Juan Camilo Cardona Castaño
Julian David Gil Botero
Santiago Salazar Osorio

UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

ÍNDICE GENERAL

1. Introducción **1**
1.1. Contexto 2
1.2. Planteamiento del Problema 2
1.3. Dataset 2

2. Análisis de variables **3**
2.1. Tipos de variables 3
2.1.1. Variables categóricas 3
2.1.2. Variables numéricas 3

3. Limpieza de datos **4**
3.1. División de Datos 5
3.2. Visualización de datos 5

4. Mejores hiperparámetros de dos algoritmos predictivos **12**
4.1. Algoritmo Predictivo: LogisticRegression 12
4.2. Algoritmo Predictivo: RandomForestClassifier 13
4.3. Matriz de confusión 13
4.4. Curvas de aprendizaje de los algoritmos 14

5. Retos Futuros **15**

6. Conclusiones **16**

7. Bibliografía **17**

1. Introducción

La inteligencia artificial (IA) ha emergido como una herramienta preeminente, desplegada con el propósito de anticipar escenarios y proporcionar soluciones en prácticamente todos los dominios del conocimiento. Su adopción ha experimentado un crecimiento exponencial en los últimos años, y esta tendencia ascendente se

proyecta con confianza hacia el futuro. Dentro del vasto espectro de herramientas de la IA, destacan el machine learning, el deep learning y la ciencia de datos, componentes fundamentales que potencian la capacidad predictiva y resolutiva de dicha tecnología.

En el contexto de un proyecto de inteligencia artificial, la esencia radica en la disponibilidad de una base de datos suficientemente amplia para nutrir y entrenar al programa. Sin embargo, la magnitud de los datos no es el único factor determinante; la calidad de los mismos juega un papel igualmente crucial. Es decir, la obtención de resultados óptimos está intrínsecamente ligada a la especificidad y precisión de los datos utilizados. Posteriormente, se procede a la elección de un modelo o plantilla de modelo, cuya calibración implica la utilización de una porción de los datos previamente definidos como datos de calibración. Este proceso es crucial para ajustar y afinar el modelo de manera que refleje de manera óptima la complejidad inherente de los datos. Finalmente, la validación del método se realiza mediante la aplicación del modelo a otro conjunto de datos, denominado en este contexto como datos de testeo.

A continuación, se presenta el desarrollo integral del trabajo final correspondiente al curso de Introducción a la Inteligencia Artificial. En este informe, se detalla minuciosamente el tratamiento de los datos y el procedimiento metodológico empleado para abordar y resolver el problema planteado, ofreciendo una visión integral del proceso implementado para alcanzar resultados efectivos.

1.1. Contexto

Australia se destaca por su clima árido y desafiante, lo que dificulta el acceso a recursos hídricos. Por esta razón, el gobierno tiene la responsabilidad de gestionar las licencias para el uso del agua y regular el mercado hídrico. Los precios del agua son volátiles, principalmente influenciados por la demanda diaria que fluctúa. El sector agrícola representa el 80 % del uso de los recursos hídricos del país, convirtiéndose en el mayor actor en el mercado del agua. La necesidad de agua por parte de los agricultores está fuertemente ligada a las precipitaciones, ya que los días lluviosos disminuyen la demanda. Esta relación entre la lluvia y la demanda agrícola dificulta la fijación de precios para este recurso debido a la incertidumbre climática.

1.2. Planteamiento del Problema

Desarrollar un modelo predictivo para discernir entre días lluviosos y no lluviosos en una fecha específica del calendario implica considerar diversas variables climáticas, como la velocidad del viento, temperatura, presión, entre otras. La salida de este modelo se presenta como una variable booleana (Sí/No), acompañada de la probabilidad asociada de precipitación. Este enfoque integrado permite anticipar con precisión las condiciones meteorológicas, proporcionando información valiosa para la toma de decisiones y la planificación.

1.3. Dataset

El presente estudio se fundamenta en una base de datos de la plataforma Kaggle. Este conjunto de datos abarca las condiciones climáticas registradas en diversas ciudades de Australia a lo largo de un extenso período de más de 10 años. La amplitud temporal de la información recopilada comprende desde el 1 de diciembre de 2008 hasta el 25 de junio de 2017.

La robustez de este dataset se refleja en su extensión, que abarca más de 145 mil días de observación. Cada día dentro de este intervalo temporal se caracteriza por un conjunto detallado de 24 atributos meteorológicos,

proporcionando una visión holística de las condiciones atmosféricas. Entre estos atributos se incluyen variables cruciales como humedad, presión atmosférica, temperaturas máximas y mínimas, así como la dirección y velocidad del viento, entre otros parámetros relevantes.

Un elemento distintivo de este conjunto de datos es la inclusión de una columna adicional que incorpora una respuesta booleana (Sí/No). Esta columna refleja de manera precisa si se registró lluvia al día siguiente, considerando afirmativa la opción cuando la precipitación excede 1 mm. Esta característica proporciona un valor adicional al conjunto de datos al permitir la exploración de patrones climáticos relacionados con la ocurrencia de lluvia, lo que contribuye significativamente al análisis y comprensión de los fenómenos meteorológicos en el contexto específico de las localidades australianas consideradas.

Para acceder al dataset seleccionado: *CLICK AQUÍ*

2. Análisis de variables

2.1. Tipos de variables

Al abordar la clasificación de variables, es crucial distinguir entre las categóricas y las numéricas en el conjunto de datos. La presencia de ambas en el dataset destaca la necesidad de una identificación y procesamiento preciso, preparando así las variables para su integración efectiva en los modelos analíticos.

2.1.1. Variables categóricas

En el dataset pueden identificar 6 variables categóricas:
Categoricas = ['Location' , 'WindGustDir' , 'WindDir9am' , 'WindDir3pm' , 'RainToday' , 'RainTomorrow']

Tabla 1. *Cantidad de datos faltantes para cada variable categórica.*

WindGustDir	10326
WindDir9am	10566
WindDir3pm	4228
RainToday	3261
RainTomorrow	3267

2.1.2. Variables numéricas

En cuanto a las variables numéricas, tenemos 16:

Numéricas = ['MinTemp' , 'MaxTemp' , 'Rainfall' , 'Evaporation' , ' Sunshine' , 'WindGustSpeed' , 'WindSpeed9am' , 'WindSpeed3pm' , 'Humidity9am' , 'Humidity3pm' , 'Pressure9am' , 'Pressure3pm' , 'Cloud9am' , ' 'Cloud3pm' , 'Temp9am' , 'Temp3pm']

Tabla 2. *Cantidad de datos faltantes para cada variable numerica.*

MinTemp	1485
MaxTemp	1261

Rainfall	3261
Evaporation	62790
Sunshine	69835
WindGustSpeed	10263
WindSpeed9am	1767
WindSpeed3pm	3062
Humidity9am	2654
Humidity3pm	4507
Pressure9am	15065
Pressure3pm	15028
Cloud9am	55888
Cloud3pm	59358
Temp9am	1767
Temp3pm	3609

2.1.2.1. Problemas en variables categóricas y numéricas

Al analizar las variables categóricas y numéricas mencionadas previamente, se observa que existen datos faltantes en algunas de ellas. Al evaluar los valores nulos, se identificó la presencia de estos vacíos en el conjunto de datos. Ante esta situación, se decidió abordar estos datos faltantes eliminándolos del dataset. Esta acción se tomó en consideración debido a la cantidad suficiente de filas disponibles para entrenar el modelo, priorizando la calidad de los datos sobre la cantidad. Las variables categóricas fueron modificadas con el propósito de mejorar su comprensión en el análisis.

