Universidad de La Habana

Facultad de Matemática y Computación

Licenciatura en Ciencias de la Computación

Procesamiento de Grandes Volúmenes de Datos. Análisis climático y modelado de eventos extremos

---

Autor: Eveliz Espinaco Milian

Año académico: 4to año

Fecha: 27 de diciembre de 2025

Índice

1. Introducción

El cambio climático representa uno de los desafíos más críticos del siglo XXI, caracterizado por patrones meteorológicos cada vez más impredecibles y eventos extremos de mayor intensidad. La capacidad de predecir y alertar sobre olas de calor, heladas, tormentas intensas y otros fenómenos climáticos extremos es fundamental para proteger vidas, infraestructura crítica y ecosistemas. Sin embargo, la magnitud de datos históricos disponibles (millones de registros de temperatura) excede la capacidad de procesamiento de herramientas convencionales, evidenciando la necesidad de arquitecturas distribuidas y escalables.

1.1. Enfoque Tecnológico: Big Data para Análisis Climático

Este proyecto adopta un enfoque integral de Big Data para construir un sistema de predicción y alerta de eventos climáticos extremos. La arquitectura se fundamenta en tecnologías de código abierto líderes en la industria:

- **HDFS (Hadoop Distributed File System):** Almacenamiento distribuido y tolerante a fallos de grandes volúmenes de datos climáticos históricos.
- **YARN (Yet Another Resource Negotiator):** Gestión centralizada de recursos del clúster, permitiendo asignación eficiente de procesamiento entre múltiples trabajos.
- **Apache Spark:** Motor de procesamiento distribuido in-memory que acelera análisis, transformaciones y entrenamientos de modelos ML sobre datos en HDFS.
- **Hive:** Capa de abstracción SQL para consultas declarativas sobre datos almacenados, facilitando acceso a usuarios sin expertise en programación distribuida.

1.2. Problemas Abordados

El proyecto plantea y resuelve dos tareas complementarias:

1. **Predicción de Temperatura:** Estimar temperatura media para períodos futuros (semanas/meses) basándose en series históricas de 12+ meses y variables geográficas, permitiendo identificación temprana de tendencias anómalas.
2. **Detección de Eventos Extremos:** Clasificar si un período experimental presentará condiciones climáticas extremas (olas de calor, heladas intensas, desviaciones significativas de la norma histórica).

1.3. Capacidades Analíticas y Visualización

El sistema entrega capacidades de análisis y visualización en tiempo real:

- **Mapas de calor climáticos:** Visualización geoespacial de anomalías de temperatura, identificando zonas críticas de calentamiento o enfriamiento anómalo.

- **Series temporales interactivas:** Evolución histórica de temperatura global y regional, con tendencias suavizadas y bandas de incertidumbre.
- **Predicciones probabilísticas:** Cuantificación de confianza en predicciones (ej. "probabilidad de ola de calor: 72 %").
- **Alertas activas:** Notificaciones automáticas cuando probabilidad de evento extremo supera umbrales operacionales, con justificación contextual.
- **Comparación histórica:** Contrastación de eventos actuales contra análogos históricos, contextualizando severidad relativa.

1.4. Dataset y Metodología

El análisis se construye sobre un dataset masivo de 8.6 millones de registros de temperatura de superficie terrestre, abarcando 173 años (1850-2023) y 7,280+ estaciones meteorológicas globales. El pipeline de procesamiento comprende:

1. **Ingesta y particionamiento:** Transferencia escalable de datos CSV a HDFS, particionamiento por década para optimización de consultas.
2. **Limpieza y normalización:** Tratamiento de valores faltantes mediante imputación temporal, detección de anomalías, normalización de rangos.
3. **Ingeniería de features:** Generación de 20+ variables derivadas (anomalías, volatilidad, tendencias, variables geográficas) a partir de atributos brutos.
4. **Modelado ML:** Entrenamiento de regresores (Random Forest, Ridge) para predicción de temperatura y clasificadores (Regresión Logística, GBM) para detección de extremos.
5. **Evaluación y deployment:** Validación temporal de modelos, análisis de importancia de features, serialización para predicción en producción.
6. **Visualización interactiva:** Dashboard con Plotly Dash, permitiendo exploración de datos y consulta de predicciones por usuarios técnicos y no-técnicos.

