“Modelos Predictivos con Inteligencia Artificial para el Análisis Climático en Zonas de Geografía Compleja del Perú”

Autor/Autores:
Johnny Orihuela (20211000)
Joel Galindo Nuñez (20206222)
Fernando Ramírez (20201858)
Marcos Alonso Frisancho Machaca (20210702)
Joseph Tintaya (20201155)

Resumen

El presente trabajo busca el análisis de datos metereológicos en terrenos complejos de las regiones con terreno complejo del país (sea el caso de la sierra peruana), mediante el uso de herramientas de Machine Learning, con el objetivo de generar predicciones atmosféricas y realizar un análisis previos a futuras consecuencias, debido al cambio climático. Para ello, se recurrió a la investigación de artículos científicos que plantean una solución para lograr estos objetivos, generando una propuesta a raíz de estos y analizar su eficiencia con la realidad. Este trabajo, también buscará poder generalizarse, para que cualquier región del país también pueda predecir estas variables climáticas.

1. Introducción

El Perú presenta una geografía singular, caracterizado por 3 regiones principales: Costa, sierra y selva. Además de presentar 12 regiones naturales conocidas, clasificadas de acuerdo a su clima y altura. Este último también es un factor fundamental en el territorio. Ante ello, la precaución ante las consecuencias del cambio climático, conlleva un riesgo importante que varía enormemente de acuerdo a la región. Además, el mal manejo de estrategias de solución, han ocasionado daños aún mayores en algunas regiones. La autora Anna Heiken en su artículo [6] propone que las consecuencias, que sufren los agricultores en la sierra, también son influenciadas por prácticas sociopolíticas y un poco conocimiento crítico del estado climático. Con esto, el problema del cambio climático empeora, aún más, porque al tratar con un país pluricultural, estandarizar una solución y recopilar datos para predecir futuros comportamientos climáticos, se vería muy complejo y costoso, además del impacto indirecto del calentamiento global, siendo que el año pasado por primera vez se ha superado el umbral de 1,5 °C según el informe de la Organización Meteorológica Mundial (OMM) y el Servicio de Cambio Climático de Copernicus. Ante ello, la inteligencia artificial ha emergido como una herramienta capaz de extraer conocimientos profundos y descubrir patrones ocultos con relativa facilidad. La IA ofrece mejores predicciones climáticas, muestra los impactos de las condiciones meteorológicas extremas, identifica la fuente real de los emisores de carbono e incluye numerosas otras contribuciones razonables (Qin Xin, Ravi Samikannu y Chuliang Wei).

Ante esto se plantea los retos a tratar:
-Tener en cuenta las características del terreno, donde se quiere realizar la predicción
-Tener en cuenta la emisión de gases que causan las actividades humanas.
-Tener en cuenta influencias externas al cambio del clima.

Con esto se propone la siguiente pregunta:
“¿Se puede proponer un modelo predictivo con inteligencia artificial que se pueda estandarizar para todo el Perú?”

Para responder esta pregunta se plantean los siguientes objetivos:

- Buscar variables comunes, a través de análisis de estudios anteriores, que tengan un gran impacto en la predicción del clima.
- Buscar un modelo capaz de predecir variables que ayuden a la toma de decisiones.

2. Trabajos relacionados

La investigación sobre la aplicación de inteligencia artificial al análisis meteorológico en la región andina ha ganado un impulso significativo. Para contextualizar y fundamentar nuestro proyecto, se ha realizado una revisión y evaluación de tres trabajos clave que, en conjunto, establecen el

problema local, proponen una metodología base y demuestran el estado del arte.

En primer lugar, el trabajo de Flores Rojas (2020)[1] es fundamental, ya que se centra directamente en el Observatorio de Huancayo, nuestro caso de estudio. La investigación identifica eventos de lluvia intensa y los vincula a fenómenos de mayor escala como El Niño. Este artículo no solo justifica la relevancia de estudiar las variables de Huancayo, sino que nos proporciona una base científica para formular hipótesis sobre los impulsores climáticos locales. Sin embargo, su enfoque es estadístico y no utiliza técnicas de machine learning, dejando abierta la oportunidad de aplicar modelos predictivos más avanzados.

