UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

APRENDIZAJE AUTOMÁTICO

Predicción de flujo de calor del suelo usando variables meteorológicas

DOCENTE: Javier Arturo Rozas Huacho

INTEGRANTES:

Aguilar Mainicta Gian Marco (174905)

Alegria Sallo Daniel Rodrigo (215270)

Muñoz Centeno Milder (211860)

Perú

Junio 2025

Tabla de Contenido

1. Introducción del problema

El monitoreo de variables meteorológicas en ecosistemas altoandinos es fundamental para comprender los procesos físicos que ocurren en la superficie terrestre y su interacción con la atmósfera. En este proyecto se analiza un conjunto de datos proveniente de la Torre de Gradiente del Instituto Geofísico del Perú (IGP), ubicada en LAMAR, Junín, a más de 3300 m.s.n.m. Esta torre recopila información meteorológica con una alta resolución temporal (cada minuto), incluyendo temperatura, humedad relativa, velocidad y dirección del viento, así como flujos de calor, lo cual ofrece una visión detallada del comportamiento climático en esta región.

El objetivo principal del proyecto es desarrollar un modelo de aprendizaje automático de tipo regresión que permita predecir el flujo de calor del suelo `soil_heat` a partir de las variables meteorológicas registradas. El flujo de calor del suelo es una variable clave en el balance energético de la superficie, con implicancias directas en procesos como la evaporación, el crecimiento vegetal, la productividad agrícola y la dinámica hídrica.

1.1. Justificación del problema abordado

Desde un enfoque técnico, la predicción del flujo de calor del suelo utilizando modelos de aprendizaje automático representa un desafío interesante debido a la naturaleza multivariada, correlacionada y altamente dinámica de los datos atmosféricos. Comparar distintos algoritmos permite identificar cuál se ajusta mejor al problema, en términos de precisión, eficiencia y capacidad de generalización. Este tipo de solución basada en datos puede contribuir al desarrollo de herramientas de pronóstico ambiental más precisas, adaptadas al contexto geográfico y climático del Perú.

Desde una perspectiva social, la mejora en la comprensión y predicción de variables como el flujo de calor del suelo tiene un impacto significativo en la agricultura andina, una actividad fuertemente dependiente del clima. En regiones como Junín, donde muchas comunidades rurales basan su subsistencia en la agricultura familiar, disponer de modelos predictivos puede ayudar en la toma de decisiones relacionadas con el riego, la siembra y la cosecha. Además, en un contexto de cambio climático, contar con herramientas que permitan anticipar comportamientos extremos o cambios en el patrón térmico del suelo es crucial para la gestión sostenible del territorio y los recursos naturales.

Por tanto, este proyecto se sitúa en la intersección entre la ciencia de datos, la meteorología aplicada y el desarrollo sostenible, proponiendo una solución técnica de valor tangible para uno de los sectores clave del país.

2. Descripción del dataset

El dataset utilizado corresponde a mediciones de la estación Torre de Gradiente del Laboratorio de Micro Física Atmosférica y Radiación (LAMAR) del Instituto Geofísico del Perú (IGP), ubicada en Junín, Perú (-12.0399°S, −75.3207°W, 3316.78 m.s.n.m.). El dataset comprende registros meteorológicos desde mayo de 2018 hasta marzo de 2025.

El dataset contiene mediciones de una torre meteorológica de 30 metros de altura con sensores distribuidos en diferentes niveles (2, 6, 12, 18, 24 y 29 metros) que registran las

principales variables atmosféricas con resolución temporal de un minuto. Las mediciones incluyen temperatura y humedad relativa mediante sondas HMP60 de Campbell Scientific, así como velocidad y dirección del viento utilizando conjuntos Wind Sentry 03002.

2.1. Estructura del dataset

El dataset se compone de **25 columnas** principales que incluyen:

- **Variables temporales**: FECHA_CORTE, UBIGEO, year, month, day, hour
- **Temperatura del aire**: temp_n1 a temp_n6 (mediciones en 6 niveles, °C)
- **Velocidad del viento**: wind_n1 a wind_n6 (mediciones en 6 niveles, m/s)
- **Humedad relativa**: RH_n1 a RH_n6 (mediciones en 6 niveles, %)
- **Dirección del viento**: dir_wind_01 y dir_wind_02 (18m y 29m, grados)
- **Variable objetivo**: soil_heat (flujo de calor del suelo a 8 cm, W/m²)

Todas las variables numéricas se almacenan con precisión de cinco decimales, con valores enteros positivos para viento y humedad, y posibles valores negativos para temperatura y flujo de calor del suelo.

