“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”

Autor/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, lo cual 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[8]. 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 la hipótesis de que el uso de dos modelos con enfoques distintos pueden mitigar los efectos del desbalance y mejorar la capacidad predictiva. 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:

El conjunto de datos utilizados en este estudio presenta aproximadamente un 95% de valores igual a cero en la variable objetivo. Este fuerte desbalance introduce un sesgo significativo en el modelo, ya que tiende a favorecer a la clase

mayoritaria. Con el objetivo de mitigar este problema y mejorar el rendimiento predictivo, se propusieron los siguientes enfoques:

A. **Modelo híbrido:** Consiste en una etapa inicial de clasificación binaria que predice la ocurrencia de precipitación, seguido por un modelo de regresión que se encargará de estimar la cantidad de precipitación en milímetros, el cual actuará dependiendo de la salida del modelo clasificador binario.
B. **Modelo de clasificación multiclase balanceado:** Transforma la variable objetivo en tres categorías discretas ("Sin lluvia" (nivel 0), "Ligera" (nivel 1) y "Moderada/Fuerte" (nivel 2)) para luego entrenar el modelo usando la técnica SMOTE[9] que sirve para balancear las clases, generando datos sintéticos de las categorías minoritarias.

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[9]

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 supera 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

En ambos enfoques se aplicó una división en el conjunto de datos en dos subconjuntos utilizando la técnica de train-test split[5], asignando el 80% de los datos al conjunto de entrenamiento y el 20% al conjunto de prueba.

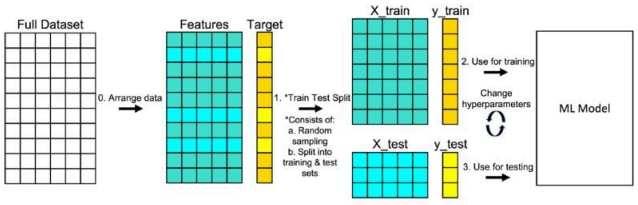

Figura 4: Train-test split
Fuente: Galarnyk, M. (s. f.). Train Test Split: What It Means and How to Use It

Para asegurar una evaluación robusta y reducir la varianza en los resultados, se empleó validación cruzada con 10 folds. Esta técnica permitió estimar de manera más confiable el desempeño de los modelos en distintos subconjuntos del conjunto de entrenamiento, mitigando el riesgo de sobreajuste y favoreciendo una mejor capacidad de generalización.

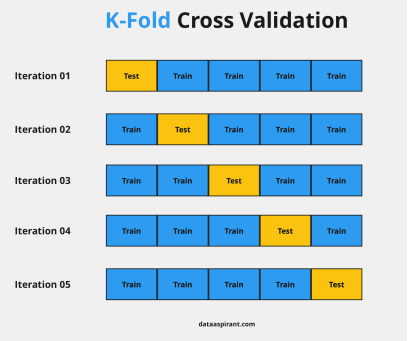

Figura 5: Validación cruzada con K-Folds
Fuente: Arif, A. (2020). How cross-validation works in machine learning

Modelo híbrido:

En la primera etapa de este enfoque que consiste en la clasificación de la ocurrencia de precipitación, se consideraron diversos modelos de machine learning con el fin de identificar cual obtendrá un mejor desempeño, tales como: regresión logística, clasificación K-Nearest Neighbors y clasificación con Decision Trees. Se utilizó como métrica principal balanced accuracy y se seleccionó el de mayor puntaje. En la etapa de regresión, se evaluaron algoritmos como: regresión lineal, K-Nearest Neighbors y regresión con Decision Trees. Se aplicaron pipelines con estandarización de variables.

Durante la implementación, el modelo de clasificación fue el primero en ser entrenado. Este recibía como entrada el vector de características y determinaba la probabilidad de ocurrencia de precipitaciones. Si la predicción era positiva, el mismo vector se transmitía al modelo de regresión para estimar la cantidad de precipitación acumulada. En caso contrario, el valor de salida se fijaba automáticamente en 0 mm, evitando así predicciones innecesarias en condiciones secas.

Posteriormente, se evaluó el rendimiento individual de cada modelo y del sistema combinado. Para el modelo de clasificación se utilizó *balanced accuracy*, mientras que para el modelo de regresión y el sistema combinado se calcularon el error cuadrático medio (MSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2).

Modelo de clasificación multiclase balanceado:

En este enfoque, se aplicó undersampling manual (reducción previa de la clase mayoritaria) en el conjunto de entrenamiento para evitar un exceso de muestras sintéticas. Luego se aplicó SMOTE[9] para equilibrar las clases, obteniendo como resultado un equilibrio entre las clases.