En segundo lugar, la tesis de Sedano Lazo (2023)[2] representa el antecedente más directo y relevante para nuestro proyecto. El autor desarrolla y compara varios modelos de redes neuronales recurrentes (LSTM, GRU, Bi-LSTM) para pronosticar la temperatura mínima diaria específicamente en el Observatorio de Huancayo. Su hallazgo de que el modelo Bi-LSTM es el más efectivo nos proporciona un sólido punto de partida y un benchmark con el cual comparar nuestros propios resultados. Este trabajo valida que los modelos de aprendizaje profundo son altamente efectivos para los datos de nuestra locación de estudio.

Finalmente, el estudio de Anochi y Shimizu (2025)[3] amplía el contexto a nivel continental, demostrando la superioridad de las redes neuronales para pronosticar la precipitación en toda Sudamérica. Aunque su enfoque es a una escala más amplia, su metodología para el tratamiento de datos de precipitación y el diseño de una red neuronal supervisada nos ofrece un marco metodológico robusto que podemos adaptar.

En síntesis, estos tres trabajos se complementan: Flores Rojas define el problema local, Sedano Lazo nos ofrece un modelo de IA de alto rendimiento probado en el mismo sitio, y Anochi & Shimizu validan el enfoque a una escala regional. Nuestro proyecto se sitúa en la intersección de estas investigaciones, buscando adaptar y extender estos enfoques para un análisis multivariado que integre tanto datos meteorológicos como de emisiones.

3. Metodología

- Enfoque(s) propuesto:

Con esta información se propone al análisis de datos para tratar de resolver la pregunta plantemos la formulación de las variables:

1) Para los datos que describen actividades humanas se usará el dataset de “Inventario Nacional de Emisiones de Gases de Efecto Invernadero, recuperado de: Inventario Nacional de Emisiones de Gases de Efecto Invernadero - [Ministerio del Ambiente - MINAM]”.[4]

2) Para datos metereológicos usaremos: “Dataset Datos de precipitación y temperatura del Centro de Investigación Científica y Tecnológica en Ecosistemas de Montaña (CICTEM) en Zona de monitoreo 1”.[5]

Paso 1: Manejo de las emisiones de actividad humana:

Se presenta el diccionario de variables que participarán en la metodología:

Variable	Descripción
FECHA_CORTE	Fecha de recolección de datos (yy/mm/dd)
ANIO	Año de emisión
SECTOR	Sector económico que se recolectó los datos
CATEGORIA	Categoría de la actividad realizada
SUBCATEGORIA	Subcategoría de la actividad realizada
FUENTE_DE_EMISION	Especificación de instante donde se recolectó la emisión de gases contaminantes
IPCC_CLASF	Clasificación según el IPCC de la actividad emisora
DIOXIDO_DE_CARBONO_GGCO2	Emisiones de Dióxido de Carbono (CO_2) expresadas en gigagramos (Gg)
METANO_GGCH4	Emisiones de Metano (CH_4) en gigagramos.
METANO_EQUIVALENTE_GGCO2EQ	Emisiones de Metano convertidas a su equivalente en CO_2 (CO_2eq), usando el Potencial de Calentamiento Global (GWP) del metano.
OXIDO_NITROSO_GGN2O	Emisiones de Óxido Nitroso (N_2O) en gigagramos.
OXIDO_NITROSO_EQUIVALENTE_GGCO2EQ	Emisiones de N_2O convertidas a su equivalente en CO_2, también usando su GWP.
HFC_GGCO2EQ	Emisiones de Hidrofluorocarbonos (HFCs) expresadas directamente en CO_2 equivalente.
EMISIONES_GEI_GGCO2EQ	Emisiones totales de Gases de Efecto Invernadero, sumando CO_2, CH_4, N_2O y HFCs, todas expresadas como CO_2 equivalente.

Tabla 1 Diccionario de datos del dataset 1.

Paso 2: Manejo de datos meteorológicos
Se presenta el diccionario de variables del dataset.

Campo	Descripción
FECHA_REGISTRO	Fecha en que se registraron los datos meteorológicos
PRECIPITACIÓN	Cantidad de precipitación registrada en milímetros.
TEMPERATURA_PROMEDIO	Temperatura promedio registrada durante el día.
TEMPERATURA_MINIMA	Temperatura mínima registrada durante el día.
TEMPERATURA_MAXIMA	Temperatura máxima registrada durante el día.
DEPARTAMENTO	Nombre del departamento donde se encuentra la estación meteorológica.
PROVINCIA	Nombre de la provincia donde se encuentra la estación.

