Data Mining [IES-411]

Clasificación y pronóstico de precipitaciones en Rapa Nui

Usando XGBoost y técnicas de minería de datos sobre registros meteorológicos.

Rossemari Gajardo González
Angel Llanos Herrera

Introducción

Contexto del problema

- **Aislamiento geográfico**: Rapa Nui no cuenta con ríos ni acuíferos superficiales.
- **Dependencia de la lluvia**: La captación pluvial es la principal fuente de agua para consumo y agricultura.
- **Irregularidad pluviométrica**: Las precipitaciones son de difícil pronóstico y de alta variabilidad temporal.
- **Desafío de gestión**: Esta incertidumbre obliga a implementar una planificación hídrica más eficiente, basada en pronósticos confiables.

Planteamiento del problema

- La alta variabilidad e irregularidad de las precipitaciones en Rapa Nui impide planificar el riego y actividades agrícolas de forma confiable.
- Este escenario conduce a un uso ineficiente del agua de lluvia, principal recurso hídrico de la isla.
- La falta de capacidades predictivas amenaza la seguridad alimentaria y la sostenibilidad agrícola local.

Objetivos y justificación del estudio

- El enfoque del estudio está en prever la ocurrencia de lluvia hora a hora a partir de datos de la estación meteorológica Mataveri. Mediante modelos de clasificación y regresión.
- Igualmente, es de interés comprender el fenómeno climático en Rapa Nui mediante análisis descriptivo.
- Buscando generar una herramienta útil mediante técnicas de minería de datos para anticipar horas con y sin lluvia y apoyar así la toma de decisiones locales (riego, siembre y cosecha)

Repositorio en GitHub

Metodología

Fuente de datos

Los datos meteorológicos fueron obtenidos desde la Dirección Meteorológica de Chile, específicamente de la estación Mataveri ubicada en Rapa Nui, con código nacional 270001. Se trata de registros en minutos recopilados de forma continua y descargables desde su plataforma pública.

Variable	Descripción	Variable	Descripción
codigoNacional	Código de la estación meteorológica	ffInst	Velocidad del viento (m/s)
momento	Fecha y hora del registro (UTC)	isSkyClear	Cielo despejado (1) o nublado (0)
rr1Hora	Precipitación en una hora (mm)	mes	Mes del año
ts	Temperatura del aire seco (°C)	estacion	Estación del año correspondiente
td	Temperatura del punto de rocío (°C)	ddInst	Dirección del viento (°)
hr	Humedad relativa del aire (%)	qff	Presión al nivel del mar (hPa)

Preprocesamiento

Se identificaron valores faltantes en variables clave: agua caída (12 casos), presión (9) y cielo visible (12). Para abordarlos, se aplicó imputación por el método del vecino más cercano (KNN Imputer), que estima los valores ausentes a partir de observaciones cercanas en un espacio multidimensional.

Los espacios utilizados para definir similitud se construyeron considerando las variables más correlacionadas con la variable a imputar, con el fin de lograr mayor precisión en la estimación:

- Agua caída: Según punto de rocío, humedad y presión.
- Presión: Según punto de rocío.
- Cielo visible: Según agua caída, humedad y punto de rocío.

Preprocesamiento

Para determinar el valor óptimo de k, se utilizó el Error Cuadrático Medio (RMSE) sobre un 20% de los datos completos, seleccionando aquel k que minimizó el error para cada variable.

Los k seleccionados fueron:

- Agua caída: $k = 15$
- Cielo visible: $k = 10$
- Presión: $k = 15$

Finalmente, se normalizaron las variables numéricas con MinMaxScaler y se transformaron las variables categóricas (mes, periodo del día y cielo visible) en dummies para su uso en el modelo.