En relación a la variable 'Location', cada ubicación se convirtió en una columna individual, lo que facilita la identificación de los datos según su ubicación geográfica. Respecto a las direcciones del viento, se optó por asociar a cada dirección un número específico. Este enfoque permitió una representación numérica de las distintas orientaciones del viento (por ejemplo, Norte: 1, Sur: 2, Este: 3, Oeste: 4, y así sucesivamente), proporcionando una comprensión más clara de las condiciones del viento sin incrementar excesivamente la cantidad de columnas en el conjunto de datos.

Además de estas transformaciones, las variables 'RainToday' y 'RainTomorrow' fueron tratadas asignándoles el valor de 1 si indican lluvia y 0 si no, lo cual simplifica su interpretación en el análisis meteorológico.

3. Limpieza de datos

La limpieza de datos es una práctica esencial en el análisis de datasets, donde se busca identificar y corregir valores nulos o inconsistentes. En este contexto, es crucial distinguir entre variables categóricas y numéricas para aplicar las estrategias de limpieza apropiadas

En el preprocesamiento de datos, se llevó a cabo la eliminación de aquellas variables que presentaban una falta de datos superior al 10%. Esta acción se tomó con el objetivo de mantener la integridad y confiabilidad del conjunto de datos, dado que la ausencia significativa de información en estas variables podría afectar negativamente la capacidad predictiva del modelo.

Adicionalmente, se procedió a la exclusión de todos los registros asociados a la variable "RainTomorrow". Esta variable se identificó como la variable respuesta, es decir, aquella que deseamos predecir. La eliminación de los datos correspondientes a esta variable se llevó a cabo para evitar cualquier sesgo en el entrenamiento del modelo, ya que incluir información sobre el evento que estamos tratando de prever podría distorsionar los resultados.

No obstante, se observó la persistencia de variables con datos faltantes después de estas operaciones. En aras de asegurar la calidad de los datos utilizados para el entrenamiento del modelo, se tomó la decisión de eliminar también estas variables restantes con valores ausentes. Este enfoque se adoptó con el propósito de garantizar que el modelo sea entrenado únicamente con datos completos y fiables, maximizando así su capacidad para realizar predicciones precisas en nuevos conjuntos de datos.

3.1. División de Datos

La división de datos es un paso crucial en el proceso de entrenamiento de modelos, y su impacto en el rendimiento final no puede subestimarse. Un enfoque comúnmente adoptado es dividir el conjunto de datos en porciones estratégicas, generalmente entre el 10% y el 30%, para su uso en evaluación y validación.

Esta metodología busca encontrar un equilibrio entre la cantidad de datos utilizados para entrenar el modelo y la cantidad reservada para evaluar su rendimiento. Inicialmente, se realiza una división donde un porcentaje significativo se destina al entrenamiento del modelo, permitiéndole aprender patrones y características fundamentales de los datos. El porcentaje restante se reserva para evaluar el rendimiento del modelo en un entorno no visto durante el entrenamiento.

Después de las iteraciones realizadas se determinó que el 25% de los datos es el porcentaje óptimo para el conjunto de testeo y el 75% restante para la sección de entrenamiento, asegurando así resultados confiables.

3.2. Visualización de datos

Para comprender mejor el panorama, se emplearon distintos tipos de gráficos.

Con la *Figura 1.* buscamos identificar la ciudad con mayores precipitaciones ('Rainfall' en mm). A simple vista, se evidencia que esta situación se concentra entre Brisbane y Cairns. Por ende, se podría inferir que son las localidades con mayor probabilidad de experimentar lluvias intensas.

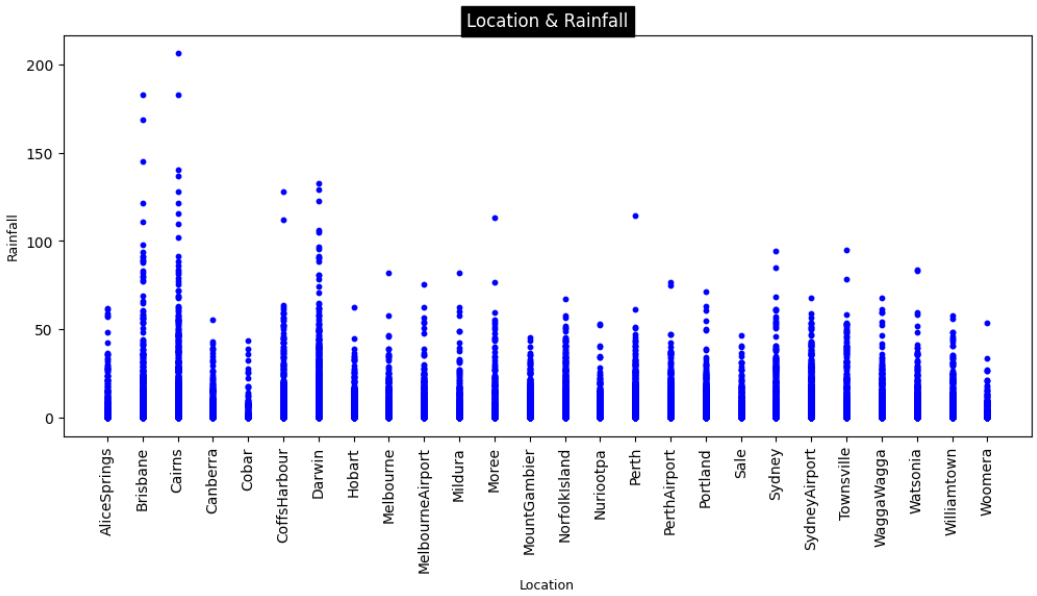
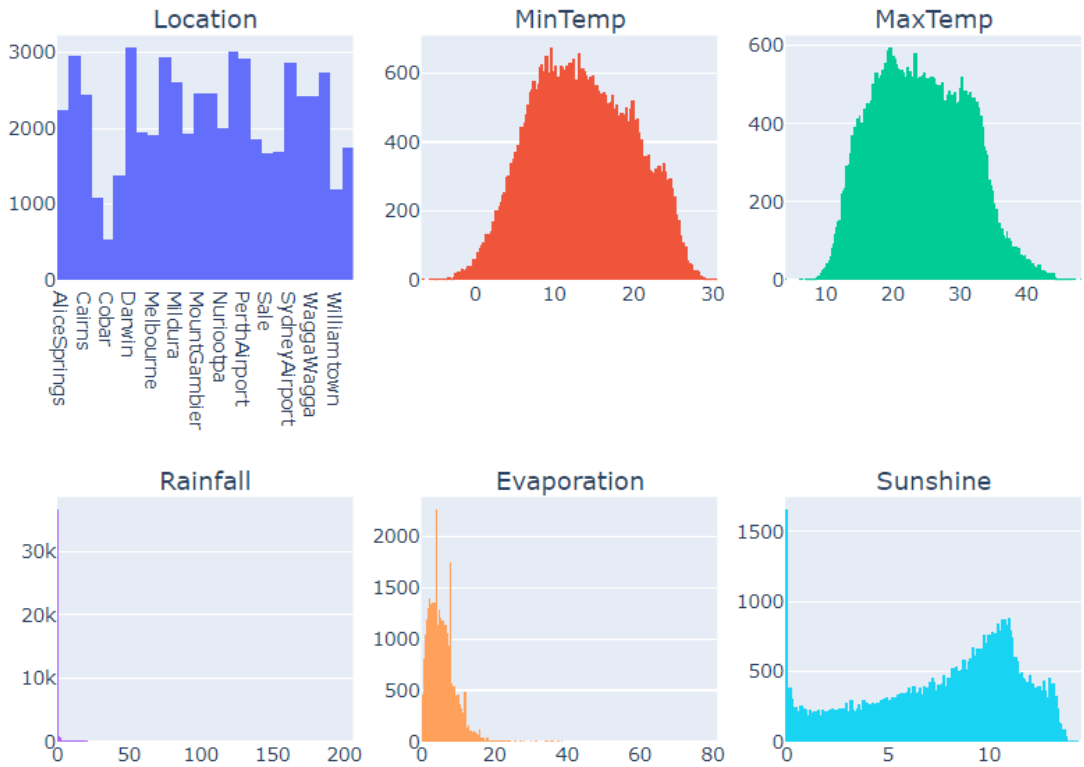
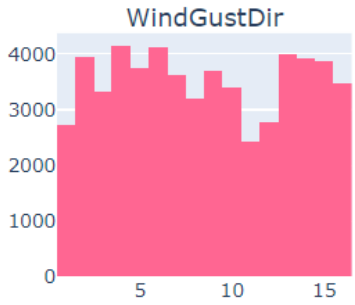
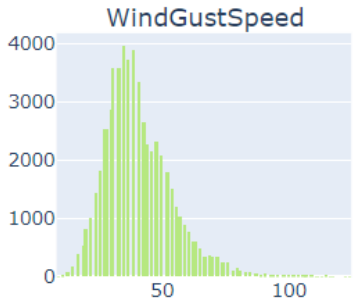