3. Formulación del problema

El problema se formula como una tarea de **regresión** donde el objetivo es predecir el flujo de calor del suelo `soil_heat` utilizando las variables meteorológicas medidas en diferentes niveles de la torre.

3.1. Definición del objetivo

El modelo busca establecer relaciones cuantitativas entre las condiciones atmosféricas (temperatura, humedad, velocidad y dirección del viento) y el intercambio de calor entre la superficie terrestre y la atmósfera. Esta predicción es fundamental para:

- Estudios de balance energético en ecosistemas de alta montaña
- Análisis de procesos de intercambio superficie-atmósfera
- Estimación de flujos de calor sensible y latente
- Modelado microclimático en regiones andinas

La variable objetivo `soil_heat` representa el flujo de calor del suelo medido a 8 cm de profundidad, expresado en W/m², donde valores positivos indican transferencia de calor desde el suelo hacia la atmósfera y valores negativos representan el flujo inverso.

% Análisis exploratorio de datos con apoyo de visualizaciones. % Limpieza y preparación del dataset. % Aplicación de técnicas básicas de ingeniería de características.

4. Metodología de preprocesamiento

4.1. Análisis exploratorio de datos

El análisis exploratorio del dataset de **54,803 registros** reveló las siguientes características:

- **Dimensiones**: 54,803 filas × 27 columnas (11.29 MB)
- **Período temporal**: 2018-2025 con distribución temporal continua

- **Variable objetivo**: soil_heat con rango de −4,000.97 a 4,132.34 W/m²
- **Valores faltantes críticos**: 25.31% en soil_heat (13,870 registros)

Distribución de valores faltantes

El análisis identificó patrones específicos de datos faltantes:

Variable	Valores Faltantes	Porcentaje
soil_heat	13,870	25.31%
temp_n2-n6	1,395	2.55%
RH_n2-n6	705	1.29%

Tabla 1: Distribución de valores faltantes por grupo de variables

Las visualizaciones implementadas incluyen análisis completo de:

- Distribución estadística de la variable objetivo (soil_heat)
- Series temporales multi-nivel de temperatura, viento y humedad
- Matriz de correlación de 27 variables meteorológicas
- Patrones diurnos y estacionales del flujo de calor
- Análisis de dispersión entre variables predictoras y objetivo

4.2. Limpieza y preparación del dataset

El proceso de preprocesamiento se decidio **ignorar** elementos de valores faltantes. Ademas de realizar:

- **Construcción de índice temporal**: Timestamps precisos usando year-month-day--hour
- **Ordenamiento cronológico**: Organización secuencial para interpolación temporal
- **Interpolación temporal**: Método «time» para aprovechat continuidad temporal
- **Imputación residual**: SimpleImputer con estrategia de mediana para valores restantes
- **Normalización**: StandardScaler aplicado a 35 características finales

Resultado del preprocesamiento: **0 valores faltantes** en todas las 27 variables.

4.3. Ingeniería de características

Se generaron **35 características** mediante técnicas avanzadas:

- **Codificación cíclica temporal** (4 características):
  - hour_sin, hour_cos: Captura periodicidad diaria
  - month_sin, month_cos: Captura estacionalidad anual
- **Gradientes verticales de temperatura** (5 características):
  - temp_gradient_1_2 hasta temp_gradient_5_6
  - Diferencias entre niveles consecutivos (2m-6m, 6m-12m, etc.)

- **Estadísticas agregadas multi-nivel** (6 características):
  - temp_mean, temp_std: Estadísticas de temperatura por timestamp
  - wind_mean, wind_std: Estadísticas de velocidad de viento
  - rh_mean, rh_std: Estadísticas de humedad relativa
- **Variables originales**: 20 variables meteorológicas originales

4.3.1. División y normalización

- **Conjunto de entrenamiento**: 43,842 muestras (80%)
- **Conjunto de prueba**: 10,961 muestras (20%)
- **Estrategia**: División aleatoria con seed=42
- **Normalización**: StandardScaler ajustado solo en entrenamiento

5. Modelo entrenado y resultados preliminares

5.1. Modelos implementados

Se entrenaron tres modelos de aprendizaje automático con los siguientes parámetros:

- **Regresión Lineal**: Modelo baseline con regularización estándar
- **Random Forest**: Ensamble de 100 árboles de decisión con paralelización
- **Gradient Boosting**: Boosting secuencial con 100 estimadores

5.2. División de datos y estrategia de validación

- **Entrenamiento**: 43,842 muestras (80%)
- **Prueba**: 10,961 muestras (20%)
- **Estrategia**: División aleatoria estratificada (seed=42)
- **Características**: 35 variables predictoras procesadas

