SEGUNDA ENTREGA AI

Santiago Salazar Osorio
Julian David Gil Botero
Juan Camilo Cardona Castaño

Para abordar el análisis de los datos, se inició examinando las propiedades fundamentales del conjunto de datos, lo que incluyó la evaluación de su forma (shape) y su estructura mediante info(). Uno de los aspectos cruciales de esta exploración temprana fue identificar la presencia de datos faltantes, lo cual es esencial para tomar decisiones informadas en el preprocesamiento de los datos.

En un primer paso, se elaboró la Tabla 1 en la que se calculó el porcentaje de datos faltantes para cada atributo. Aquellos atributos que exhibieron un porcentaje de datos faltantes superior al 10% fueron excluidos del análisis. Esta decisión se basó en la premisa de que un alto nivel de datos faltantes puede afectar la integridad y la utilidad de las variables.

Además, se observó que las columnas "RainToday" y "RainTomorrow" contenían datos faltantes. Dado el carácter crítico de estas variables para el problema de predicción de lluvia, se optó por eliminar las filas que contenían datos faltantes en estas columnas.

Esto se hizo para garantizar que el modelo se entrenara con datos de calidad y para evitar la especulación o imputación de valores en relación con la variable de interés.

A pesar de los esfuerzos realizados para limpiar y preparar el conjunto de datos, se encontró que todavía existían 5747 filas con valores faltantes en diferentes atributos. En lugar de intentar llenar estos datos faltantes, se tomó la decisión de eliminar por completo estas filas. Esta elección se basó en la premisa de que, a pesar de la reducción en el tamaño del conjunto de datos, se lograban resultados sólidos y confiables con el modelo, evitando la introducción de "ruido" al tratar de imputar valores en lugar de contar con observaciones completas y de alta calidad.

Este enfoque de preprocesamiento de datos fue fundamental para garantizar que el modelo de predicción se entrenara con información confiable y de calidad, lo que a su vez contribuyó a la obtención de resultados sólidos en la tarea de predicción de lluvia en Australia.

Tabla 1. Información general del DataFrame.

COLUMN	NON-NULL COUNT DTYPE	NULL DATA	%MISSING
Location	145460	0	0,00%
MinTemp	143975	1485	1,02%
MaxTemp	144199	1261	0,87%
Rainfall	142199	3261	2,24%
Evaporation	82670	62790	43,17%
Sunshine	75625	69835	48,01%

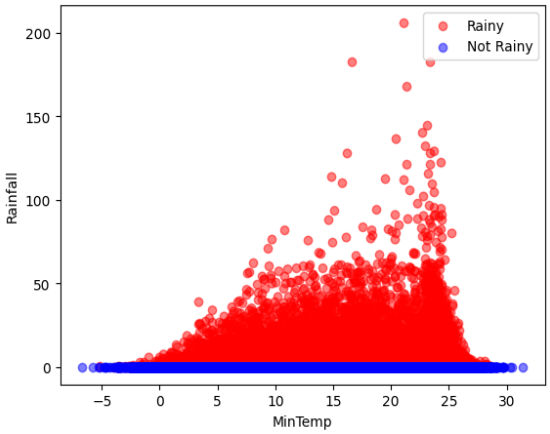
WindGustDir	135134	10326	7,10%
WindGustSpeed	135197	10263	7,06%
WindDir9am	134894	10566	7,26%
WindDir3pm	141232	4228	2,91%
WindSpeed9am	143693	1767	1,21%
WindSpeed3pm	142398	3062	2,11%
Humidity9am	142806	2654	1,82%
Humidity3pm	140953	4507	3,10%
Pressure9am	130395	15065	10,36%
Pressure3pm	130432	15028	10,33%
Cloud9am	89572	55888	38,42%
Cloud3pm	86102	59358	40,81%
Temp9am	143693	1767	1,21%
Temp3pm	141851	3609	2,48%
RainToday	142199	3261	2,24%
RainTomorrow	142193	3267	2,25%

