“Evaluación de dos enfoques para la predicción de precipitaciones en la región de Junín: un modelo híbrido de clasificación-regresión y un modelo de clasificación multiclase balanceado”

Autores: Lino, P., Calderón, C., Chavarria, M., Izarra, L., Mascco, P.

Resumen- Este trabajo aborda el problema de la predicción de precipitaciones en la región de Junín, la cual radica en la alta proporción de registros sin lluvia que introduce un sesgo significativo en modelos tradicionales de predicción. El objetivo general fue desarrollar y evaluar dos enfoques alternativos: un modelo híbrido que combina clasificación binaria para determinar la ocurrencia de lluvia y regresión para estimar su cantidad; y un modelo de clasificación multiclase entrenado con técnicas de balanceo para predecir directamente niveles discretos de precipitación. Aunque el modelo híbrido permite una estimación más detallada del volumen de lluvia, los resultados muestran que el modelo multiclase balanceado obtuvo mejor rendimiento general, especialmente en la detección de eventos de lluvia significativa.

1. Introducción

Las precipitaciones extremas en la región de Junín constituyen un factor climático determinante para el sector agrícola, especialmente en lo relacionado con la planificación del riego, la gestión de cultivos y la productividad de las tierras. Durante la temporada de lluvias, la variabilidad en la cantidad e intensidad de las precipitaciones afecta directamente la disponibilidad de agua para uso agrícola, pudiendo generar tanto déficits hídricos como excesos que dañan los cultivos.

Según el Instituto Nacional de Defensa Civil (INDECI), entre 2012 y 2023 se registraron 136 emergencias por inundaciones en la región de Junín, afectando a más de 13 000 personas, 2200 viviendas y 2300 hectáreas agrícolas, principalmente en los distritos de Chilca, Jauja, Concepción y Huancayo. Estos eventos extremos no solo ponen en riesgo las cosechas, sino que también dificultan una gestión eficiente del riego, lo que evidencia la necesidad de contar con herramientas predictivas que permitan anticipar la ocurrencia y magnitud de las lluvias para una mejor toma de decisiones en el agro.

Además, en base a los datos obtenidos de condiciones atmosféricas y temporales, se ha observado que alrededor del 95% de los registros presentan valor cero en la variable de precipitación, la cual se expresa en términos de la profundidad vertical de agua que cubriría una superficie horizontal del terreno en una hora, medida en milímetros (mm).

Ante este panorama y ante la marcada desproporción de registros con valor cero en la variable objetivo de precipitación, este trabajo plantea como hipótesis que el uso de dos modelos con enfoques distintos pueden mitigar los efectos del desbalance y mejorar la capacidad predictiva del sistema. Con ese fin, se desarrollaron y evaluaron dos enfoques alternativos: un modelo híbrido, compuesto por una etapa de clasificación binaria seguida de una regresión para estimar la cantidad de lluvia, y un modelo de clasificación multiclase entrenado con técnicas de balanceo para predecir directamente niveles discretos de precipitación. El objetivo principal fue determinar cuál de estos enfoques resulta más eficaz para anticipar la ocurrencia e intensidad de las lluvias. Este estudio busca, en última instancia, contribuir a una mejor gestión agrícola en regiones vulnerables como Junín, donde una predicción adecuada de las precipitaciones es clave para optimizar el riego, prevenir pérdidas en cultivos y tomar decisiones informadas sobre el uso del recurso hídrico.

2. Trabajos relacionados

A continuación se presenta un resumen de artículos científicos relacionados al objetivo del artículo.

Título: Predicción de precipitación mensual mediante Redes Neuronales Artificiales para la cuenca del río Cali, Colombia.

El estudio aplica Redes Neuronales Artificiales para predecir la precipitación mensual en la cuenca del

río Cali bajo escenarios climáticos RCP. Se usaron datos de 35 estaciones (1972–2016) y se completaron faltantes con redes tipo Multilayer Perceptron. Las redes se entrenaron con funciones hiperbólicas y hasta 6000 iteraciones logrando hasta un 98% de precisión, útil para la planificación local.