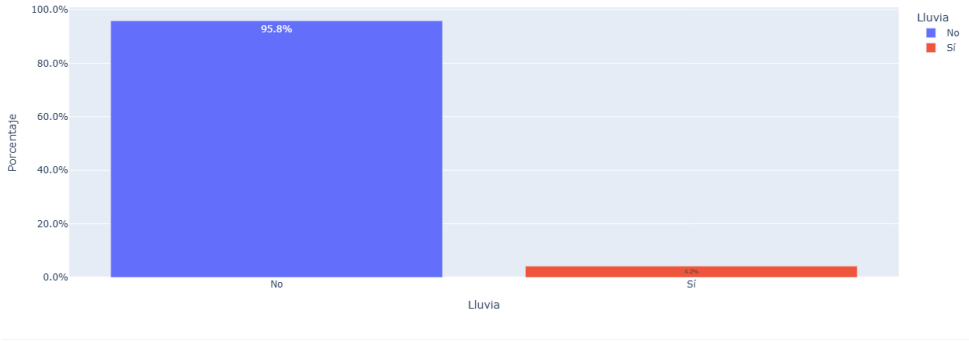
Preprocesamiento

Se crea una nueva variable a partir de agua caída (mm/h) en una variable binaria lluvia/no lluvia para un modelo de clasificación.

- Umbrales de referencia:
  - WMO clasifica llovizna por tamaño de gota (< 0.5 mm) sin umbral de intensidad fijo.
  - Gultepe (2008) y AMS: llovizna hasta ≈ 0.3 mm/h.
- Criterio utilizado:
  - Umbral de 0.4 mm/h para diferenciar lluvia real de llovizna en Rapa Nui (alta humedad y microclima costero).

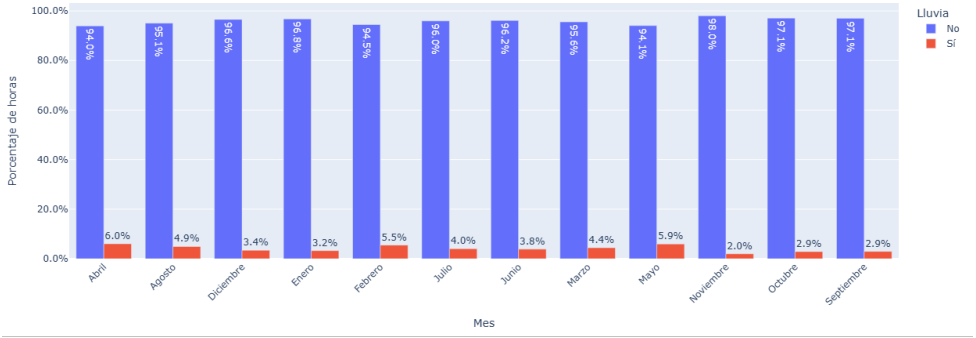
Evita falsas alarmas por lloviznas insignificantes y mejora la calidad de la clasificación binaria en XGBoost.

Análisis exploratorio: Distribución de lluvia



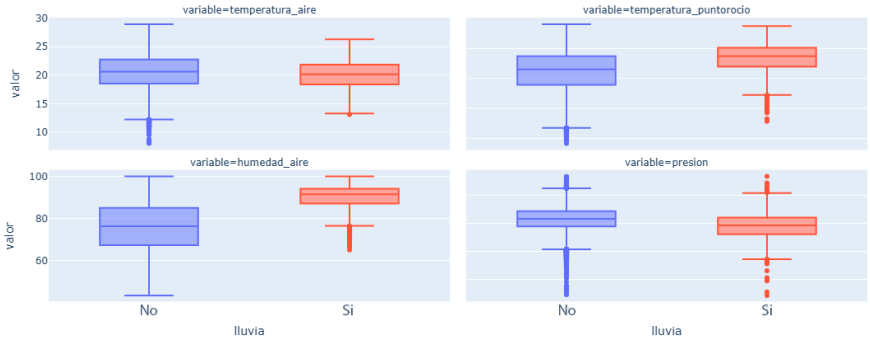
- La mayoría de los registros presentan horas sin lluvia.

