Análisis de Datos, TP Integrador

Alumno: Silva Plata, Bruno Fernando

1.- Análisis exploratorio inicial

En esta etapa se exploró el dataset:

- La información:

```
 #   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
```

- La descripción:

	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

Variables categóricas
Primero se vio acerca de la cardinalidad de las columnas categóricas, donde se obtuvo los siguientes datos:

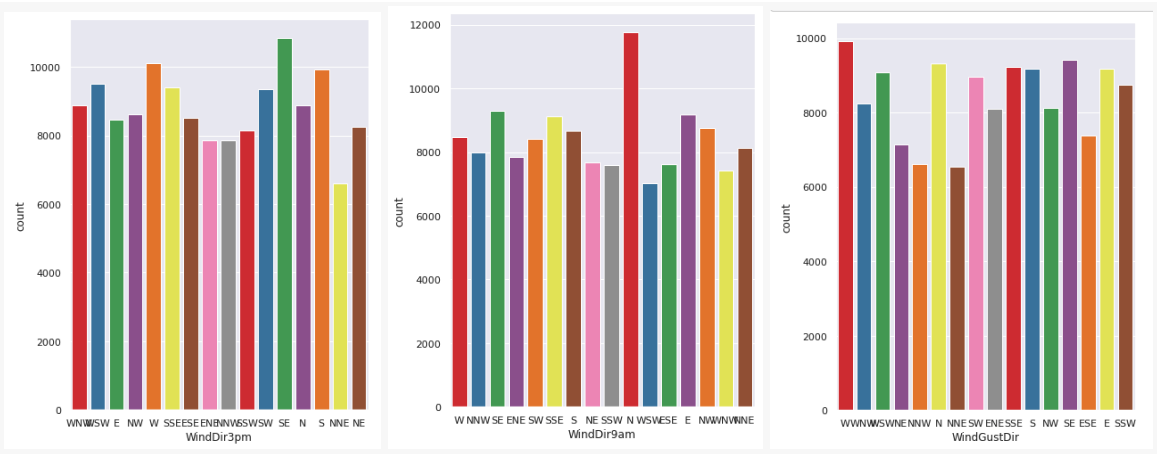
```
Date: 3436 etiquetas
Location: 49 etiquetas
WindGustDir: 17 etiquetas
WindDir9am: 17 etiquetas
WindDir3pm: 17 etiquetas
RainToday: 3 etiquetas
RainTomorrow: 3 etiquetas
```

Para facilitar el feature engineering más adelante se realizó un tratamiento para Date, donde se obtiene directamente el mes y así evitar la alta cardinalidad.

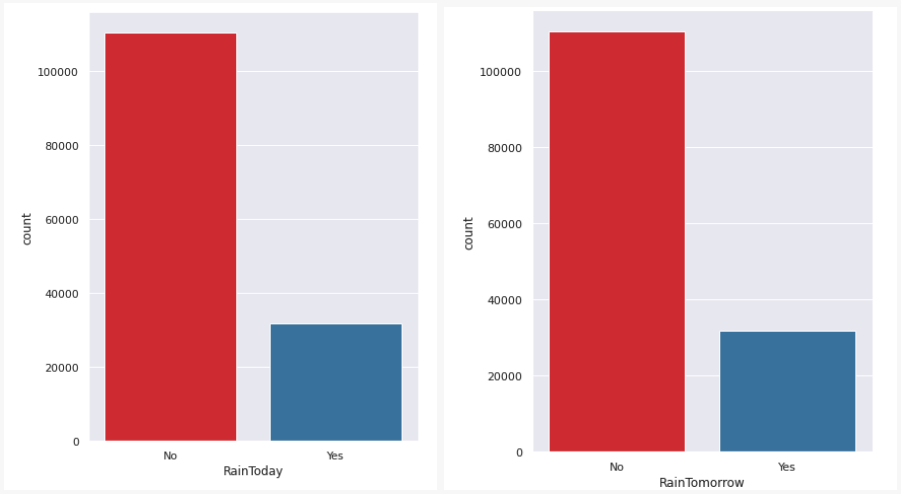
En el caso de location (la segunda columna con mayor cardinalidad) se realizó su transformación a coordenadas con Geopy.goecorders, según la siguiente literatura

- https://amaral.northwestern.edu/blog/getting-long-lat-list-cities
- https://peterhaas-me.medium.com/how-to-geocode-with-python-and-pandas-4cd1d717d3f7
- https://geopy.readthedocs.io/en/stable/

Se visualizó la frecuencia de etiquetas en las columnas: WindGustDir, WindDir9am, WindDir3pm.