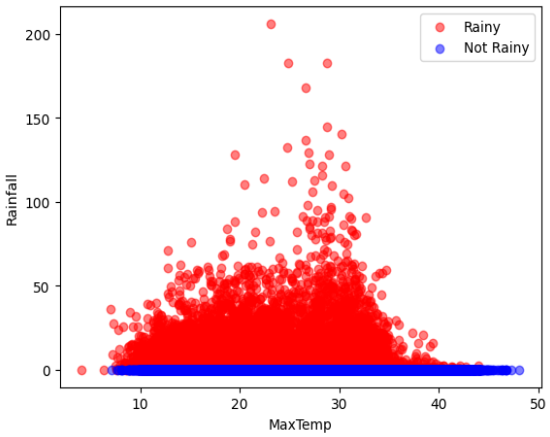
Se realizaron diversas visualizaciones gráficas con el propósito de explorar y entender más profundamente los datos. Entre las representaciones visuales empleadas, destacan un "pairplot" creado con la biblioteca Seaborn (sns), así como un mapa de calor.

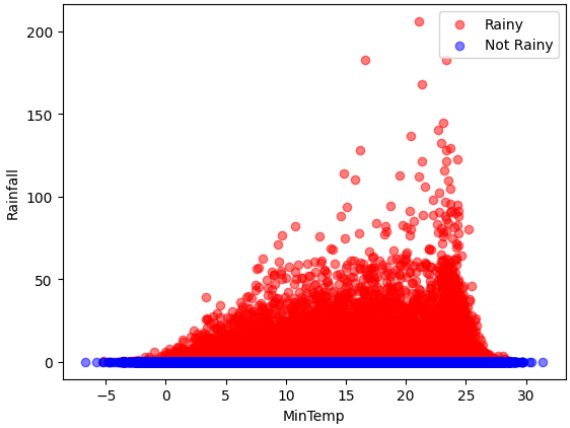
Adicionalmente, se llevó a cabo un análisis para determinar en qué ciudad se registra una mayor cantidad de lluvia. Para esta tarea, se realizaron gráficas que representaban todas las variables con "Rainfall" en el eje Y. Este enfoque permitió una visualización más efectiva de los datos en comparación con la variable "RainToday." La premisa subyacente en esta representación gráfica es que cuando el valor de "Rainfall" es igual a 0, generalmente se considera que no ha llovido.

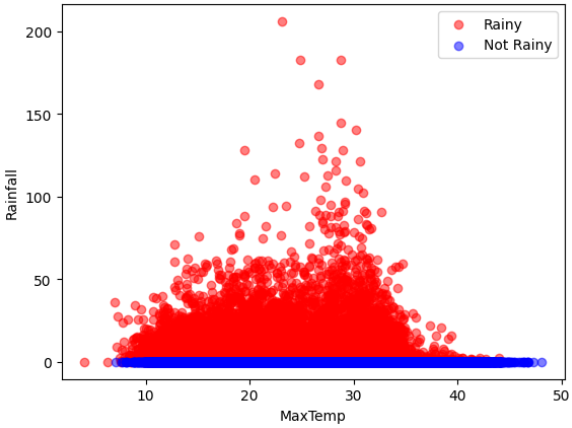
Esta técnica facilitó la distinción entre los eventos de lluvia y las condiciones de ausencia de lluvia.Estas visualizaciones desempeñaron un papel esencial en la fase exploratoria del análisis de datos, brindando información valiosa sobre la distribución de lluvia en diferentes localidades y permitiendo una mayor comprensión de las relaciones entre las variables clave, lo que a su vez facilitó la toma de decisiones en el proceso de modelado.

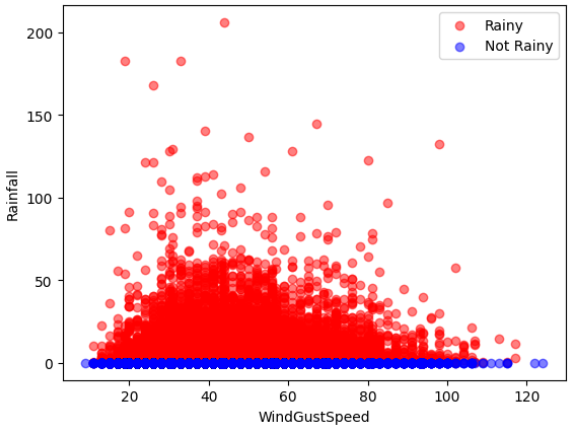
Graficas de la distribución entre diferentes variables numericas.