Luego de finalizar el proceso de balanceo de clases, se procedió al entrenamiento y selección del modelo óptimo mediante un procedimiento similar al del primer enfoque.

4. Experimentación y Resultados

- Setup experimental:

Los modelos fueron desarrollados en Python, utilizando Google Colab como entorno.

Luego del preprocesamiento de datos se procedió con el desarrollo de cada enfoque:

A. Modelo híbrido:

La primera etapa del modelo híbrido de machine learning, consiste en la regresión para predecir la cantidad de precipitación. De los diferentes algoritmos usados, se obtuvo el mejor resultado con el algoritmo Ridge, debido a su bajo error cuadrático medio y tener la menor varianza, luego se complementó con la adición de pipelines con el escalador StandardScaler.

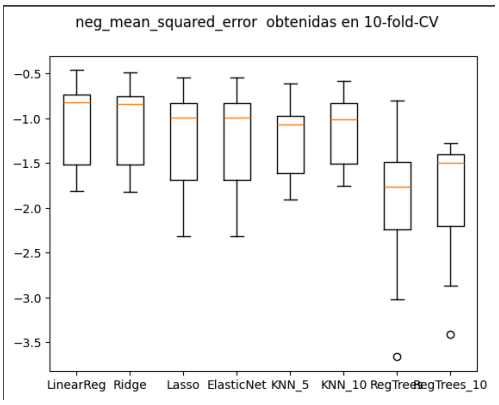

Figura 6: Resultados de los algoritmos usados con pipelines en el modelo híbrido

```
Mean squared error: 2.1502887146155114
Mean absolute error: 0.7969104336050022
Explained variance score: 0.09984637718972023
R2 score: 0.084124878114899
```

Figura 7: Resultados de la etapa de regresión usando el algoritmo ridge en el modelo híbrido

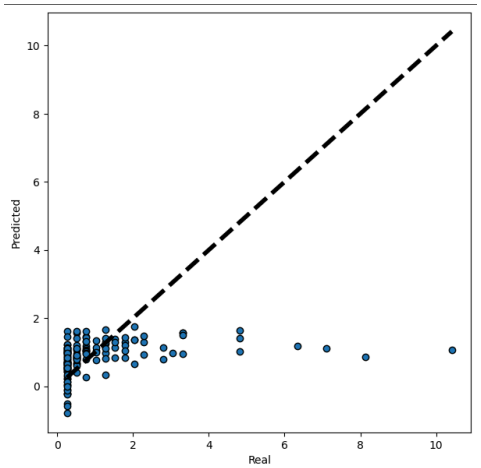

Figura 8: Scatter plot, comparación entre valores predichos y reales.

Para el algoritmo de clasificación se aplicaron los mismos algoritmos que se usaron en la etapa de regresión con el objetivo de tener una visión clara de cuál de ellos obtendrá el mejor resultado posible para luego implementarlo en el modelo híbrido.

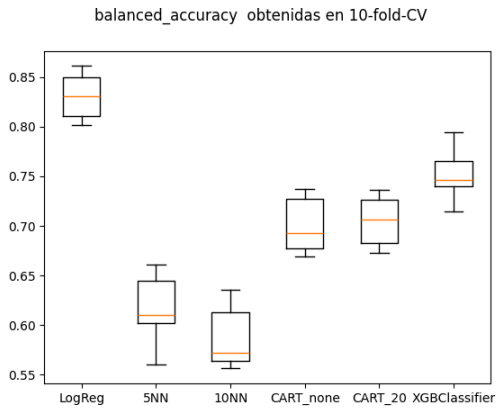

Figura 9: Boxplot de métricas en algoritmos de clasificación

B. Modelo de clasificación multiclase balanceado:

Se procedió a realizar el balanceo de clases aplicando el undersampling manual para luego utilizar la técnica SMOTE[9] obteniendo la siguiente distribución de datos:

Distribución de datos original

```
Counter({0: 1195, 1: 128, 2: 14})
```

Distribución de datos luego de aplicar SMOTE

```
Counter({0: 3700, 1: 3700, 2: 3700})
```

Figura 10: Distribución de los datos antes y después del balance.