Análisis exploratorio: Distribución de lluvia por meses



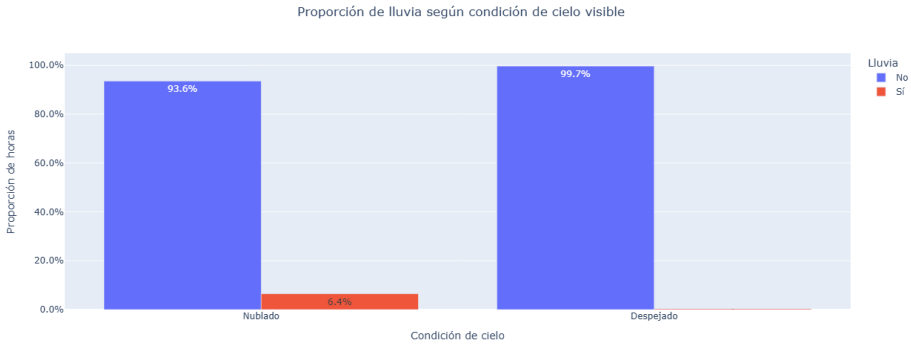
- Meses con más lluvia: Abril (6.0%), mayo (5.9%), febrero (5.5%) y agosto (4.9%).

Análisis exploratorio: Boxplot por lluvia



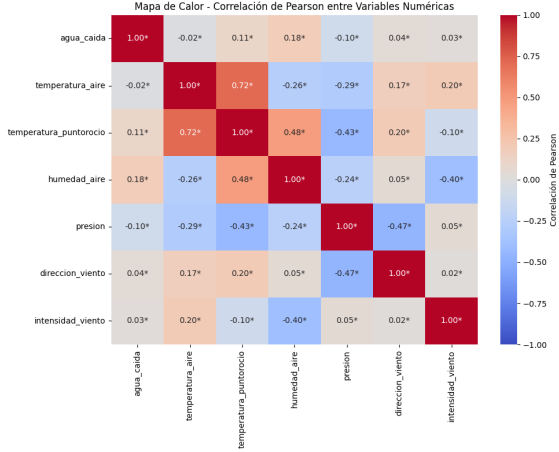
- Durante la lluvia se registra mayor humedad y temperatura del punto de rocío, junto con menor presión y temperaturas del aire.

Análisis exploratorio: Distribución de lluvia por condición del cielo



- Durante horas con condición nublada, el 6.4% de las horas llueve.

Análisis exploratorio: Correlaciones



Análisis exploratorio: Resultados

- La lluvia en Rapa Nui es un fenómeno poco frecuente, pero de alta intensidad, por lo que se emplearon métricas robustas para abordar el desbalance de clases en el modelo.
- Las variables de viento (dirección e intensidad) fueron eliminadas por su baja correlación con la lluvia y su comportamiento disperso, lo que las hacía poco útiles para la predicción.
- Las variables a utilizar para los modelos son: Temperatura del punto de rocío, humedad relativa del aire, presión, condición del cielo, periodo del día y mes.

Modelo XGBoost: Descripción esencial

- Historia:

  En 2001 se propuso inicialmente por Friedman una técnica de aprendizaje supervisado basada en "Gradient Boosting"

  Luego, en 2016 esta técnica fue optimizada por Chen y Guestrin y llamada formalmente como "Extreme Gradient Boosting" (XGBoost)

- Mecanismo básico:

  Secuencia de árboles que corrigen los errores del conjunto previo vía optimización de gradientes.

  Penalización L1/L2 para limitar la complejidad y evitar sobreajuste.

  Captura no linealidades e interacciones (ideal para datos meteorológicos).

Modelo XGBoost

Cada árbol individual f_k, entrega una parte de la predicción del modelo final.