DISTRITO	Nombre del distrito donde se encuentra la estación.
UBIGEO	Código UBIGEO del distrito (código geográfico nacional de Perú).
ESTACION	Código identificador de la estación meteorológica.
MARCA	Marca del equipo utilizado para el registro de datos.
MODELO	Modelo del equipo utilizado.
FECHA_CORTE	Fecha hasta la cual se tienen datos disponibles (fecha de corte del registro).

Tabla 2 Diccionario de datos del dataset 2.

Para ambos dataset, se haze el procesamiento de valores nulos, y eliminación de campos no requeridos (Dataset 1: Fecha_Corte, Anio, Sector, Categoría, Subcategoría; Dataset2: Fecha_Registro, Departamento, Distrito, Marca y Modelo).

Para el caso de Provincia, se realizará un encoding en base a 12, de acuerdo a la ubicación en las regiones climáticas, cambiando la columna a Región Climática.

Se seleccionó TEMPERATURA_PROMEDIO como variable objetivo del modelo predictivo por su estabilidad y representatividad del comportamiento térmico diario, a diferencia de los valores extremos. Es crucial para analizar tendencias de cambio climático y se correlaciona fuertemente con otras variables atmosféricas. Su relevancia es mayor en geografías complejas como la sierra peruana, donde influye en la agricultura y la gestión del agua. Finalmente, su disponibilidad y alta calidad de datos la hacen ideal para entrenar modelos de aprendizaje automático precisos y aplicables.

Para manejar ambos datasets, se utilizará la librería pandas de python.

Paso 3: Procesamiento de datos:
Para el primer dataset, se buscará identificar el modelo que mejor prediga las variables relacionadas con la emisión de contaminantes. Para ello, se aplicará un enfoque basado en redes neuronales, utilizando previamente una selección de variables mediante el algoritmo de Random Forest, siguiendo la metodología propuesta en la tesis de Alessandra Meza Lázaro. Se trabajará con datos de prueba y técnicas de muestreo, buscando optimizar los resultados mediante la minimización del error cuadrático medio (MSE):

$$mse = \frac{1}{n}\sum(y_n - \bar{y})^2$$

Con esto se busca optimizar el modelo de “Random Forest”. Una vez obtenidos el modelo, se pasa a la construcción de la red neuronal. Para la optimización del modelo de emisiones, se piensa usar valores similares a los propuestos en el trabajo de Alessandra Messa Lazaro.

Para establecer una línea base de comparación (baseline) con la cual evaluar el rendimiento de nuestro modelo de red neuronal, se estrenarán varios modelos de regresión más simples. El objetivo de estos modelos será predecir la TEMPERATURA_PROMEDIO utilizando las variables meteorológicas del segundo dataset. Se empleará una validación cruzada de 10 pliegues (k-fold cross-validation) para obtener una medida robusta del rendimiento de los siguientes algoritmos de prueba:

- Regresión Lineal(con Ridge Regression)
- Regresión Lineal(con LASSO)
- Regresión Lineal (con Elastic Net)
- K-Nearest Neighbour (K=5)
- Decision Tree

Estos modelos, implementados con la biblioteca Scikit-learn en Python, servirán como punto de referencia para determinar si el enfoque más complejo de la red neuronal ofrece una mejora significativa en la precisión del pronóstico.

Para reducir la probabilidad de tendencias del valor objetivo, se realizará una estandarización de los datos empleando pipelines, para evitar cualquier cambio de objetividad en los datos.

Paso 4: Predicción de variables meteorológicas

Inspirados en los resultados de la tesis de Sedano Lazo (2023), que demostró la alta efectividad de las redes neuronales recurrentes para datos de Huancayo, se construirá un modelo basado en Long Short-Term Memory (LSTM). Nuestra principal adaptación algorítmica consistirá en diseñar un modelo LSTM multivariado que recibirá como entrada (input) las características más importantes seleccionadas en el Paso 4 (incluyendo datos meteorológicos y de emisiones). El modelo será entrenado para minimizar el error cuadrático medio (MSE) y tendrá como salida (output) la predicción de la TEMPERATURA_PROMEDIO. Para mejorar el rendimiento, los datos de entrada serán estandarizados antes de alimentar el modelo.

Variables a considerar para la formulación de la red:

- Gas contaminante Seleccionado
- Ubicación Geográfica
- Precipitación
- Temperatura Promedio
- Región Climática

4. Experimentación y Resultados

- Setup experimental:
  - Describir datos usados (o método para obtenerlos) (si aplica).
  - Describir las métricas de evaluación.
  - Describir los experimentos hechos (qué componentes/parámetros/escenarios se probaron, qué valores, qué estrategia de validación).