Se dividió la variable objetivo en 3 clases tomando como referencia los niveles de lluvia, ("Sin lluvia" (nivel 0), "Ligera" (nivel 1) y "Moderada/Fuerte" (nivel 2)). Luego se entrenó el modelo con diferentes algoritmos tal y como se realizó con el primer enfoque para luego seleccionar el que tenga mejor desempeño, obteniendo los siguientes resultados:

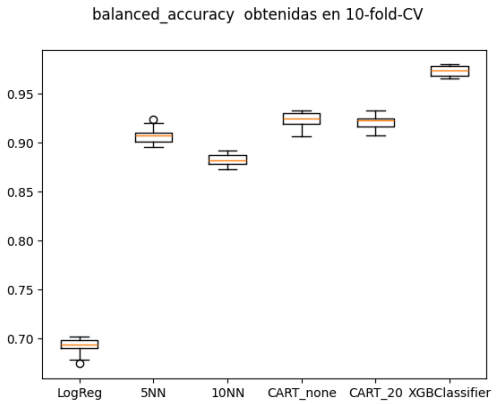

Figura 11: Boxplot de métricas de los algoritmos usados para el entrenamiento

Resultados y discusión:

A. Modelo híbrido:

Luego de los resultados obtenidos en las elecciones del algoritmos para cada etapa del primer enfoque, se eligió el modelo XGBClassifier en lugar de LogisticRegression, ya que, al combinarse con Ridge en el enfoque propuesto, obtuvo mejores resultados en comparación, según las métricas MSE, MAE y R^2 descritas en la metodología.

```
Clasificación:
Balanced Accuracy: 0.7373929391556627
Tamaño del conjunto de entrenamiento: 498
Tamaño del conjunto de prueba: 147

Regresión (solo en casos con lluvia predicha):
MAE: 0.858106719047044
MSE: 1.9534813808753446
R²: 0.08989580935995156

Clasificación + regresión:
MAE: 0.08896594601192197
MSE: 0.16984548503922015
R²: 0.3237407238283293
```

Figura 12: Resultados de las métricas luego de entrenar el modelo híbrido

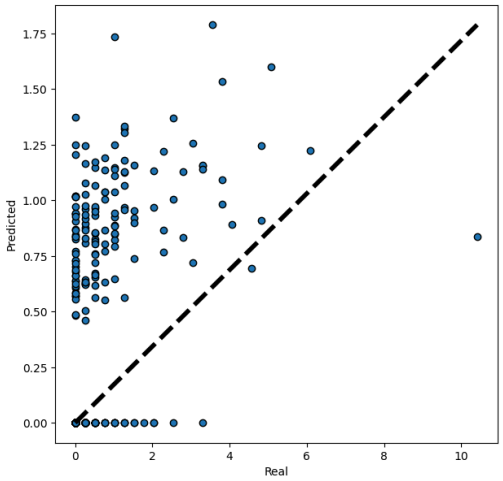

Figura 13: Scatter plot, resultados del modelo híbrido, comparación entre valores predichos y reales.

El resultado obtenido en el modelo no fue el esperado, ya que obtuvo un pobre rendimiento para la predicción de la precipitación.

B. Modelo de clasificación multiclase balanceado:

Se eligió el modelo LogisticRegression por su mayor balanced accuracy y su excelente capacidad para detectar lluvia moderada o fuerte (recall = 0.93), lo cual es prioritario, ya que minimizar los falsos negativos en esa clase es más importante que evitar falsas alarmas. El resultado obtenido fue el siguiente:

```
Balanced Accuracy score: 0.7678723973899183

Matriz de Confusión:
[[999 153  43]
 [ 23  69  36]
 [  0   1  13]]

Reporte de clasificacion en conjunto de validacion:

              precision    recall  f1-score   support

           0       0.98      0.84      0.90      1195
           1       0.31      0.54      0.39       128
           2       0.14      0.93      0.25        14

    accuracy                           0.81      1337
   macro avg       0.48      0.77      0.51      1337
weighted avg       0.90      0.81      0.85      1337
```

Figura 14: Matriz de confusión y tabla de resultados del entrenamiento.

Como se pudo apreciar en la Figura 14, los resultados obtenidos en el segundo enfoque presentaron una mejor predicción para la precipitación en comparación con el modelo híbrido.

5. Conclusión

En síntesis, este trabajo logró demostrar que, en primera instancia, nuestro primer enfoque, el modelo híbrido, obtiene mejores resultados si utiliza el modelo de clasificación mediante el algoritmo XGBClassifier y al modelo de regresión lineal mediante el algoritmo Ridge. Sin embargo, se observa que este enfoque otorga resultados por debajo de lo esperado e, inclusive, se pueden catalogar como malos resultados.