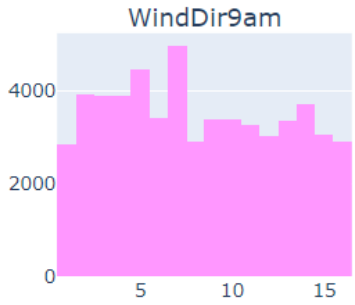
Figura 1. *Precipitación en las diferentes ciudades.*

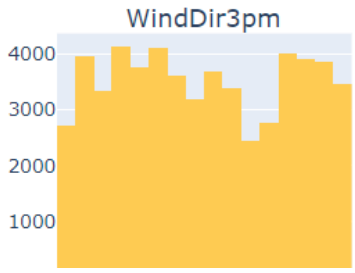
El *Figura 2*. corresponde a un histograma que muestra la distribución de todas las variables. Destaca que, en el caso de 'RainTomorrow', alrededor del 22% de los 56,420 días analizados presentaron lluvia, mientras que el resto no. En consecuencia, se puede concluir que, en general, es más probable que no llueva.

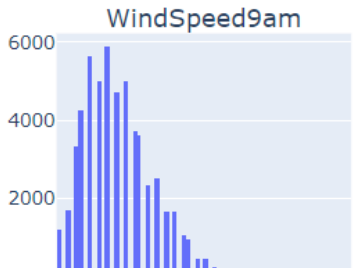


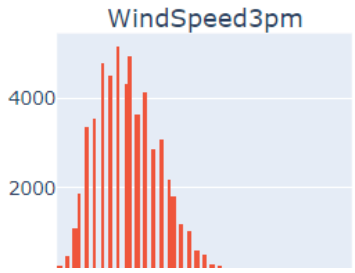

WindGustDir
4000
3000
2000
1000
0
5
10
15

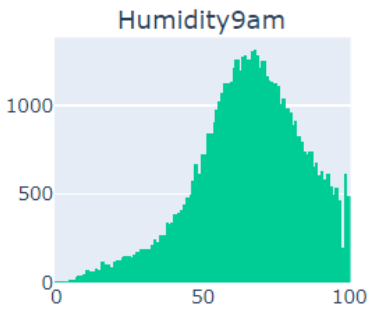

WindGustSpeed
4000
3000
2000
1000
0
50
100

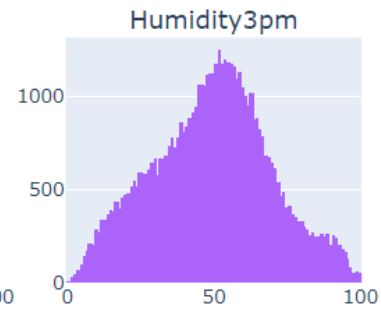

WindDir9am
4000
2000
0
5
10
15

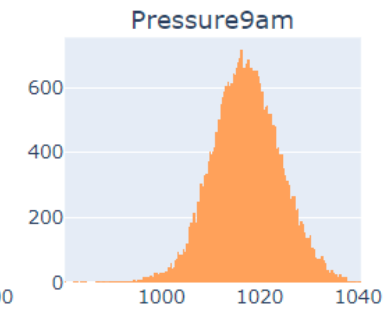

WindDir3pm
4000
3000
2000
1000

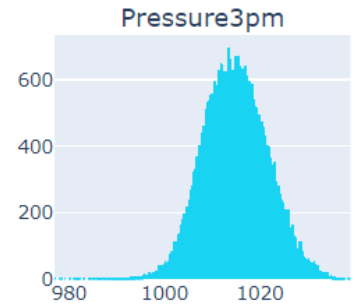

WindSpeed9am
6000
4000
2000

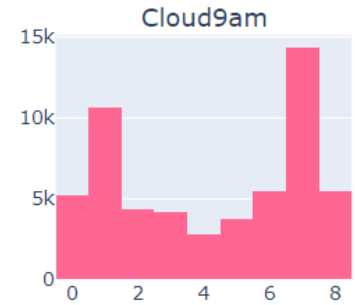

WindSpeed3pm
4000
2000

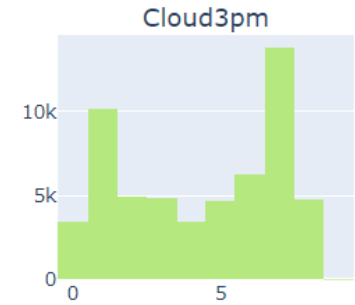

Humidity9am
1000
500
0
0
50
100


Humidity3pm
1000
500
0
0
50
100


Pressure9am
600
400
200
0
1000
1020
1040


Pressure3pm
600
400
200
0
980
1000
1020


Cloud9am
15k
10k
5k
0
0
2
4
6
8


Cloud3pm
10k
5k
0
0
5

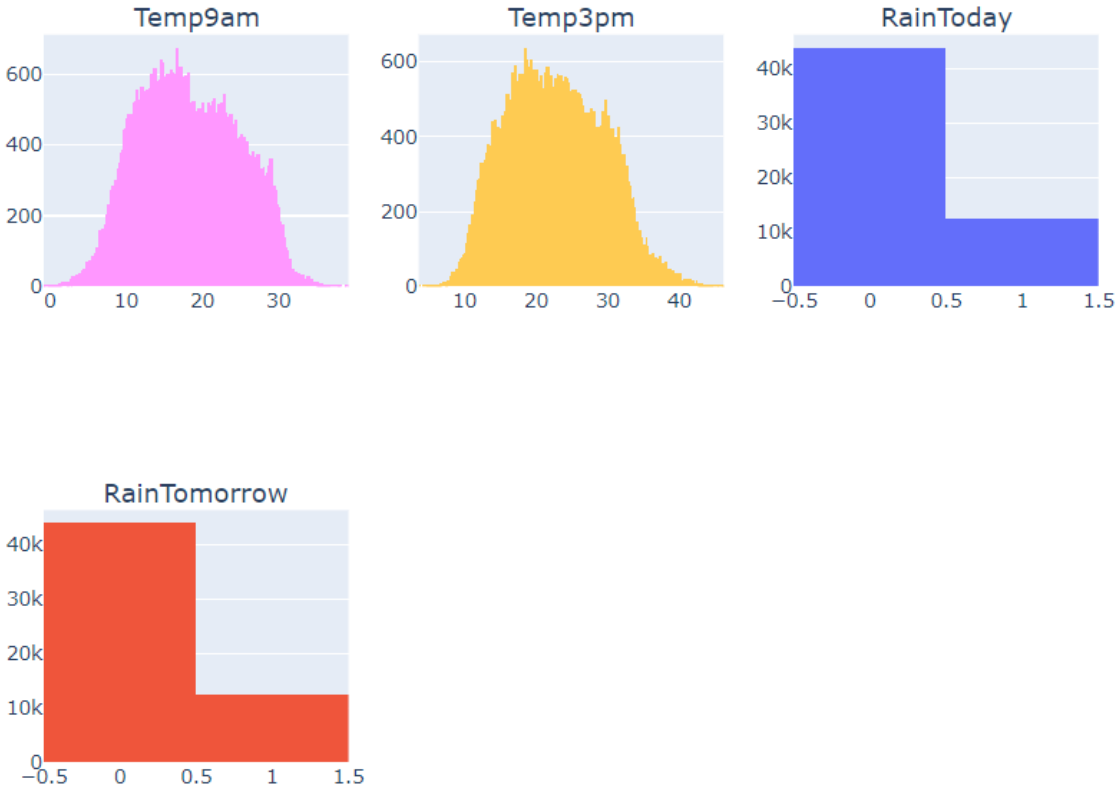


Figura 2. *Histograma para cada una de las variables.*

En la *Figura 3.* se emplea un pairplot para explorar relaciones entre variables y comprender mejor el problema. Sin embargo, es en el *Figura 4.* el mapa de calor o matriz de correlaciones, donde se resaltan las conexiones más significativas. Este análisis indica que la variable 'Sunshine' guarda una relación inversa con 'RainTomorrow'. Además, se observa que, en general, a mayor humedad, aumenta la probabilidad de precipitaciones. También se destaca una relación directa entre la lluvia de un día y la probabilidad de lluvia al día siguiente.