5.3. Métricas de evaluación

Para evaluar el rendimiento de los modelos de regresión se utilizaron:

- **R^2 (Coeficiente de determinación)**: Proporción de varianza explicada
- **MSE (Error cuadrático medio)**: Penalización cuadrática de errores grandes
- **MAE (Error absoluto medio)**: Medida robusta de error promedio

5.4. Resultados experimentales

Los modelos mostraron el siguiente rendimiento diferenciado:

Modelo	R^2 Train	R^2 Test	MSE Train	MSE Test	MAE Train	MAE Test
Regresión Lineal	0.2460	0.2187	640,539	656,706	548.18	545.68
Random Forest	**0.9375**	**0.5588**	53.124	370.880	112.42	298.30
Gradient Boosting	0.3379	0.3229	562.h21	569.143	504.24	509.49

Tabla 2: Métricas de rendimiento comparativo de los tres modelos

5.5. Análisis de resultados

5.5.1. Modelo ganador: Random Forest

Random Forest emergió como el modelo superior con:

- **R^2 = 0.5588**: Explica 55.88% de la variabilidad del flujo de calor
- **MSE = 370,880 W^2/m^4**: Error cuadrático moderado
- **MAE = 298.30 W/m^2**: Error absoluto promedio aceptable

5.5.2. Detección de sobreajuste

El análisis revela **sobreajuste moderado en Random Forest**:

- R^2 Train $(0.9375) \gg R^2$ Test (0.5588)
- Diferencia de 0.38 puntos indica memorización de patrones de entrenamiento
- MSE aumenta significativamente de entrenamiento (53K) a prueba (371K)

5.5.3. Rendimiento comparativo

- **Regresión Lineal**: Rendimiento consistente pero limitado ($R^2 \approx 0.22$)
- **Gradient Boosting**: Mejor generalización que Random Forest ($R^2 = 0.32$)
- **Random Forest**: Mayor capacidad predictiva pero con sobreajuste

Las visualizaciones implementadas permiten evaluar:

- **Predicciones vs. valores reales**: Correlación visual del modelo ganador
- **Análisis de residuos**: Detección de heterocedasticidad y patrones
- **Distribución de errores**: Verificación de normalidad en residuos
- **Importancia de características**: Ranking de variables meteorológicas influyentes

5.5.4. Interpretación de resultados

El **R^2 = 0.5588** indica que el modelo Random Forest captura efectivamente:

- Patrones de intercambio energético superficie-atmósfera
- Relaciones no-lineales entre variables meteorológicas
- Efectos de gradientes verticales de temperatura y viento
- Ciclos temporales diurnos y estacionales

El error absoluto medio de **298.30 W/m^2** es razonable considerando:

- Rango de soil_heat: [−4,000, +4,132] W/m^2 (8,132 W/m^2 total)
- Error relativo: $\sim 7.3\%$ del rango total
- Variabilidad natural alta en flujos de calor en ecosistemas montanos

6. Conclusiones parciales y próximos pasos

Durante esta primera etapa del proyecto se ha completado con éxito el análisis exploratorio, preprocesamiento y entrenamiento de modelos para abordar el problema de regresión orientado a la predicción del flujo de calor del suelo (**`soil_heat`**) utilizando datos meteorológicos obtenidos de la Torre de Gradiente del IGP en Junín, Perú.

El modelo con mejor desempeño lo obtuvo el modelo **Random Forest**, con un R^2 en prueba de **0.5588**, una reducción significativa del *error cuadrático medio (MSE)* en comparación con los otros modelos y un *MAE* de 298.30, lo cual indica una mejor capacidad de ajuste sin sobreajuste severo.

Próximos pasos

1. **Optimizar hiperparámetros** del modelo Random Forest con **GridSearchCV** o **RandomizedSearchCV**.
2. **Explorar modelos adicionales** como XGBoost, LightGBM o redes neuronales para comparar su rendimiento.
3. **Aplicar validación cruzada k-fold** para una evaluación más robusta del rendimiento.
4. **Analizar la importancia de características** del modelo para identificar qué variables tienen mayor influencia sobre el flujo de calor del suelo.
5. **Desarrollar visualizaciones interactivas** y/o un prototipo de sistema que permita visualizar predicciones en tiempo real con base en los datos atmosféricos.
6. Documentar todo el *pipeline* para facilitar la reproducibilidad del análisis.

Anexos

A Repositorio del Proyecto

https://github.com/daniel-alegria3/Proyecto_ML

B Fuente del Dataset

https://www.datosabiertos.gob.pe/dataset/dataset-de-la-estaci%C3%B3n-torre-de-gradiente--para-estimar-los-flujos-de-calor-sensible-y-calor