Título: MT-HCCAR: Multi-Task Deep Learning with Hierarchical Classification and Attention-based Regression for Cloud Property Retrieval

Este artículo propone una metodología basada en el diseño de un modelo de aprendizaje profundo multitarea denominado MT-HCCAR, el cual aborda simultáneamente la clasificación jerárquica de nubes y la estimación del grosor óptico de la nube (COT). El modelo se compone de dos subredes principales: una subred de clasificación jerárquica que primero determina la presencia de nubes y luego identifica su fase (líquida, sólida o mixta), y una subred de regresión asistida por atención que estima el grosor óptico únicamente en los casos donde se ha detectado una nube. Esta última incorpora un mecanismo de atención cruzada que integra la información extraída por la subred de clasificación para mejorar la precisión de la regresión. Ambas subredes comparten una arquitectura convolucional base y se entrenan de forma conjunta mediante un esquema de optimización multitarea, lo que permite aprovechar las sinergias entre las tareas para mejorar el desempeño global. El entrenamiento se llevó a cabo con datos simulados de sensores satelitales (OCI, VIIRS y ABI), aplicando validación cruzada con K-fold y utilizando la regla de un error estándar (1SE) para la selección del modelo final.

Título: A Methodology Based on Random Forest to Estimate Precipitation Return Periods: A Comparative Analysis with Probability Density Functions in Arequipa, Peru.

Este estudio compara Random Forest (RF) con funciones de densidad de probabilidad para estimar precipitaciones extremas en Arequipa, Perú. Con datos de 26 estaciones (1965–2013), RF mostró mayor precisión en 61–69% de los casos frente a PDFs como Gumbel o GEV. RF destacó por su robustez ante datos ruidosos y su menor error en periodos de retorno de 2 a 100 años.

Título: Variabilidad climática sobre el rendimiento de los cultivos de seguridad alimentaria en la sierra - Junín

Este estudio analiza cómo la variabilidad climática (temperatura, precipitación y humedad) afecta el rendimiento de 27 cultivos en la sierra de Junín entre los años 2015 y 2022. Se utilizaron datos climáticos y de rendimiento obtenidos de MIDAGRI y SENAMHI, y se evaluaron estrategias de adaptabilidad (EA) aplicadas por los agricultores, como medidas técnicas y agroecológicas. Para el análisis se usaron métodos estadísticos como pruebas t, correlación de Pearson y regresión logística binaria.

Título: Advancing preeclampsia prediction: a tailored machine learning pipeline integrating resampling and ensemble models for handling imbalanced medical data

La metodología del estudio consistió en el desarrollo de un pipeline especializado para la predicción temprana de preeclampsia en conjuntos de datos médicos desequilibrados. Para abordar el desbalance de clases, se aplicaron ocho técnicas de remuestreo, entre ellas SMOTE, ADASYN y el novedoso IWGMM, variando la proporción minoría/mayoría (MMR) entre 0.05 y 1. Posteriormente, se entrenaron modelos con seis algoritmos de aprendizaje de ensamble (como GBDT y Random Forest), generando 4,608 configuraciones. La evaluación de desempeño utilizó métricas adaptadas al desbalance, donde destaca G-mean, MCC, AP y AUC. Finalmente, se aplicó una optimización iterativa que ajustó resampling, algoritmo y MMR en el orden más efectivo, y se comprobó la eficacia del pipeline en tres datasets públicos adicionales para confirmar su capacidad de generalización.

Título: SMOTE for Handling imbalanced Data Problem: A Review

Este artículo aborda el problema de la desigualdad en la distribución de clases en tareas de clasificación, donde una clase está representada por muchas menos instancias que otras, lo que perjudica la precisión del modelo, especialmente en la clase minoritaria.