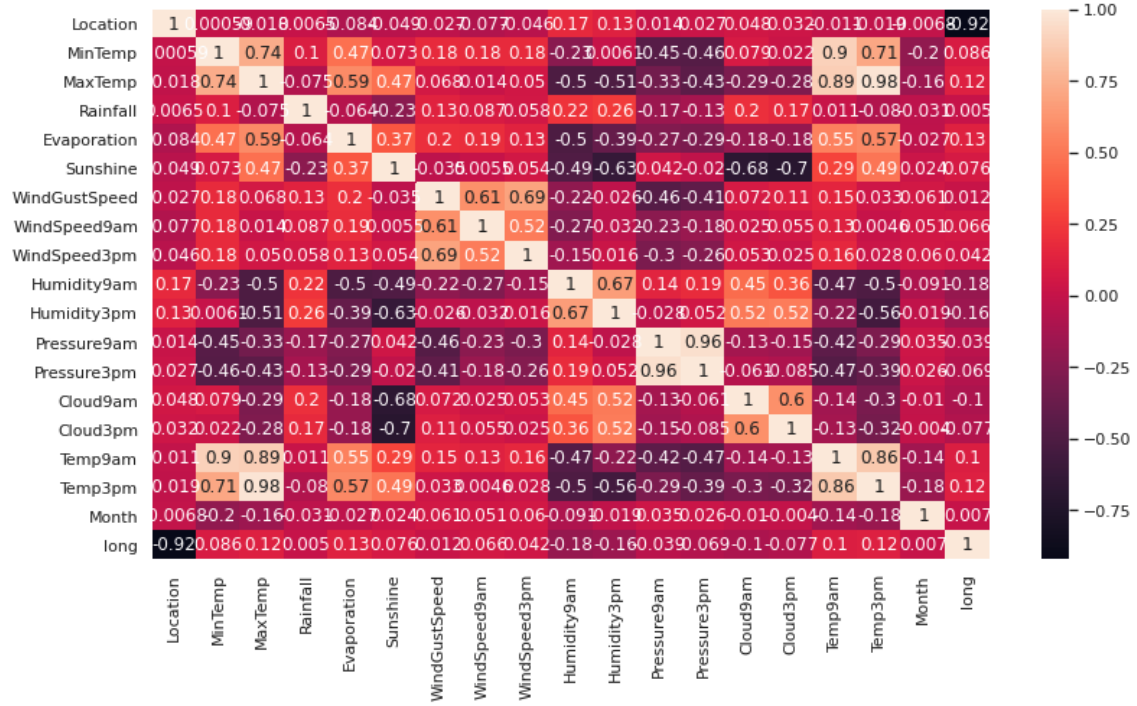
Se visualizo a su vez la frecuencia de etiquetas en RainToday, RainTomorrow:



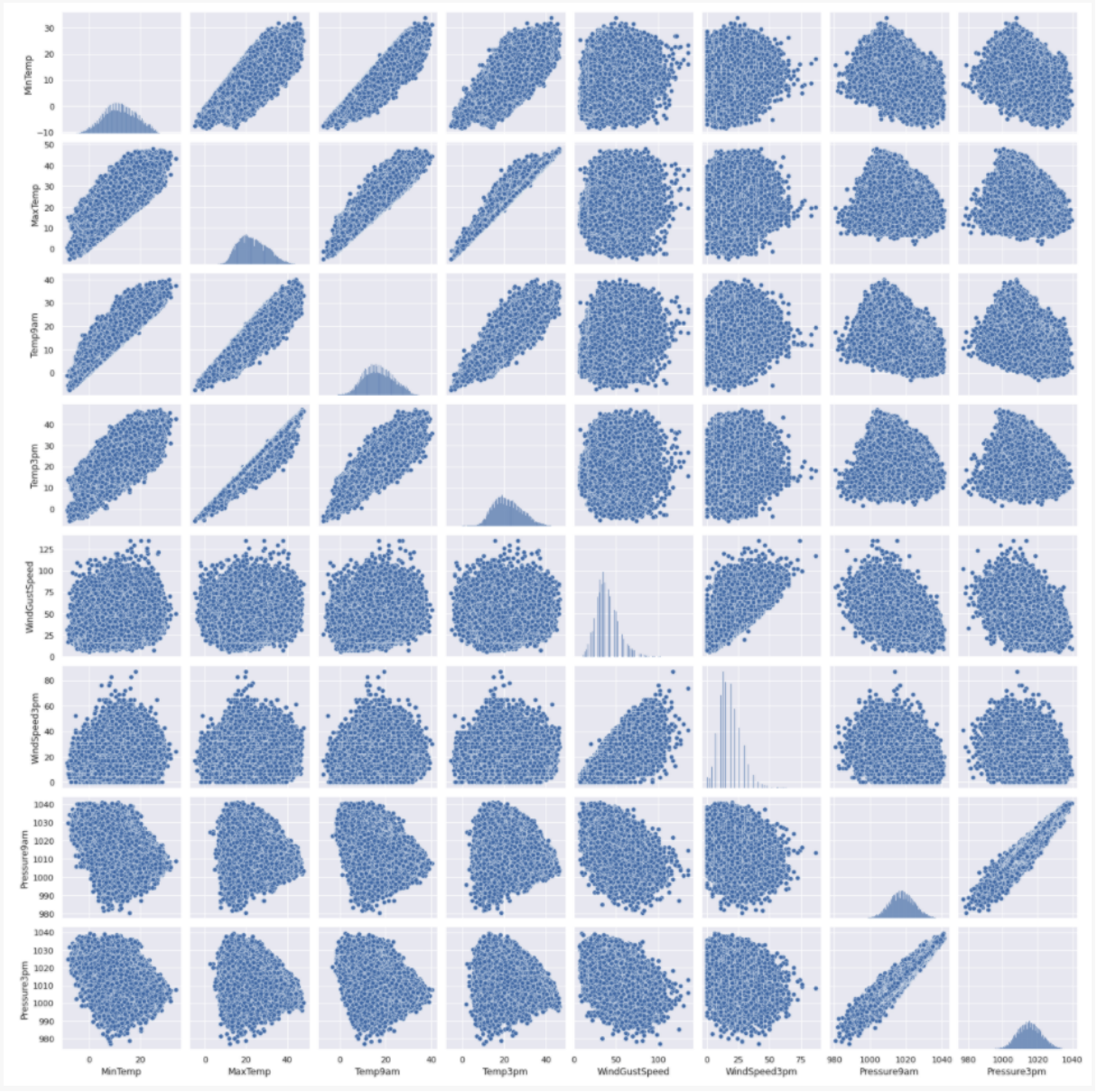
Vemos que la columna “RainTomorrow” tiene un desbalance en los datos. Posteriormente en la etapa de las métricas se utiizará F1 Score para contrarrestar el desbalance.

Variables numéricas

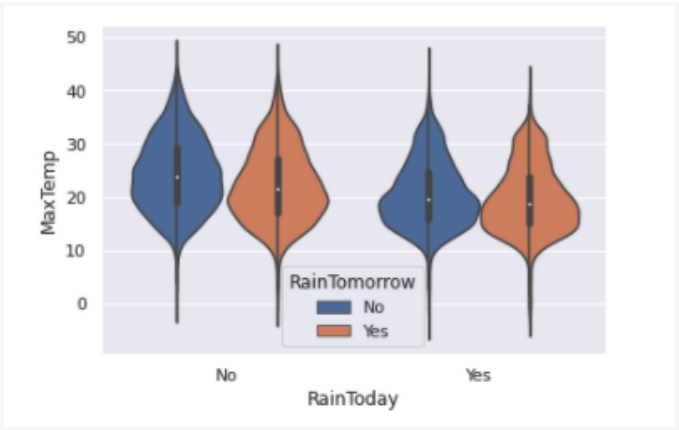
Se visualizó la matriz de correlación entre las variables:



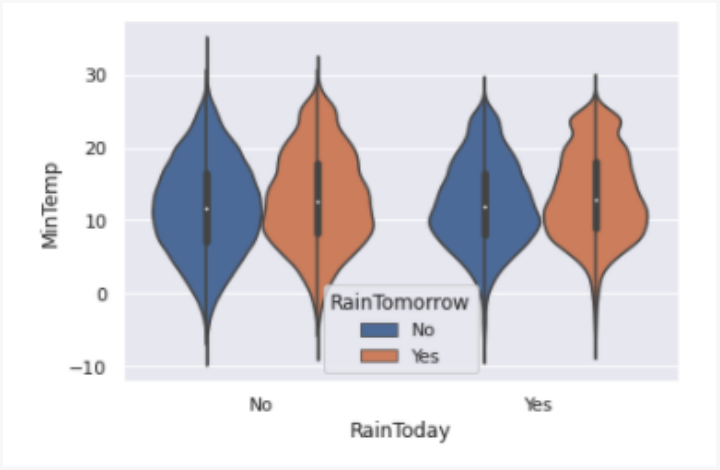
Se observó que las columnas 'MinTemp', 'MaxTemp', 'Temp9am', 'Temp3pm', 'WindGustSpeed', 'WindSpeed3pm', 'Pressure9am', 'Pressure3pm' son las que más correlación tienen, por lo que se visualizó una gráfica del mismo:



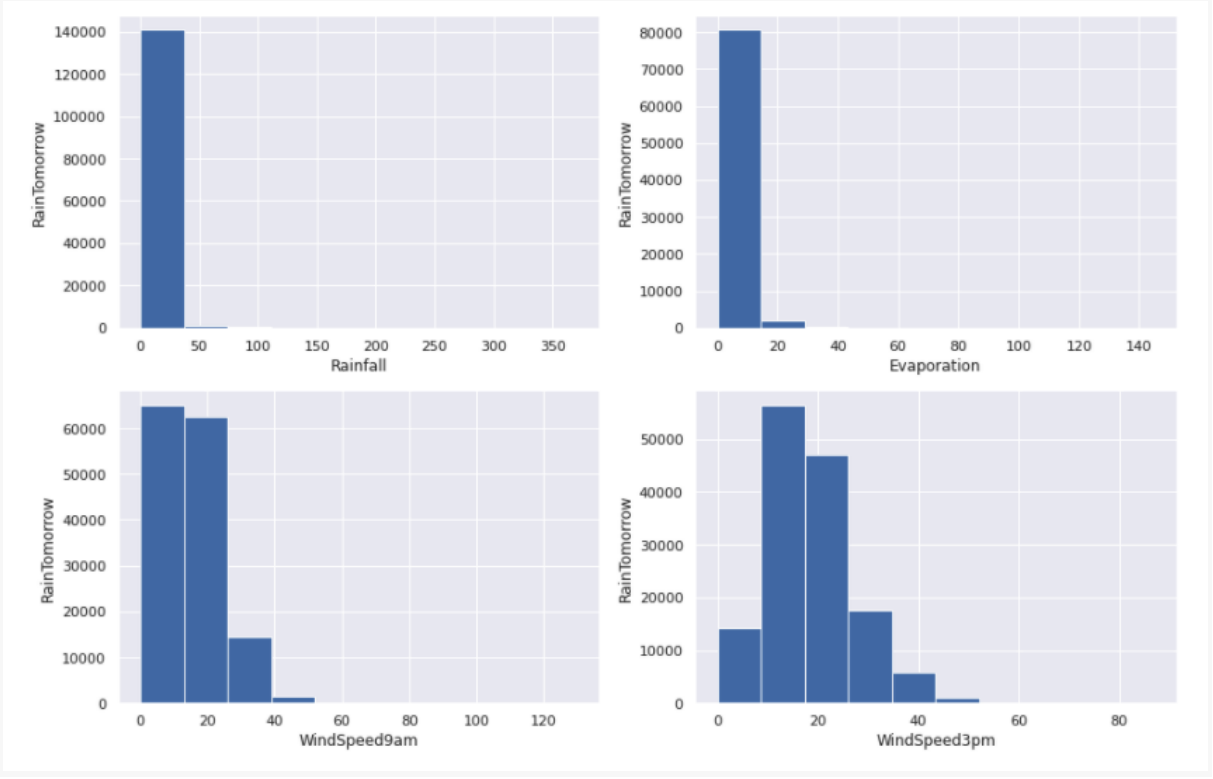
Se realizó un Análisis Bivariable entre:
MaxTemp y RainToday