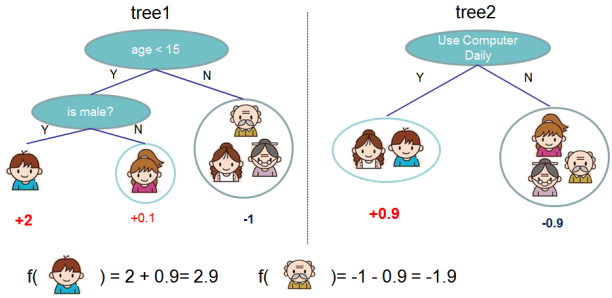


Figure 1: Esquema de XGBoost (adaptado de Chen & Guestrin, 2016).

Se repite hasta un número predefinido o hasta que la mejora sea mínima.

Modelo XGBoost

Cada árbol $f_t(x_i)$ en el modelo se entrena para corregir los errores residuales de la predicción acumulada hasta el árbol $t-1$.

Matemáticamente, el nuevo árbol $f_t(x_i)$ ajusta esta predicción según:

$$\hat{y_i}^t = \hat{y_i}^{(t-1)} + f_t(x_i)$$

Donde,

$\hat{y}_i^{(t-1)}$ es la predicción para la observación i hasta la iteración $t-1$.

$$\text{Predicción final} = \hat{y}_i = \sum_{k=1}^{T} \left(f_k(x_i)\right)$$

Modelo XGBoost

El entrenamiento sigue la esencia, el esquema Gradient Boosting, donde los pasos clave está el uso de expansión de Taylor de segundo orden para optimización. El algoritmo recorre iterativamente etapas:

- Inicialización:

  Se inicializa con un modelo con predicción constante (base_score) inicial (usualmente el promedio del target, árbol(0)).

- Cálculo de gradientes (primer orden) y Hessianos (segunda derivada):

  Para cada muestra i, se calculan el gradiente g_i y el Hessiano h_i de la función de perdida l con respecto a la predicción actual $\hat{y}_i^{(t-1)}$:

$$g_i = \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}}, \; h_i = \frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^2}$$

Modelo XGBoost

- Construcción de un nuevo árbol f_t:

  Se usan gradientes y Hessianos para seleccionar divisiones con ganancia positiva; el crecimiento se detiene al alcanzar límites predefinidos.

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma$$

- Cálculo de pesos de las hojas:

  Para comparar y regularizar el árbol, para cada hoja j se calcula el peso óptimo $w_j = -\frac{G_j}{H_j+\lambda}$, y la ganancia total, que penaliza el número de hojas:

$$Gain(Obj)_{tree} = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda} + \gamma T$$

Modelo XGBoost

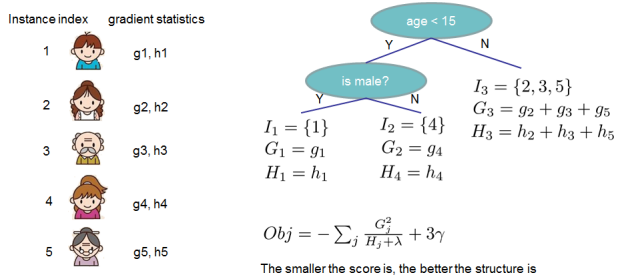


Figure 2: Ejemplo de árbol de decisión (adaptado de Chen & Guestrin, 2016).

- El primer término (negativo) mide la ganancia (cuanto menor sea mejor ajusta la hoja).
- El segundo término penaliza la complejidad (3 hojas)

Modelo XGBoost

- Aplicación del árbol:

  El árbol f_t entrenado se escala con la tasa de aprendizaje η (eta) para controlar su impacto, multiplicando sus predicciones por η (por defecto 0.3). Este “shrinkage” reduce el sobreajuste al limitar la contribución de cada árbol. Luego, las predicciones acumuladas se actualizan como:

$$\hat{y}_i^t = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

- Iteración y parada:

  El proceso (cálculo de gradientes, construcción del árbol y actualización) se repite por un número fijo de rondas o hasta cumplir un criterio de parada, sumando árboles que corrigen errores previos.

Modelo XGBoost

El modelo fue entrenado con las variables ya preprocesadas descritas anteriormente.