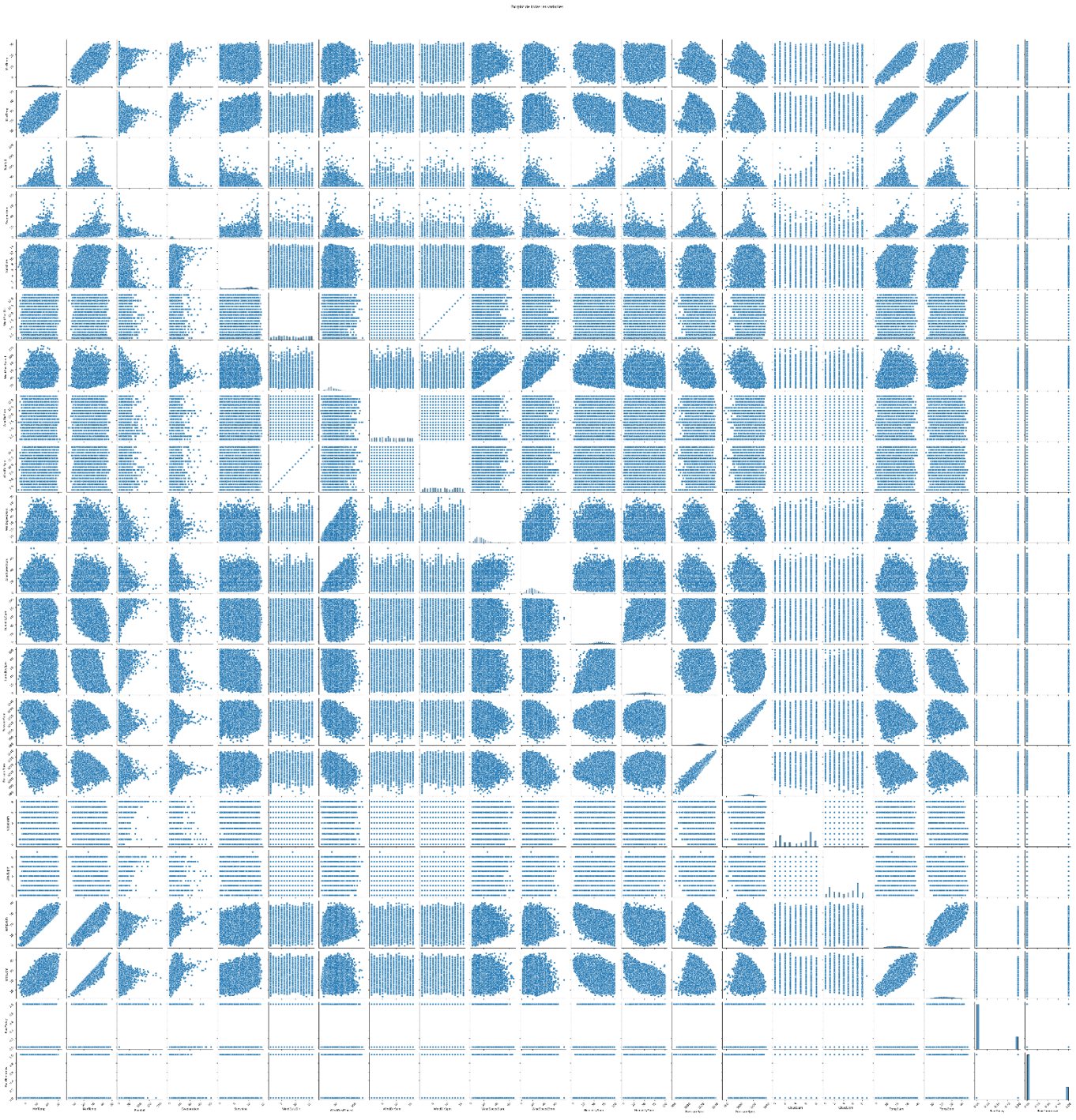


Figura 3. *Correlación de las variables del dataset.*

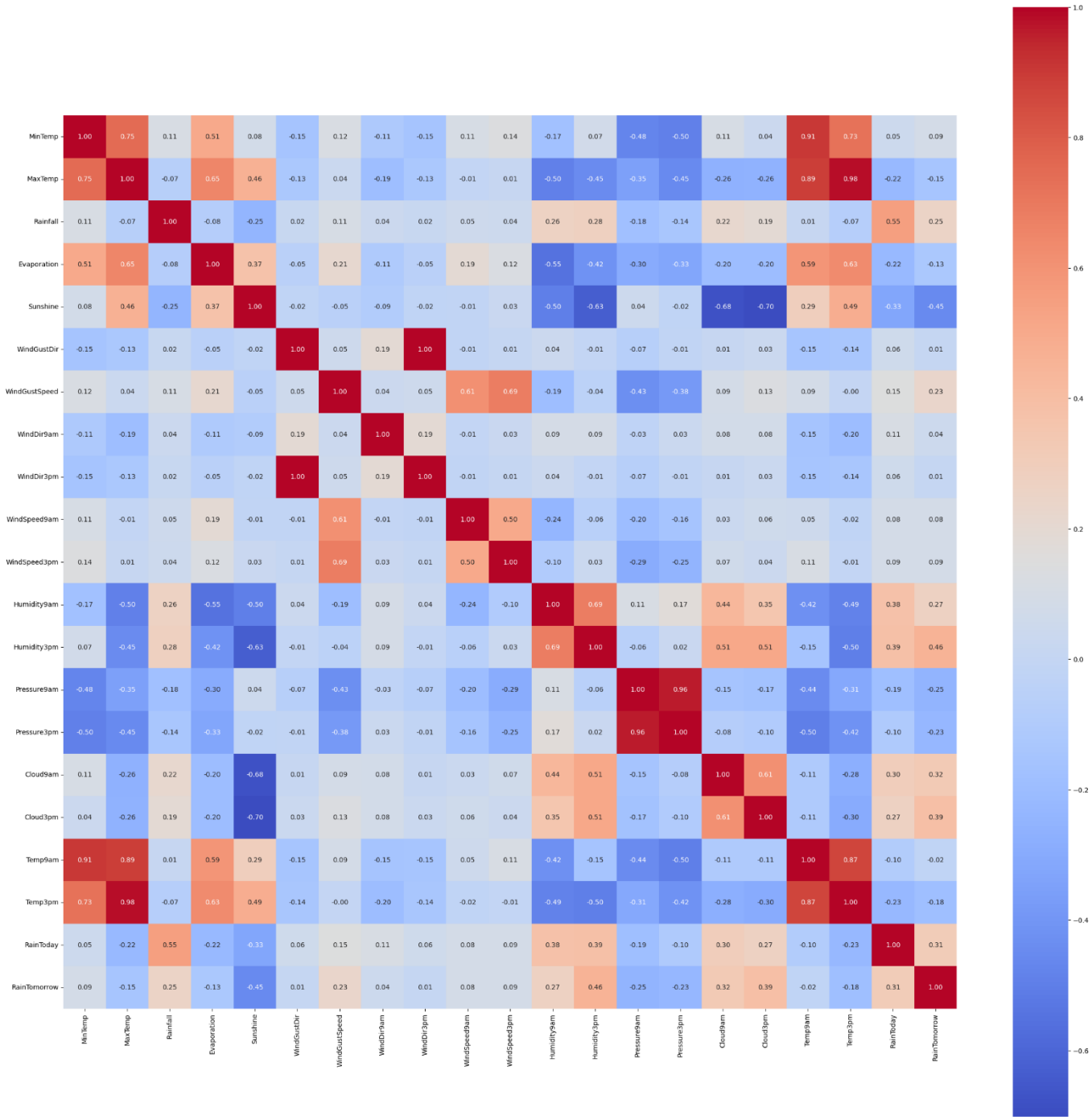


Figura 4. *Correlación y mapa de calor de las variables.*

En la *Figura 5*. que corresponde a los gráficos de humedad a las 9 am y 3 pm, se percibe que una humedad superior al 60% a las 9 am indicaría una mayor probabilidad de lluvia ('Rainfall'). De manera