Los experimentos deben ser planificados para poder caracterizar/comparar los enfoques desarrollados. Aquí algunas preguntas que deben responder los experimentos:

- ¿El enfoque desarrollado resuelve siempre el problema?
- ¿Qué tan eficientemente lo resuelven?
- ¿Cuál es el desempeño comparado con otros modelos o técnicas de referencia?
- ¿Cómo influyen los parámetros del enfoque en su desempeño?

- Resultados y Discusión:
  - Presentar resultados numéricos generados en los experimentos. Hacer un análisis de dichos resultados.

La presentación de los resultados debe facilitar el entendimiento. En general se deben usar figuras y/o tablas. Recuerde, todos los resultados deben interpretarse. Esforzarse para explicar el formato de las curvas presentadas, dar detalles del tiempo de simulación.

5. Conclusión

Dar las conclusiones principales con base en los resultados obtenidos y a lo que fue planteado en su hipótesis, ¿qué se puede decir del o los enfoques desarrollados y/o del problema abordado?

6. Sugerencias de trabajos futuros

Indicar, por ejemplo: ¿qué cosas se pueden mejorar del enfoque?, ¿qué otros posibles problemas podrían abordarse con el enfoque?

7. Implicancias éticas

Indicar qué implicancias éticas podría generar el trabajo desarrollado de ser escalado (sesgos, posible afectación a la seguridad/privacidad de usuarios, posibilidad de ataques y robo de datos, etc). Sugiera formas de abordar dichos problemas

8. Link del repositorio del trabajo

Puede ser Github, Gitlab, u otro. Dar las credenciales para poder tener acceso.

9. Declaración de contribución de cada integrante

Describir los aportes de cada integrante al proyecto.

10. Referencias

[1]. ANOCHI, J. A. y M. H. SHIMIZU 2025 “Precipitation Forecasting and Drought Monitoring in South America Using a Machine Learning Approach”. Basilea, volumen 4, número 1. Consulta: 1 de junio de 2025 .https://www.mdpi.com/2674-0494/4/1/1

[2] FLORES, J. L. 2020 “Identificación de eventos de lluvia intensa en el Observatorio de Huancayo, Valle del Mantaro”. Boletín científico El Niño. Lima, volumen 7, número 5, pp. 4-13. Consulta: 1 de junio de 2025. http://hdl.handle.net/20.500.12816/5145

[3] MESA LÁZARO, Alessandra 2024 Aplicación de redes neuronales para estimar la emisión de gases contaminantes con relación al consumo de gasolina de vehículos livianos circulando en Lima Metropolitana en el 2022 . Consulta: 1 de junio de 2025. https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/24759

[4] *MINISTERIO DEL AMBIENTE (MINAM)* 2024. *Inventario Nacional de Emisiones de Gases de Efecto Invernadero*. Lima: Plataforma Nacional de Datos Abiertos. Consulta: 2 de junio de 2025. https://datosabiertos.gob.pe/dataset/inventario-nacional-de-emisiones-de-gases-de-efecto-invernadero-ministerio-del-ambiente-0

[5] INSTITUTO NACIONAL DE INVESTIGACIÓN EN GLACIARES Y ECOSISTEMAS DE MONTAÑA (INAIGEM) 2024. Datos de precipitación y temperatura del Centro de Investigación Científica y Tecnológica en Ecosistemas de Montaña (CICTEM) en Zona de Monitoreo 1. Plataforma Nacional de Datos Abiertos. Consulta: 2 de junio de 2025. https://datosabiertos.gob.pe/dataset/datos-de-precipitaci%C3%B3n-y-temperatura-del-

centro-de-investigaci%C3%B3n-cient%C3%ADfica-y-tecnol%C3%B3gica-2

[6] HEIKKINEN, A. 2023. Cambio climático, poder y vulnerabilidades en la sierra peruana. Allpanchis, año 50, número 91, pp. 111–157. Consulta: 6 de junio de 2025. https://dialnet.unirioja.es/servlet/articulo?codigo=9124512

Nota importante:

Figuras y tablas

- Cada figura debe ser leída y comprendida sin necesidad de recurrir a la descripción a lo largo del texto.
- Todos los símbolos usados deben ser explicados en la leyenda.