Para optimizar el rendimiento, se aplicó RandomizedSearchCV, evaluando 30 combinaciones de hiperparámetros con validación cruzada. Dado el fuerte desbalance observado entre las clases, se ajustó el parámetro scale_pos_weight para penalizar errores en la clase minoritaria.

$$scale\_pos\_weight = \frac{\text{número de horas sin lluvia en entrenamiento}}{max(1, \text{número de horas con lluvia en entrenamiento})}$$

Modelo XGBoost

Se eligió XGBoost por su buen rendimiento en tareas de clasificación binaria, su eficiencia para trabajar con datos numéricos y categóricos, y su capacidad para manejar clases desbalanceadas.

Se configuró con la función objetivo binary:logistic, y se utilizó la métrica AUC para evaluar el rendimiento global del modelo. Durante el entrenamiento, se usó la balanced accuracy como métrica de optimización, por ser más adecuada en contextos con clases desbalanceadas.

Clase	Cantidad	Porcentaje
0 (No lluvia)	52.594	95.84%
1 (Lluvia)	2.282	4.16%

Table 1: Distribución de clases en la variable lluvia

Desarrollo de modelos (XGBoost)

Modelo de clasificación: Búsqueda de parámetros (para lluvia)

Se define una función de búsqueda de los parámetros óptimos del modelo de clasificación (binary:logistic) con validación cruzada (5 k-folds) y métrica de evaluación AUC ($balanced\_accuracy = 0.854$)

Parámetro	Valor
estimator	base_clf
param_distributions	param_dist
n_iter	30
scoring	'balanced_accuracy'
cv	cv
n_jobs	-1
random_state	42
verbose	1
refit	True

Búsqueda clasificación

Parámetro	Valor óptimo
n_estimators	200
max_depth	3
learning_rate	0.05
subsample	0.6
colsample_bytree	1.0
gamma	0
scale_pos_weight	24.051

Mejores parámetros

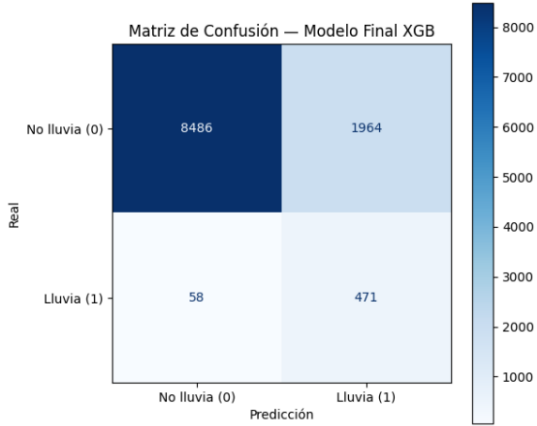
Modelo de clasificación: Definición del modelo final

Usando los parámetros encontrados y estableciendo el 80% de los datos para entrenamiento y 20% para validación.

Parámetro	Valor
params	(valores óptimos encontrados best_params)
objective	'binary:logistic'
eval_metric	'auc'
use_label_encoder	False
random_state	42

Definición modelo de clasificación final

Modelo de clasificación: Matriz de confusión (para lluvia)



Matriz de confusión para clasificación de lluvia

Modelo de regresión: Búsqueda de parámetros (para agua caída)

Se define una función de búsqueda de los parámetros óptimos del modelo de regresión (reg:squarederror) con validación cruzada (5 k-folds) y métrica de evaluación RMSE ($RMSE = 0.5986$)

Parámetro	Valor
estimator	base_reg
param_distributions	param_dist
n_iter	30
scoring	'neg_root_mean_squared_error'
cv	cv
n_jobs	-1
random_state	42
verbose	1
refit	True

Búsqueda regresión

Parámetro	Valor óptimo
n_estimators	200
max_depth	5
learning_rate	0.05
subsample	1.0
colsample_bytree	0.6
gamma	1
reg_alpha	1
reg_lambda	1