similar, en el caso de la humedad a las 3 pm, se encuentra una relación directa, aunque menos pronunciada.

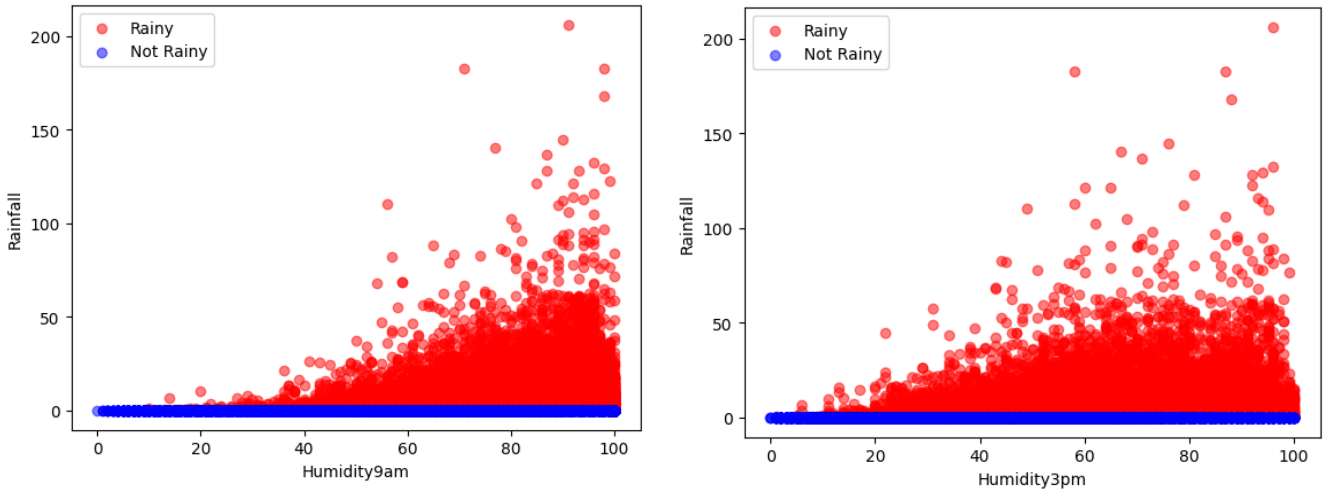


Figura 5. *Humedad a distintas horas.*

Además, se generó un gráfico para determinar los meses del año con un mayor promedio de lluvia a lo largo de los datos analizados *Figura 6.* Visualmente, se destaca que estos meses son junio (mes 6) y julio (mes 7).