Como solución, se destaca el uso de SMOTE (Synthetic Minority Oversampling Technique), un método pionero de sobremuestreo que genera instancias sintéticas combinando una instancia minoritaria con sus *K* vecinos más cercanos en el espacio de características. Esto ayuda a evitar el sobreajuste y mejora la capacidad del clasificador para encontrar límites de decisión más precisos entre clases.

3. Metodología

- Enfoque(s) propuesto:

Para abordar el problema de predicción de precipitaciones en un contexto altamente desbalanceado, se evaluaron dos enfoques distintos. El primero es un modelo híbrido que combina clasificación binaria y regresión: el modelo de clasificación predice la ocurrencia o no de lluvia, y solo en los casos positivos, el modelo de regresión

estima la cantidad acumulada en milímetros. El segundo enfoque consiste en un modelo de clasificación multiclase entrenado con la técnica SMOTE. Esta técnica permite generar registros sintéticos de las clases minoritarias a partir de un punto de la clase minoritaria y sus vecinos.

La motivación para aplicar estos enfoques radica en que aproximadamente el 95 % de los registros del conjunto de datos presentan valor cero en la variable objetivo *Precipitación*, lo que introduce un sesgo significativo que afecta el rendimiento de los modelos de regresión tradicionales. Ante esta distribución altamente desbalanceada, se consideró que dividir la tarea o transformar la variable objetivo permitiría mejorar la capacidad predictiva.

Los datos utilizados como entrada para los modelos comprenden diversos tipos de variables, que incluyen tanto información temporal, como las fechas de muestreo, como variables atmosféricas tales como temperatura, humedad y velocidad del viento. En el primer enfoque, la salida del modelo de clasificación es una predicción binaria que indica si se espera o no la ocurrencia de precipitaciones. Por su parte, la salida del modelo de regresión es la cantidad de precipitación acumulada en una hora, expresada en milímetros. En el segundo enfoque, la salida corresponde a una clasificación multiclase que predice directamente el nivel de precipitación esperada categorizada en intervalos definidos, los cuales son “Sin lluvia” (nivel 0), “Ligera” (nivel 1) y “Moderada/Fuerte” (nivel 2).

La fórmula para la generación de registros sintéticos en la técnica SMOTE es la siguiente:

$$x_{\text{nuevo}} = x_i + \delta \cdot (x_{vecino} - x_i)$$

Figura 1: Fórmula de generación para registros sintéticos

Donde:
x_i = Un registro de la clase minoritaria. En este caso, puede ser un registro de caso de lluvia moderada/fuerte.
x_{vecino} = Otro registro de la clase minoritaria cercano al registro x_i.
x_{nuevo} = Registro sintético de la clase minoritaria.
δ = Un número aleatorio entre 0 y 1

A. Adquisición de los datos.

La información empleada para entrenar los modelos proviene de datasets disponibles en la Plataforma Nacional de Datos Abiertos del Gobierno del Perú, que provienen de la Estación Meteorológica Automática (EMA) y de la Estación BSRN para el Balance Energético Solar Tierra. En total, se utilizaron estos dos conjuntos de datos que sirvieron como base para el proceso de entrenamiento de los modelos, los cuales cuentan con los siguientes atributos:

VARIABLE	DESCRIPCIÓN	TIPO
FECHA_CORTE	Fecha de corte de información	Numérico
UBIGEO	Código de Ubicación Geográfica que denotan "DDppdd" (Departamento, provincia, distrito)	Alfanumérico
YY	Year, año de registro	Numérico
MM	Month, mes de registro	Numérico
DY	Day, día de registro	Numérico
HH	Hour, hora promedio del registro	Numérico
TT	Air Temperature 2 meters (temperatura del aire a 2 metros) - [°C]	Numérico
HR	Relative Humidity (Humedad Relativa) - [%]	Numérico
RR	Precipitation (precipitación) - [mm]	Numérico
PP	Atmospheric pressure (presión atmosférica) - [hPa]	Numérico
FF	Wind speed (velocidad del aire) - [m/s]	Numérico
DD	Wind direction (dirección del aire) - [grados]	Numérico