MinTemp y RainToday

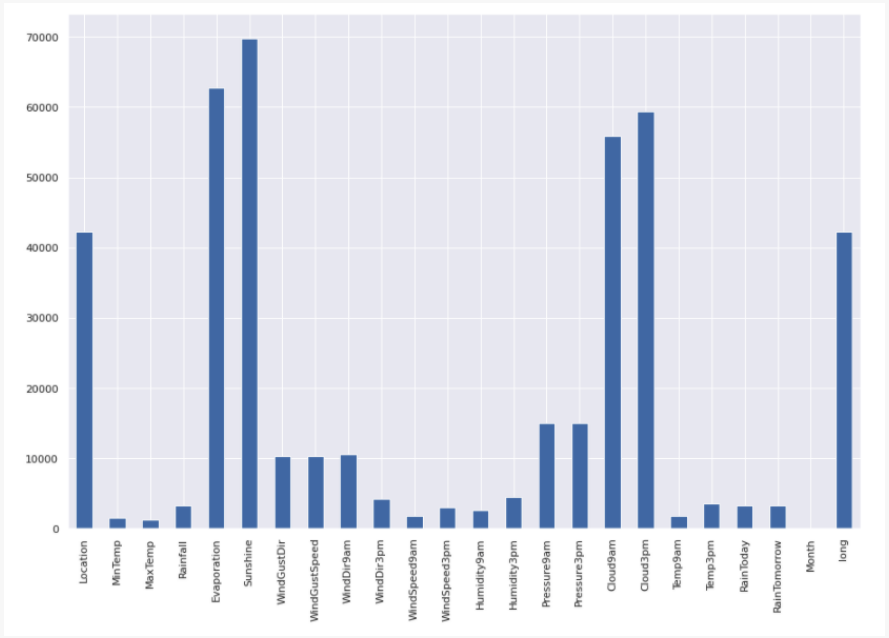


Se vio por histograma la distribución de ciertas columnas:



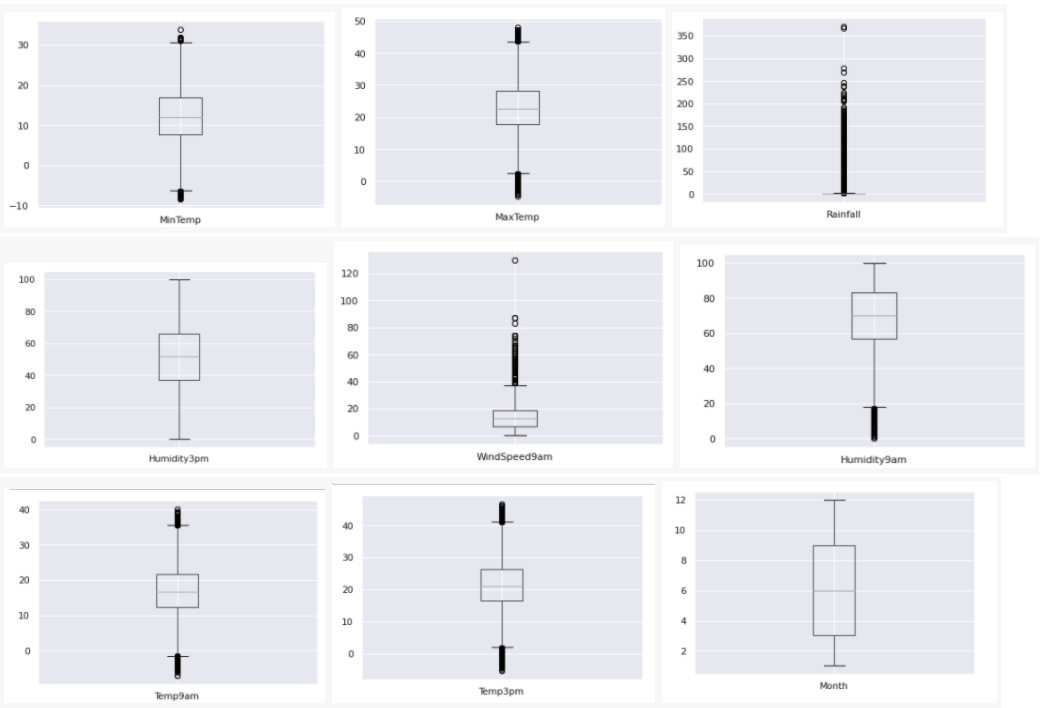
2.- Limpieza y preparación de datos / ingeniería de features