Por este motivo se recurrió al segundo enfoque, el modelo de clasificación multiclase entrenado con la técnica de balanceo SMOTE, la cual permite que el modelo de clasificación aprenda mejor los patrones asociados a la ocurrencia de lluvia.

Finalmente, se obtuvo que este modelo logra los mejores resultados utilizando el algoritmo LogisticRegression, ya que es el que mejor predice los casos donde hay una lluvia moderada o fuerte. Por lo cual, concluimos que el segundo enfoque es el que obtiene los resultados más óptimos.

6. Sugerencias de trabajos futuros

Mejora del preprocesamiento de datos:

- Implementar técnicas avanzadas de imputación para manejar datos faltantes o erróneos en variables metereológicas.
- Incorporar fuentes de datos satelitales o radares meteorológicos para enriquecer la calidad y resolución espacial del conjunto de datos.

Evaluación de modelos de Deep Learning:

- Explorar arquitecturas como LSTM, GRU o Transformers para modelar secuencias temporales y capturar patrones complejos de precipitación.
- Comparar su rendimiento frente a los enfoques híbrido y multiclase.

Validación en otras regiones agrícolas en Perú:

- Replicar el enfoque en otras regiones con características agroclimáticas como Arequipa, Cusco o Piura para evaluar su generalización.

Explicabilidad y confianza del modelo:

- Aplicar herramientas como SHAP o LIME para interpretar las decisiones del modelo, esto resulta importante para su adopción por agricultores o instituciones.

Predicción a múltiples horizontales temporales:

- Desarrollar modelos que no solo predigan la precipitación horaria, sino también a corto y mediano plazo.

7. Implicancias éticas

Uso responsable de los datos:

- Se deben garantizar principios como la calidad y privacidad de los datos metereológicos. Aunque no se trate de datos personales un mal procesamiento puede generar decisiones agrícolas equivocadas. Para abordar este problema, es fundamental implementar procesos de validación automática de datos, control de calidad en tiempo real y mecanismos de auditoría que aseguren la integridad y confiabilidad de la información utilizada por los modelos predictivos.

Riesgo de confianza ciega en modelos automatizados:

- Sí agricultores o instituciones se basan exclusivamente en las predicciones, sin entender sus limitaciones o márgenes de error, se corre el riesgo de tomar decisiones inadecuadas que perjudiquen cosechas. Para evitar esto, se debe promover la capacitación en interpretación de modelos y exigir el acompañamiento de informes de incertidumbre o márgenes de error.

Responsabilidad y transparencia en las predicciones:

- Es fundamental que los modelos sean explicables y auditables, para evitar decisiones arbitrarias o injustificadas en la asignación de recursos. Para ello, se deben emplear técnicas de interpretabilidad como SHAP o LIME, mantener registros detallados de los procesos de entrenamiento y validación, y documentar las decisiones del modelo para facilitar su revisión por expertos y partes interesadas.

Ética del sesgo y la representatividad:

- Si el modelo fue entrenado con datos limitados o no representativos de toda la región, puede generar sesgos que afecten más a ciertas zonas o tipos de cultivo. De existir, es una obligación ética revisar y corregir estos sesgos. Para ello, es necesario realizar análisis de equidad en el desempeño del modelo entre diferentes subregiones o tipos de cultivo y complementar los datos con fuentes más diversas.

Sostenibilidad a largo plazo del sistema:

- Si el sistema se implementa sin un plan de sostenibilidad, podría fallar con el tiempo, generando dependencia de una tecnología que luego quede obsoleta. La ética del diseño exige pensar en el impacto a largo plazo y en la capacidad local para sostener la herramienta. Para abordarlo, se debe incluir desde el inicio un plan de mantenimiento, formación continua para los usuarios locales, uso de tecnologías abiertas y escalables, y la integración con capacidades técnicas existentes en la región.

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. Redacción de experimentación y resultados, conclusiones, sugerencias de trabajos futuros e implicaciones éticas.
Izarra, L.: Síntesis de un trabajo relacionado. Contribución en la etapa temprana del desarrollo de la metodología. Redacción de experimentación y resultados, conclusiones, sugerencias de trabajos futuros e implicaciones éticas.
Mascco, P.: Obtención de fuentes académicas que sirven como guía en el modelo, redacción de la introducción y aporte de ideas en el código. Redacción de experimentación y resultados, conclusiones, sugerencias de trabajos futuros e implicaciones éticas.

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. pp. 1-8. https://ieeexplore.ieee.org/abstract/document/9632912