Mejores parámetros

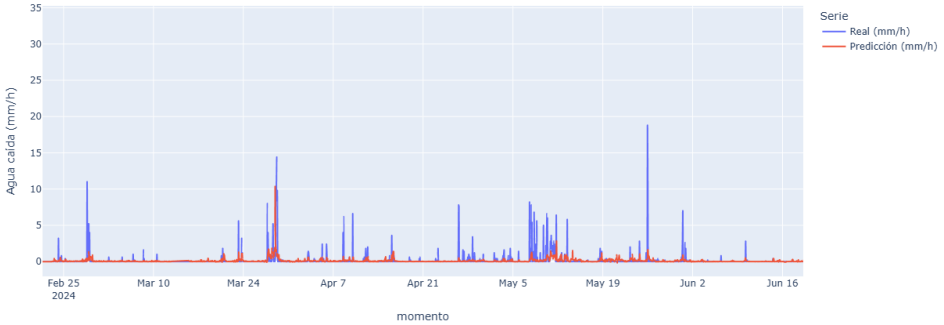
Modelo de regresión: Definición del modelo final (para agua caída)

Usando los parámetros encontrados y estableciendo el 80% de los datos para entrenamiento y 20% para validación.

Parámetro	Valor
params	(valores óptimos encontrados best_params)
objective	'reg:squarederror'
eval_metric	'rmse'
use_label_encoder	False
random_state	42

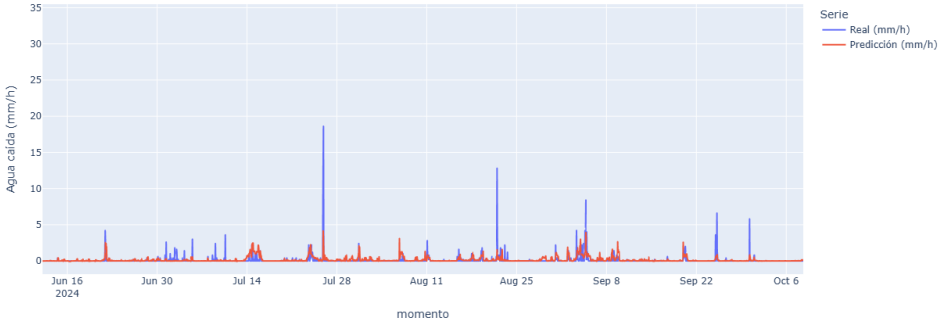
Definición modelo de regresión final

Modelo de regresión: Serie de tiempo pronósticos (para agua caída)



Validación de predicción para agua caída entre febrero y junio de 2024

Modelo de regresión: Serie de tiempo pronósticos (para agua caída)



Validación de predicción para agua caída entre junio y octubre de 2024

Modelo de regresión: Métricas (para agua caída)

El modelo definido entregó las siguientes métricas en la validación.

Métrica	Valor
RMSE Test	0.8903
R^2 Test	0.1161

Desempeño en conjunto de prueba

Discusión

Discusión

- Clasificación:

  XGBoost detectó el 81% de las horas sin lluvia y alcanzó un recall del 89% para lluvia, pero su precisión en la clase lluvia fue solo del 19%, generando una alta tasa de falsos positivos. Esto limita su uso como sistema de alerta, aunque resulta muy útil para identificar condiciones secas (precisión sin lluvia: 99%) y así reducir pérdidas por sequía.

- Regresión:

  Se logró anticipar momentos en que hay precipitación, pero no se estimó bien su magnitud. El bajo $R^2 = 0.1161$ y un RMSE de 0.89 mm/h indican que hacen falta variables relevantes o más datos para mejorar la estimación precisa de la cantidad de lluvia.