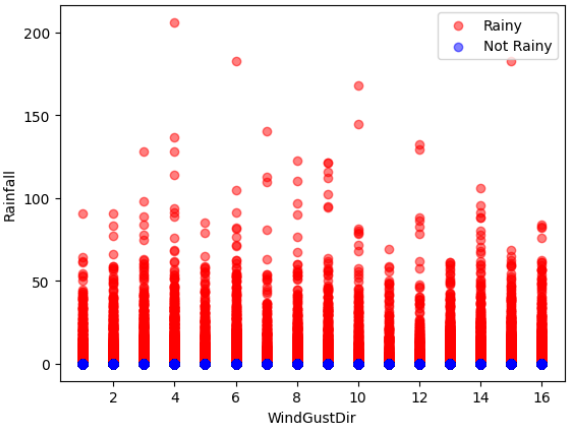


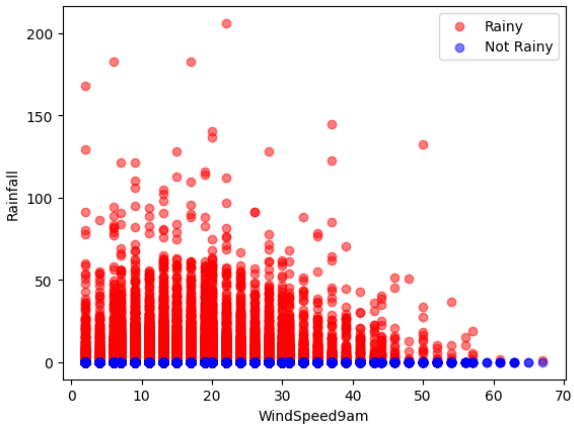


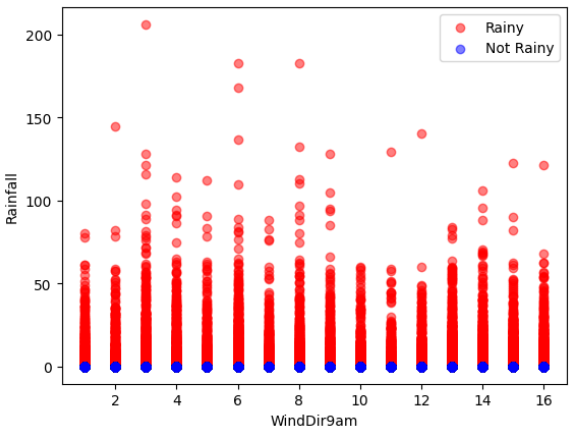










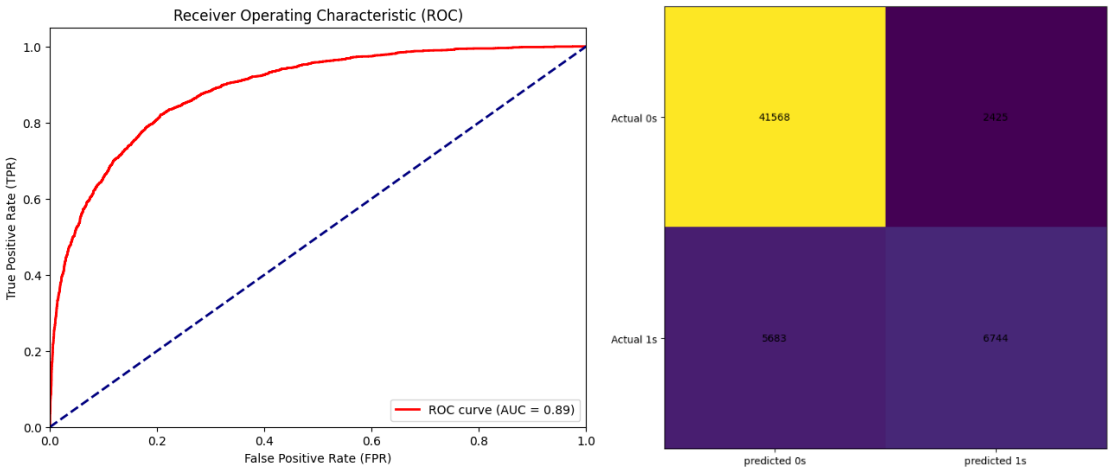


Modelo de predicción

En el contexto del problema de clasificación que busca predecir si va a llover o no en Australia, se han aplicado modelos de regresión lineal con diversos solvers, entre ellos "sag," "newton-cg," "liblinear," y "saga." Cada uno de estos solvers es un algoritmo de optimización utilizado para entrenar el modelo. A continuación, se describen brevemente:

- "sag": Stochastic Average Gradient Descent. Es un algoritmo de descenso de gradiente estocástico que se utiliza para entrenar modelos de regresión logística.
- "newton-cg": Este algoritmo utiliza el método de Newton-Conjugate Gradient para la optimización de la función de costo. Es especialmente útil cuando se trabaja con problemas de clasificación.
- "liblinear": Es una biblioteca de software para la optimización de problemas de clasificación lineal. Es adecuada para conjuntos de datos pequeños y medianos.

Para evaluar el rendimiento de estos modelos, se realizaron divisiones de los datos en diferentes porcentajes de entrenamiento y prueba: 10%, 15%, 20%, 25%, y 30%. En cada caso, se entrenó el modelo utilizando el solver correspondiente y se midió el rendimiento utilizando dos métricas clave:
Accuracy (Exactitud): Esta métrica indica la proporción de predicciones correctas realizadas por el modelo en el conjunto de prueba. En otras palabras, mide con qué precisión el modelo clasifica las muestras.
Score (Puntuación): En el contexto de la regresión logística, la puntuación se refiere a la precisión general del modelo en todo el conjunto de datos, no solo en el conjunto de prueba. Es una medida del rendimiento general del modelo.
Este enfoque permite comparar cómo varía el rendimiento de los modelos en función del solver utilizado y el tamaño de los conjuntos de entrenamiento y prueba. La elección del solver y la división de datos adecuada son fundamentales para lograr un buen rendimiento en la predicción de lluvia en Australia.



Curva ROC y Matriz de confusión.

La representación de la curva ROC (Receiver Operating Characteristic) junto con el valor del área bajo la curva (AUC) resultaron ser herramientas fundamentales en la evaluación del modelo de clasificación. En particular, se observó que el modelo con el solver

"newton-cg" logró un desempeño superior en comparación con otras configuraciones, obteniendo un AUC de 0.89.

Para contextualizar, es importante tener en cuenta que el AUC es una métrica que varía entre 0 y 1, donde un valor de 0.5 indica un modelo completamente aleatorio y un valor de 1 representa un modelo perfecto. En este caso, el AUC de 0.89 sugiere que el modelo tiene una capacidad sustancial para distinguir entre las clases objetivo, lo que es un indicativo positivo de su rendimiento.

Sin embargo, es vital recordar que a medida que se aumenta la Tasa de Verdaderos Positivos (TPR) en la curva ROC, también se incrementa la Tasa de Falsos Positivos (FPR). Este equilibrio es crucial, ya que un alto TPR podría llevar a un aumento no deseado en el FPR, lo que podría resultar en clasificaciones incorrectas o falsos positivos. Por lo tanto, se debe abordar el análisis de la curva ROC con precaución y considerar el equilibrio entre la sensibilidad y la especificidad.

$$TPR = \frac{TP}{TP + FN} = 0.54$$

$$FPR = \frac{FP}{FP + TN} = 0.055$$

- TPR (Tasa de Verdaderos Positivos): Con un valor de 0.54, esto indica que el modelo ha identificado correctamente el 54% de los casos positivos en comparación con el total de casos positivos reales. En otras palabras, el modelo logra capturar un porcentaje considerable de los eventos que realmente ocurrieron, lo que es positivo.
- FPR (Tasa de Falsos Positivos): Con un valor de 0.055, esto significa que el modelo ha cometido un error del 5.5% al clasificar incorrectamente casos negativos como positivos en comparación con el total de casos negativos reales. Un FPR bajo es un indicativo positivo, ya que significa que el modelo no está generando un gran número de falsas alarmas al predecir eventos que no ocurrieron.