Primero se visualizó la cantidad de valores nulos por columna:



Se eliminaron las columnas: 'Evaporation', 'Sunshine', 'Cloud9am', 'Cloud3pm' por tener más del 35% de registros nulos.

Se visualizaron Box Plots para las columnas numéricas, y en base a eso realizar el llenado de datos con la media o mediana, todo acorde al gráfico:



Para el tratamiento de la variables categóricas se realizaron dos técnicas:

- La primera fue one hot encoding en las columnas de ‘WindGustDir","WindDir9am","WindDir3pm"

- La segunda fue convertir a 0,1 los valores en la columna target (RainTomorrow) y rellenar los datos faltantes con KNN imputer. Se siguió la siguiente literatura:
-
    - https://towardsdatascience.com/preprocessing-encode-and-knn-impute-all-categorical-features-fast-b05f50b4dfaa
    - https://machinelearningmastery.com/knn-imputation-for-missing-values-in-machine-learning/

3.- Esquema de validación

Se realizo la partición de los datos en 30% para el conjunto de testeo y 70% para el conjunto de entrenamiento.

4.- Entrenamiento del modelo

Para el entrenamiento del modelo se seleccionó:

- Regresión Logística
- Random Forest

Donde se obtuvo los siguientes resultados:

Modelo	Accuracy
Logistic Regression	83,59%
Random Forest	84,91%

5.- Resultados del modelo

Se usó como métrica F1 Score, esto por el desbalance que se vio previamente en la columna target. Los resultados de F1 Score con average “weighted” son los siguientes:

Modelo	F1 Score, average “weighted”
Logistic Regression	81,91%
Random Forest	83,41%

También se visualizó la tabla de classifcation report:

Regresión Logística:

	precision	recall	f1-score	support
0	0.86	0.95	0.90	23907
1	0.00	0.00	0.00	237
2	0.00	0.00	0.00	102
3	0.71	0.48	0.57	6721
accuracy			0.83	30967
macro avg	0.39	0.36	0.37	30967
weighted avg	0.82	0.83	0.82	30967

Random Forest:

	precision	recall	f1-score	support
0	0.86	0.96	0.91	23907
1	0.83	0.16	0.27	237
2	0.71	0.15	0.24	102
3	0.76	0.49	0.59	6721
accuracy			0.85	30967
macro avg	0.79	0.44	0.50	30967
weighted avg	0.84	0.85	0.83	30967

6.- Conclusiones

Se visualizaron los datos para tener una imagen general de que procedimientos aplicar a la misma.

Se implementaron distintos métodos en el tratamiento de los datos para obtener un resultado óptimo. Por el lado de las columnas categóricas se trato de reducir la cardinalidad en las columnas que presentaban mayor número de etiquetas. Por el lado de las columnas numéricas se reemplazaron los datos faltantes con la media y mediana, dependiendo de cada caso.

Se entrenó con dos modelos para clasificación ya que la columna target era de este tipo de Machine Learning Supervisado. Se usó F1 Score para tener una métrica más realista, ya que los datos en la columna target estaban desbalanceados.

Los resultados obtenidos son buenos tanto en Regresión Logística como en Random Forest, aunque pudo ser mejor. Un factor que pudo haber influido es el desglose de las columnas correspondientes a Wind, un distinto tratamiento quizá influiría en los resultados.

.