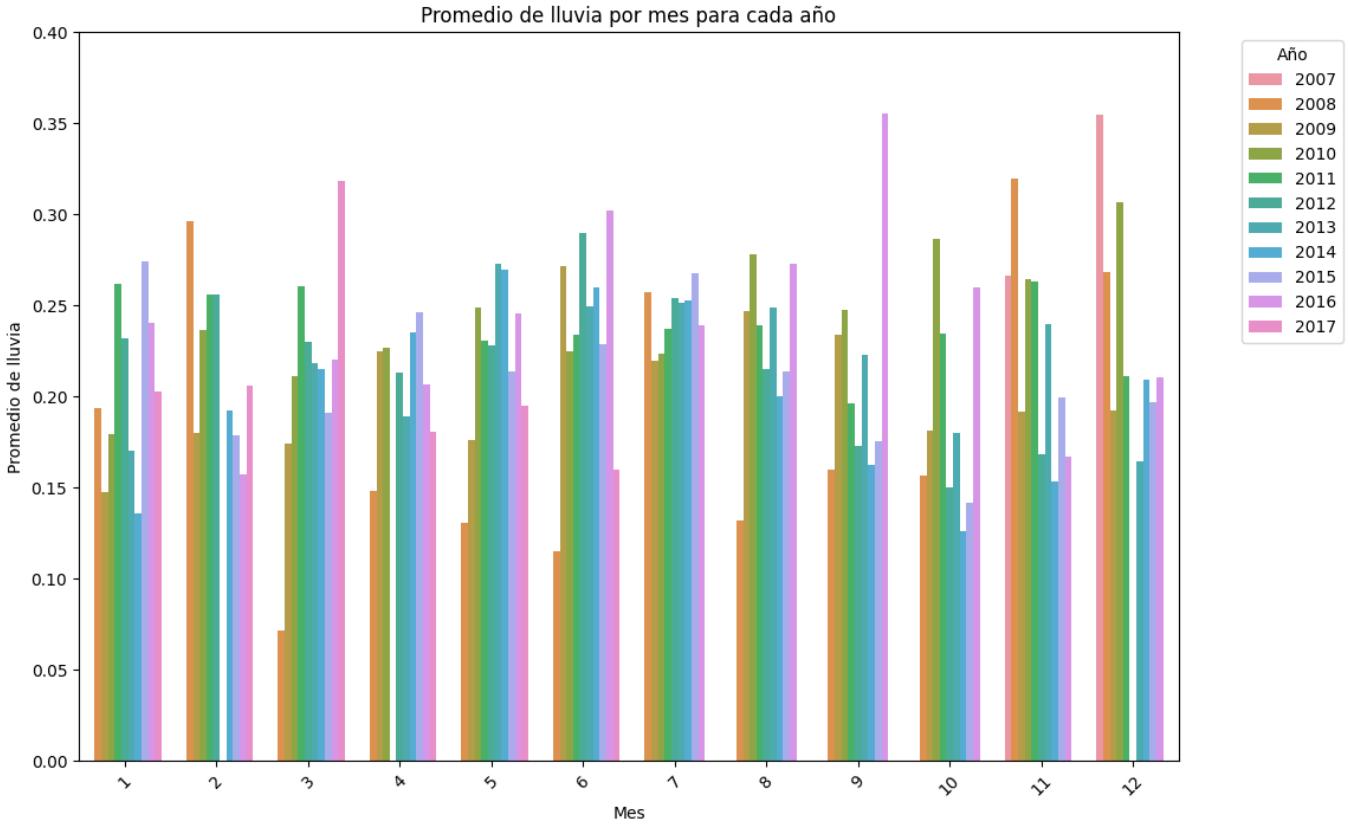


Figura 6. *Promedio de lluvia por mes para cada año.*

4. Mejores hiperparámetros de dos algoritmos predictivos

4.1. Algoritmo Predictivo: LogisticRegression

El LogisticRegression es un algoritmo simple y ampliamente utilizado en la clasificación supervisada. Su aplicación se centra en predecir la probabilidad de una variable categórica dependiente [2]. Utilizando la librería sklearn, se entrena y aplica este modelo. Se realizó una búsqueda exhaustiva de hiperparámetros para optimizar los resultados del algoritmo, siguiendo el esquema establecido en el Notebook para encontrar la mejor configuración.

Para este algoritmo, se exploraron y combinaron diversos conjuntos de parámetros con el fin de encontrar la configuración óptima. La meta inicial consistía en alcanzar una precisión superior al 80% en la predicción de eventos de lluvia en Australia, manteniendo el FPR por debajo del 15%. No obstante, se observó en la matriz de confusión que el FPR se situó aproximadamente en un 5%, cumpliendo con la meta establecida. Aunque este valor es aceptable, desde una perspectiva económica, se busca reducir aún más este parámetro para minimizar cualquier margen de error y optimizar la predicción.

Es evidente la necesidad de una mejora continua en la metodología empleada para perfeccionar la precisión del modelo, lo cual no solo es crucial para cumplir con los estándares predefinidos, sino también para garantizar un impacto económico positivo al mitigar posibles consecuencias derivadas de predicciones erróneas.

La curva ROC es una curva de evaluación que indica qué tan bien se ajusta un modelo en términos de la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). Si la pendiente de la curva es de 1, significa que el modelo es completamente aleatorio. Además, a medida que aumenta la TPR, también aumenta la FPR, por lo que encontrar un equilibrio entre ambas es importante. En este caso, una FPR del 5% y una TPR del 55% aproximadamente se considera ganador por la baja cantidad de FPR que puede afectar de manera negativa al mercado del agua. Además, el área bajo la curva (AUC =0.89) es una medida de qué tan bien se desempeña el modelo en general, siendo el valor máximo 1.

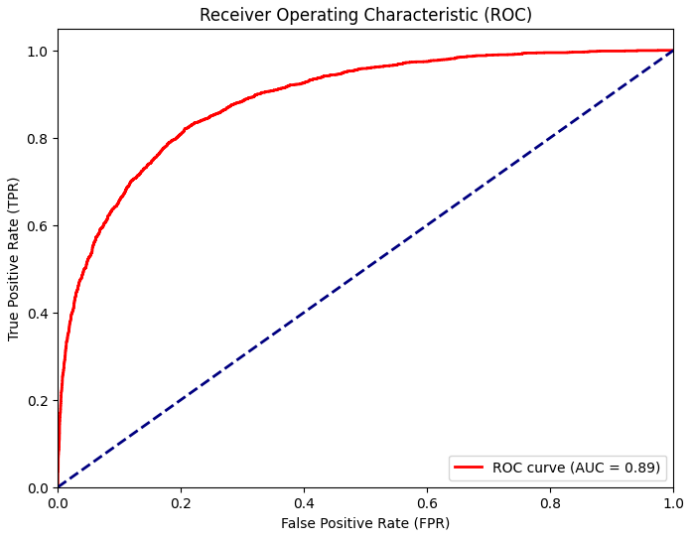


Figura 7. *Curva ROC modelo predictivo LogisticRegression.*

4.2. Algoritmo Predictivo: RandomForestClassifier

Los árboles de decisión dividen las observaciones según atributos para obtener resultados deseados. Diferentes árboles con configuraciones distintas generan predicciones variadas, y la respuesta final se obtiene combinando las respuestas de todos los árboles [4]. Este método puede emplearse tanto para clasificación como para regresión [5].