Conclusiones

Conclusiones

- XGBoost resultó óptimo en clasificar la ocurrencia de lluvia, especialmente en casos sin lluvia. La detección de momentos secos puede apoyar decisiones agrícolas en Rapa Nui.
- La detección de lluvias es óptima, pero la alerta de eventos lluviosos aún requiere mejora.
- El modelo de regresión suele anticipar cuándo llueve, pero subestima consistentemente la cantidad de precipitación.
- En general, los modelos captan correctamente horas con lluvia. No obstante, la cantidad de falsos positivos y al ser clases desbalanceadas derivan a un acierto de lluvia del 19%.

Limitaciones o sesgos

Imputación

Vecinos más cercanos (KNN) puede suavizar lluvias extremas al basarse en vecinos mayoritarios.

Subestimación

Se sugiere la falta de predictores clave para explicar magnitudes altas de lluvia.

Desbalance

Mayor cantidad de horas sin lluvia complica el ajuste de los modelos.

Umbral fijo

Un umbral de 0.4 mm/h fijo para lluvia no refleja variaciones estacionales.

Líneas de trabajo futuro

- **Mayor cobertura**: Incluir datos satelitales (anomalías, densidad de nubes, etc.).
- **Modelos estacionales**: Entrenar versiones de modelos por estación del año.
- **Actualización automática**: Automatizar la actualización de datos para adaptar el modelo.
- **Umbrales dinámicos**: Evaluar umbrales dinámicos para definir “lluvia” según la temporada.

Referencias

[1] American Meteorological Society. (s.f.). Drizzle. En *Glossary of Meteorology*. Recuperado el 19 de junio de 2025, de *https://glossary.ametsoc.org/wiki/Drizzle*

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. *https://doi.org/10.1145/2939672.2939785*

[3] Chen, T. (2024). *XGBoost model tutorial*. XGBoost Documentation. Recuperado de *https://xgboost.readthedocs.io/en/stable/tutorials/model.html*

[4] Chen, T. (2025). XGBoost Documentation. Recuperado el 24 de junio de 2025, de *https://xgboost.readthedocs.io/en/latest/*

Referencias

[5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. *https://doi.org/10.1214/aos/1013203451*

[6] Gultepe, I. (2008). Measurements of light rain, drizzle and heavy fog. En S. Michaelides (Ed.), *Precipitation: Advances in Measurement, Estimation and Prediction* (pp. 59–82). Springer. *https://doi.org/10.1007/978-3-540-77655-0_3*

[7] International Civil Aviation Organization & World Meteorological Organization. (2021). Considerations on precipitation intensity thresholds (AMOFSG/10-IP/6). Recuperado de *https://www.icao.int/safety/meteorology/amofsg/amofsg%20meeting%20material/amofsg.10.ip.006.5.en.pdf*

Referencias

[8] Kumar, G. D., Tyagi, S., Pradhan, K. C., & Shah, A. (2025). District-Level Rainfall and Cloudburst Prediction Using XGBoost: A Machine Learning Approach for Early Warning Systems. *Informatica*, 49(2), 375–396. *https://doi.org/10.31449/inf.v49i2.7612*

[9] López-Fernández, J. J., García-Torres, M., Luna, J. M., Ventura, S., & Riquelme, J. C. (2024). *Evaluating missing data imputation techniques in multivariate time series with attention mechanisms*. Pattern Analysis and Applications. *https://doi.org/10.1007/s10044-024-01262-3*

[10] Mujahid, A., & Rafique, M. (2024). *Comparative analysis of missing value imputation techniques for time series data*. Journal of Data Analysis and Information Processing, 12(2), 151–172. *https://doi.org/10.4236/jdaip.2024.122009*

Referencias

[11] Syah, M. S. I., Nardi, & Rachmawardani, A. (2025). Evaluation of the XGBoost Model for Rainfall Prediction and Classification Using BMKG Data and OpenWeather API. *Journal of Computation Physics and Earth Science*, 5(1), 21–30. *https://doi.org/10.63581/JoCPES.v5i1.03*

[12] World Meteorological Organization. (s.f.). Drizzle. En *WMO Cloud Atlas*. Recuperado el 19 de junio de 2025, de *https://cloudatlas.wmo.int/es/drizzle.html*