Figura 2: Diccionario de datos del dataset EMA

VARIABLE	DESCRIPCIÓN	TIPO
FECHA_CORTE	Fecha de corte de información	Numérico
UBIGEO	Código de Ubicación Geografica que denotan "DDppdd" (Departamento, provincia,distrito)	Alfanumérico
year	Year, año de registro	Numérico
month	Month, mes de registro	Numérico
day	Day, dia de registro	Numérico
hour	Hour, hora promedio del registro	Numérico
incide_LW_rad	Incident longwave irradiance (irradiancia incidente de onda larga) W m^{-2}	Numérico
emitte_LW_rad	Emitted longwave irradiance (irradiancia emitida de onda larga) W m^{-2}	Numérico
direct_SW_rad	Direct shortwave irradiance (irradiancia directa de onda corta) W m^{-2}	Numérico
global_SW_rad	Global shortwave irrandiance (irradiancia glocal de onda corta) W m^{-2}	Numérico
diffus_SW_rad	Diffuse shortwave irradiance (irradiancia difusa de onda corta) W m^{-2}	Numérico
reflec_SW_rad	Reflected shortwave irradiance (irradiancia reflejada de onda corta) W m^{-2}	Numérico

Figura 3: Diccionario de datos del dataset BSRN

Luego, se renombraron las columnas de año, mes, día y hora del dataset EMA a los nombres que aparecen en el dataset BSRN a fin de realizar la unión de las columnas de los dos datasets en un solo dataset.

B. Preprocesamiento de los datos.

Primero, se realizó la eliminación de filas con valores nulos dado que el porcentaje de nulos de ninguna columna superaba el 25%. Luego, se observó qué meses son los que presentan mayor frecuencia de lluvias, a fin de reducir el porcentaje de registros con valor 0 en la variable objetivo y centrar el trabajo en la predicción de lluvias en períodos de lluvias. Adicionalmente, se realizó la conversión de la variable hora a dos variables radiales por su naturaleza cíclica y la eliminación del resto de variables temporales. Luego, se realizó la proyección de las variables en PC1 y PC2 usando PCA para ver la variabilidad que provee cada variable en sus dos proyecciones. Después, se generó una matriz de correlación a fin de observar si existían columnas redundantes. Las columnas redundantes encontradas se eliminaron del dataset y se mantuvieron las que se consideraron más genéricas para la característica.

C. Entrenamiento y selección del modelo óptimo

Para cada etapa del primer enfoque, previamente se dividió el conjunto de datos en dos subconjuntos utilizando la técnica de train-test split, asignando el

80% de los datos al conjunto de entrenamiento y el 20% al conjunto de prueba. En el caso de la etapa de selección de modelo de regresión, todos los registros deben presentar un valor mayor a 0 en la variable objetivo Precipitación.

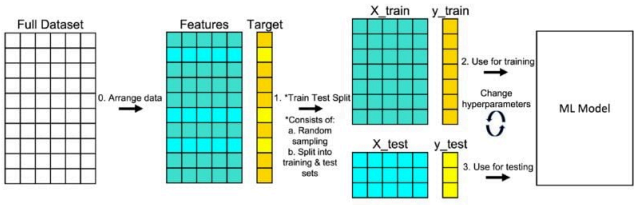

Figura 4: Train-test split
Fuente: Galarnyk, M. (s. f.). Train Test Split: What It Means and How to Use It

Luego, en cada etapa se entrenaron diversos modelos de machine learning con el fin de identificar el modelo que ofreciera el mejor desempeño. En la etapa de clasificación, se evaluaron algoritmos como regresión logística, clasificación K-Nearest Neighbors y clasificación con Decision Trees. El criterio utilizado para seleccionar el mejor modelo fue el valor de balanced accuracy, eligiéndose aquel con el mayor puntaje en esta métrica. En la etapa de regresión, se evaluaron algoritmos como regresión lineal, regresión K-Nearest Neighbors y regresión con Decision Trees. En este caso, el criterio utilizado fue el error cuadrático medio (MSE), eligiéndose aquel con menor valor en esta métrica.

