a lo enunciado en la tabla anterior, agregamos otro caso donde, luego de imputar las columnas con un % de NaN >10, hacemos un análisis de caso completo. Los resultados de las técnicas propuestas se muestran en el punto 4.

Respecto a los métodos de codificación elegidos, se mencionaron en las tablas X Y Z.

Para comprobar las relaciones entre las señales de entrada utilizamos la matriz de correlación, que por defecto se basa en el coeficiente de Pearson. Para poder aplicar este método, dado que se utiliza para relaciones entre variables numéricas, previamente codificamos las variables de entrada que lo precisaban. Esta codificación nos puede afectar la relación pero dado que la correlación la utilizaremos como un indicativo no es problema.

PONER MATRIZ

Para destacar:

* Las temperaturas mínimas, máximas, 9am y 3pm están altamente correlacionadas.
* La velocidad máxima del viento esta levemente correlacionada con la velocidad a las 9am y 3pm.
* Las humedades a las 9am y 3pm están levemente correlacionadas.
* Las presiones 9am y 3pm están altamente correlacionadas.
* Las nubosidades a las 9am y 3pm están correlacionadas.

Para identificar las variables de entrada de mayor importancia se aplicó información mutua y método de Lasso. En información mutua se mide el coeficiente de “información mutua” entre las variables de entrada y la variable de salida y se selecciona aquellas variables de entrada con mayor dependencia con la salida.

PONER GRAFICA

Utilizando este estadístico, concluimos que las variables de mayor interés son Humidity3pm, Cloud3pm, RainToday, Rainfall y Sunshine.

Como segundo método proponemos aplicar Lasso. Lasso es un método de regularización de regresión lineal cuya característica es que tiende a cero la componente de multiplicación de aquellos parámetros que no son importantes en la relación lineal. Observando los parámetros finales de la regresión lineal, podemos obtener una idea de las variables más importantes.

PONER IMAGEN

Aplicando este método, observamos que las variables más importantes son RainToday, Humidity3pm, Pressure3pm Cloud3pm, WindGustSpeed y Rainfall. La diferencia con el método anterior es que Lasso se basa en relaciones lineales. Dada la diferencia podemos suponer que un método lineal no va a tener buena performance.