En el contexto específico de este trabajo, el RandomForestClassifier se usó para la clasificación binaria (llueve/no llueve). Se exploraron diferentes conjuntos de parámetros para encontrar la configuración óptima del modelo, manteniendo como objetivo la mejora continua de la precisión y la fiabilidad de las predicciones.

Esta curva ROC también es ganadora siendo su área bajo 0.88 muestra una menor taza de FPR 4% sin embargo se sacrifica mucho el TPR siendo 50% pero se considera que se ajusta bien el modelo y hay un buen balance según la necesidad del problema.

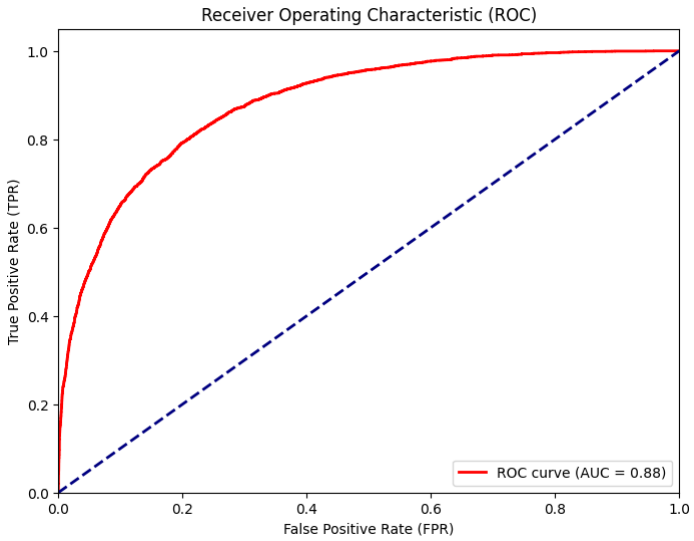


Figura 8. *Curva ROC modelo predictivo RandomForestClassifier.*

4.3. Matriz de confusión

Como se postuló en la primera entrega de este trabajo, se requería una precisión mayor al 75% en los resultados para la lluvia en Australia. Además, se esperaba que el False Positive Rate (FPR) fuera menor al 15%, esto con la finalidad de evitar infravaloraciones en el precio del agua como bien se expone en la primera entrega y en el contexto de este trabajo. Para analizar estos valores, se usará la matriz de confusión expuesta en la *Figura 9.* Así bien, se tiene FPR de aproximadamente el 5% en el modelo predictivo de LogisticRegression y 3% para el modelo predictivo Random Forest logrando así un resultado cercano al requerido, sin embargo, sería mucho mejor, en términos económicos, disminuir lo que más se pueda este parámetro de precisión.

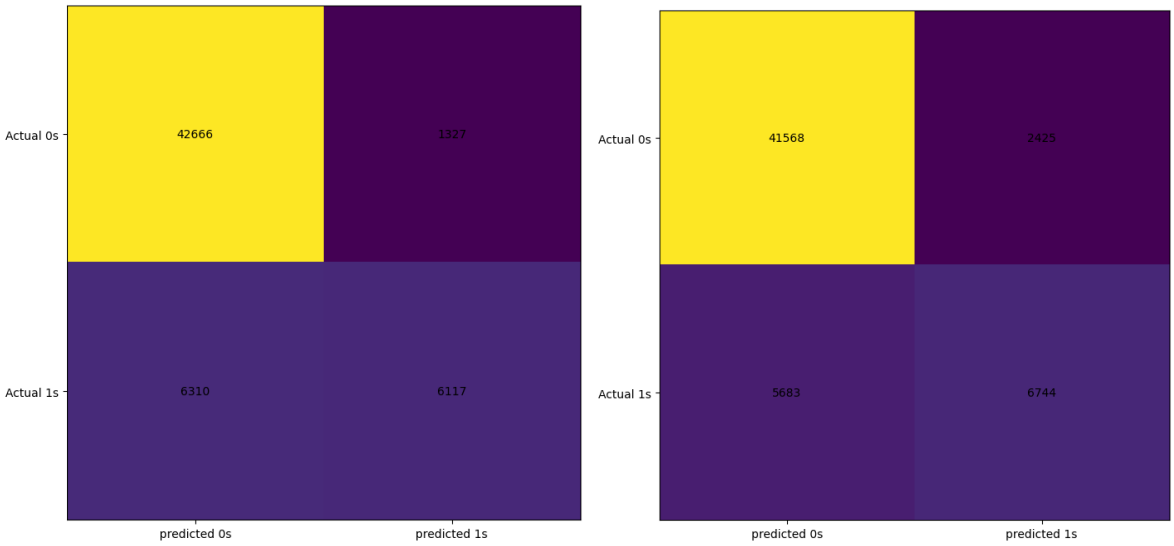


Figura 9. *Matriz de confusión para las predicciones de las lluvias en Australia. a) Random Forest , b) LogisticRegression*

4.4. Curvas de aprendizaje de los algoritmos

Al analizar las curvas de aprendizaje, se observa que para la LogisticRegression, hay una desviación estándar bastante alta, indicando una falta de convergencia del modelo. Esta falta de convergencia puede atribuirse a dos posibles causas principales: una posible carencia de complejidad en el modelo o una insuficiencia en las columnas descriptivas utilizadas.

Por otro lado, al evaluar el RandomForestClassifier, se destaca que, aunque tiende a estabilizarse después de alcanzar los 15,000 datos, muestra una convergencia limitada. Se podría experimentar con una max_depth mayor para este modelo, ya que esto podría ayudar a mejorar su capacidad predictiva.