Para ambas etapas, se usó la estrategia de validación cruzada con 10 folds para asegurar una evaluación robusta y reducir la varianza en los resultados. Esta técnica permitió estimar el desempeño general de cada modelo en distintos subconjuntos del conjunto de datos con el fin de evitar el sobreajuste y garantizar una mejor generalización.

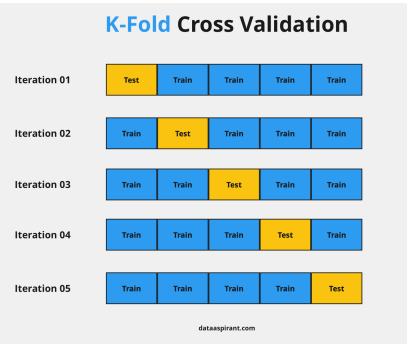

Figura 5: Validación cruzada con K-Folds
Fuente: Arif, A. (2020). How cross-validation works in machine learning

Además, para la etapa de selección de modelo de regresión, se utilizaron adicionalmente pipelines que permiten el escalado o estandarización de las variables previo a la aplicación de cada algoritmo.

Finalmente, los modelos seleccionados en cada etapa fueron evaluados sobre el conjunto de prueba, con el fin de validar su capacidad predictiva en datos no vistos durante el entrenamiento. Para ello, se utilizaron las métricas de error cuadrático medio, error absoluto medio, varianza explicada y coeficiente de determinación (R^2).

D. Combinación de modelos

Previamente, se había entrenado de manera independiente cada modelo seleccionado, con datos de entrenamiento y de prueba distintos. Ahora, en esta última etapa, se entrenó ambos modelos con los mismos datos de entrenamiento y de prueba, con el fin de integrarlos en un pipeline conjunto que permita realizar predicciones secuenciales de forma coherente.

Específicamente, el modelo de clasificación fue el primero en ejecutarse, recibiendo como entrada el vector de características y determinando si se espera la ocurrencia de precipitaciones. En caso de una predicción positiva, el mismo vector fue transmitido al modelo de regresión para estimar la cantidad de precipitación acumulada. En caso contrario, el valor de salida correspondiente a la regresión fue automáticamente asignado como 0 mm para evitar predicciones innecesarias en condiciones secas.

Finalmente, en esta etapa se evaluaron las métricas de desempeño correspondientes para ambos modelos y el modelo combinado: en el modelo de clasificación se analizó la métrica balanced accuracy, mientras que para el modelo de regresión y para el modelo combinado se calcularon el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2).

En el segundo enfoque, previamente se dividió el conjunto de datos en dos subconjuntos utilizando la técnica de train-test split, asignando el 80% de los datos al conjunto de entrenamiento y el 20% al conjunto de prueba.

Luego, se redujo el número de los registros de la clase mayoritaria en el conjunto de entrenamiento para lograr un conjunto más equilibrado antes de aplicar SMOTE con el fin de evitar generar muchos registros sintéticos de las clases minoritarias. Luego, se generaron los registros sintéticos de las clases minoritarias. Cada clase presenta exactamente la misma cantidad de registros, el cual es igual a la cantidad de la clase mayoritaria inicialmente.

Finalmente, se realizó el entrenamiento y la selección del modelo óptimo de manera similar al modelo de clasificación binaria del primer enfoque.

4. Experimentación y Resultados

- Setup experimental:

5. Conclusión

6. Sugerencias de trabajos futuros

7. Implicancias éticas

8. Link del repositorio del trabajo

https://github.com/MrLinoP/Modelos-Predictivos-Precipitacion

9. Declaración de contribución de cada integrante

Lino, J.: Planteamiento de la hipótesis, obtención de datasets, desarrollo de la metodología y el código (primer enfoque).
Calderón, C.: Redacción de la metodología, hipótesis, objetivo y enfoque a aplicar. Síntesis de dos trabajos relacionados. Desarrollo de la metodología y el código (segundo enfoque).
Chavarria, M.: Adición de dos fuentes académicas para su uso en la parte de trabajos relacionados. Aporte de visión para la metodología.
Izarra, L.: Síntesis de un trabajo relacionado. Contribución en la etapa temprana del desarrollo de la metodología.
Mascco, P.: Obtención de fuentes académicas que sirven como guía en el modelo, redacción de la introducción y aporte de ideas en el código.

10. Referencias

[1]. MONTENEGRO-MURILLO, Daniel David, Mayra Alejandra PÉREZ-ORTIZ y Viviana VARGAS-FRANCO
2019 "Predicción de precipitación mensual mediante Redes Neuronales Artificiales para la cuenca del río Cali, Colombia". *Revista DYNA*. Medellín, volumen LXXXVI, número 211, pp. 122-130. Consulta: 2 de junio de 2025. https://revistas.unal.edu.co/index.php/dyna/article/view/76079/73107

[2]. LI, Xingyan, Andrew SAYER, Ian CARROLL, Xin HUANG y Jianwu WANG
2024 "MT-HCCAR: Multi-Task Deep Learning with Hierarchical Classification and Attention-based Regression for Cloud Property Retrieval". *Lecture Notes in Computer Science*. Heidelberg, volumen XIVCML, número 1, pp. 3–18. Consulta: 6 de junio de 2025. https://arxiv.org/html/2401.16520v1#S2

[3]. PAREDES, Victor
2023 Variabilidad climática sobre el rendimiento de los cultivos de seguridad alimentaria en la sierra-Junín. Tesis de doctorado en Ciencias Ambientales y Desarrollo Sostenible. Huancayo: Universidad Nacional del Centro del Perú, Facultad de Ciencias Forestales y del Ambiente. Consulta: 6 de junio de 2025. https://repositorio.uncp.edu.pe/handle/20.500.12894/10334

[4]. MA, Yinyao, Hanlin LV, Yanhua MA, Xiao WANG, Longting LV, Xuxia LIANG y Lei WANG
2025 "Advancing preeclampsia prediction: a tailored machine learning pipeline integrating resampling and ensemble models for handling imbalanced medical data". *Revista BioData Mining*. Volumen XVIII, artículo 25, pp. 1–16. Consulta: 6 de junio de 2025. https://link.springer.com/article/10.1186/s13040-025-00440-1

[5]. GALARNYK, Michael
Train Test Split: What It Means and How to Use It. Consulta: 6 de junio de 2025. https://builtin.com/data-science/train-test-split

[6]. ARIF, Anber
2020 How cross-validation works in machine learning. Consulta: 6 de junio de 2025. https://dataaspirant.com/cross-validation/

[7]. ANCO-VALDIVIA, Johan, Sebastián VALENCIA-FÉLIX y otros
2025 "A Methodology Based on Random Forest to Estimate Precipitation Return Periods: A Comparative Analysis with Probability Density Functions in Arequipa, Peru". *Water*. Volumen XVII, número 1, artículo 128. Consulta: 6 de junio de 2025. https://www.mdpi.com/2073-4441/17/1/128

[8]. Instituto Nacional de Defensa Civil (INDECI).
2023 Compendio Estadístico del INDECI 2023: Gestión Reactiva. Lima, Perú. pp 43-44. Consulta: 6 de junio de 2025. https://cdn.www.gob.pe/uploads/document/file/5591372/4965310-compendio-final-af-2023-indeci.pdf

[9]. Pradipta, G. A., Wardoyo, R., Musdholifah, A., Sanjaya, I. N. H., y Ismail, M.
2021 "SMOTE for handling imbalanced data problem: A review". En 2021 sixth international conference on informatics and computing (ICIC). IEIE, 2021. p 1-8.

https://ieeexplore.ieee.org/abstract/document/9632912