A pesar de estos hallazgos, ambos modelos muestran un puntaje aceptable, lo que indica que, a pesar de las dificultades para converger completamente, aún pueden ofrecer resultados adecuados en la predicción

Tabla 3. *Resumen.*

	Models	Accuracy	TestSize
0	logreg_Newton-cg	0.856505	0.25
1	logreg_Sag	0.854094	0.25
2	logreg_Liblinear	0.855075	0.30
3	logreg_Saga	0.853775	0.20
4	RF	0.854484	0.25

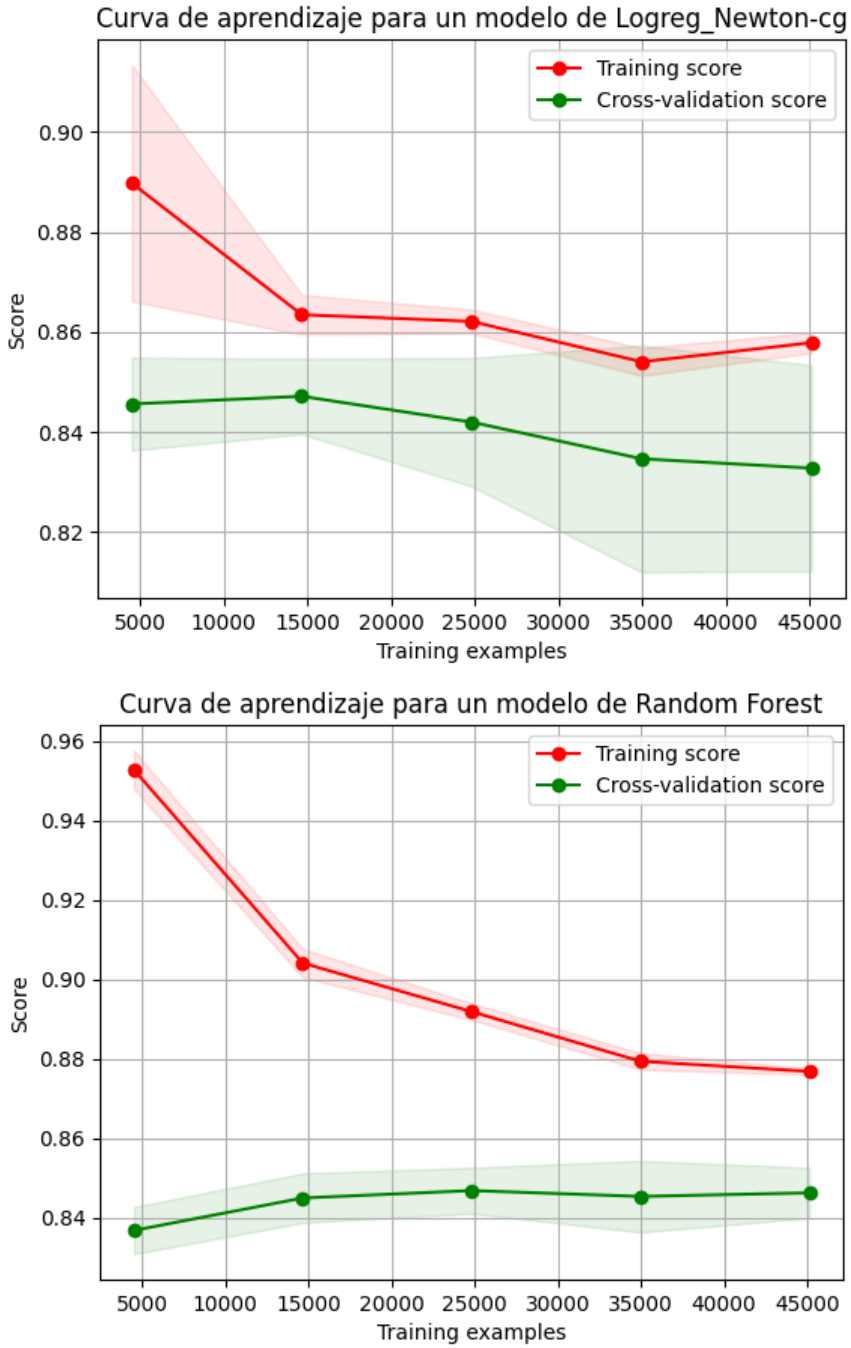


Figura 10. *Curva de aprendizaje para los modelos predictivos LogisticRegression y RandomForestClassifier.*

5. Retos Futuros

En pro de mejorar los resultados, se plantean varios retos adicionales:

- **Explorar Modelos más Complejos:** Considerar modelos como SVM o aumentar la profundidad del

RandomForest para evaluar si se obtienen mejoras significativas en las métricas.

- **Manejo de Datos Faltantes:** Emplear estrategias de imputación para datos faltantes, ya sea a través de la media, moda o con valores del día anterior, agrupados por la ciudad. Posteriormente, volver a entrenar los modelos y comparar su desempeño.
- **División Adicional de Datos:** Tomar un 10% adicional de los datos completos (56,420 filas) para usarlos exclusivamente como conjunto de prueba en los modelos, tanto en los casos donde se rellenen los datos como en los que no.

6. Conclusiones

- Cumplimiento de Métricas Iniciales: Todos los modelos evaluados alcanzaron un accuracy superior al 80% Tabla 3. Además, se logró mantener un bajo FPR del 5% y un TPR del 50% para la clasificación de días de lluvia.
- Aunque el modelo de LogisticRegression con el solver newton-cg mostró el mejor rendimiento entre los modelos probados, es crucial observar la curva de aprendizaje. Esta curva señala una disminución en el rendimiento del modelo de regresión logística a partir de aproximadamente 25,000 datos. Esta disminución podría sugerir ciertas limitaciones del modelo conforme aumenta el volumen de datos.
- Considerando este comportamiento, se plantea una alternativa interesante: el RandomForestClassifier podría ser una elección viable para este problema de clasificación. Este modelo no solo cumple con las condiciones iniciales requeridas (80% de accuracy y un FPR del 5%), sino que además su curva de aprendizaje indica la posibilidad de mejorar los resultados. Aunque se limitó la profundidad del árbol a 10 debido a restricciones computacionales, se vislumbra la oportunidad de obtener mejoras significativas mediante la exploración de árboles más profundos.
- Dicho esto, la decisión final dependerá de ponderar la capacidad de generalización de cada modelo y la factibilidad computacional en función de las necesidades del proyecto.
- Área Bajo la Curva (AUC): El mejor desempeño en AUC corresponde nuevamente al modelo Newton, pero esto se debe a que no se exploró la complejidad del RandomForest.
- Entrenamiento con el 75% de los Datos: Se logró entrenar satisfactoriamente los modelos utilizando el 75% de los datos disponibles. La mejor matriz de confusión se obtuvo con el modelo Newton un TPR de aproximadamente 55% y FPR de 5%.

7. Bibliografía

[1] J. M. Parra, "Estadística descriptiva e inferencial i," *Recuperado de: http://www. academia. edu/download/35987432/ESTADISTICA_DESCRIPTIVA_E_INFERENCIAL. pdf*, 1995.

[2] Iartificial.net, "¿cómo usar regresión logística en python?." https://www.iartificial.net/como-usar-regresion-logistica-en-python/. [Online; accedido 03-Octubre-2021].

[3] towards data science, "Building a logistic regression in python, step by step." https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8. [Online; accedido 03-Octubre-2021].

[4] Cienciadedatos.net, "Árboles de decisión, random forest, gradient boosting y c5.0." https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting. [Online; accedido 03-Octubre-2021].

[5] freecodecamp.org, "Tutorial para un clasificador basado en bosques aleatorios: cómo utilizar algoritmos basados en árboles para el aprendizaje automático." https://www.freecodecamp.org/espanol/news/ random-forest-classifier-tutorial-how-to-use-tree-based-algorithms-for-machine-learning/. [Online; accedido 03-Octubre-2021].