1.5. Objetivos Específicos del Proyecto

a Diseñar e implementar pipeline robusto de ingesta, limpieza y transformación de datos climáticos a escala de millones de registros en entorno distribuido.

b Desarrollar modelos predictivos de temperatura y clasificadores de eventos extremos con exactitud demostrada en conjunto de validación temporal.

c Crear dashboard interactivo que comunique insights del análisis a múltiples audiencias (público general, meteorólogos, investigadores).

d Documentar metodología y resultados con rigor académico, permitiendo reproducibilidad y extensión futura del sistema.

2. Sobre el Dataset

En esta sección se describe el conjunto de datos elegido para el desarrollo del proyecto.

2.1. Nombre, Fuente y Formato

- **Nombre:** Cambio climático: datos de temperatura de la superficie terrestre
- **Fuente:** https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data
- **Formato:** CSV

2.2. Variables Estadísticas

A continuación, se describen las variables incluidas en el dataset, detallando su significado, los valores que pueden tomar, y su clasificación correspondiente.

- **Fecha:** Indica el año de la medición. Los valores comienzan en 1750 para la temperatura media de la tierra y en 1850 para las temperaturas máximas y mínimas, así como para las temperaturas globales de océanos y tierra.
  Valores: Años (1750 en adelante).
  Tipo: Variable cuantitativa, escala ordinal.
- **País:** País donde se realizó la medición.
  Valores: Nombres de países (por ejemplo, España, México, Argentina).
  Tipo: Variable cualitativa, escala nominal.
- **Ciudad:** Ciudad específica de la estación meteorológica.
  Valores: Nombres de ciudades (por ejemplo, Madrid, Buenos Aires).
  Tipo: Variable cualitativa, escala nominal.
- **Latitud:** Coordenada geográfica norte-sur de la estación.
  Valores: Números reales en grados decimales (por ejemplo, 40.4168).
  Tipo: Variable cuantitativa, escala de razón.
- **Longitud:** Coordenada geográfica este-oeste de la estación.
  Valores: Números reales en grados decimales (por ejemplo, -3.7038).
  Tipo: Variable cuantitativa, escala de razón.
- **Temperatura media del terreno:** Representa la temperatura media global de la superficie terrestre, expresada en grados Celsius.
  Valores: Números reales (por ejemplo, 13.5°C).
  Tipo: Variable cuantitativa, escala de razón.
- **Incertidumbre de la temperatura media del terreno:** Intervalo de confianza del 95 % alrededor del promedio de la temperatura media del terreno.
  Valores: Números reales positivos (por ejemplo, 0.12°C).
  Tipo: Variable cuantitativa, escala de razón.

2.3. Justificación del Dataset

El dataset seleccionado contiene registros históricos de temperatura de la superficie terrestre, recolectados a lo largo de varias décadas mediante distintos métodos e instrumentos. Su uso está justificado por la complejidad inherente a los datos, que exige una gran cantidad de limpieza, normalización y procesamiento para poder extraer conclusiones válidas sobre las tendencias climáticas a largo plazo. Los primeros registros fueron obtenidos por técnicos que utilizaban termómetros de mercurio, donde incluso pequeñas variaciones en la hora de la medición podían alterar significativamente los valores registrados. Posteriormente, en la década de 1940, la construcción de aeropuertos obligó al traslado físico de muchas estaciones meteorológicas, introduciendo discontinuidades espaciales en las series de datos. Más adelante, en los años 80, se incorporaron termómetros electrónicos, los cuales presentan un sesgo sistemático de enfriamiento que debe ser corregido para garantizar la coherencia del análisis.

Este contexto histórico y técnico convierte al dataset en un excelente candidato para evaluar los conocimientos adquiridos en la asignatura de Procesamiento de Grandes Volúmenes de Datos. A pesar de que el volumen original es relativamente pequeño, la riqueza estructural, la heterogeneidad temporal y la presencia de sesgos lo hacen ideal para simular escenarios reales de Big Data, donde la calidad y la gobernanza de los datos son tan importantes como la cantidad.

2.3.1. Volumen

El dataset cuenta con aproximadamente 8.6 millones de registros, lo que lo convierte en un ejemplo representativo de escenarios reales de Big Data. Este volumen masivo permite aplicar técnicas avanzadas de procesamiento distribuido, como particionamiento, replicación y procesamiento paralelo, utilizando herramientas como Hadoop o Spark. La gran cantidad de datos facilita la segmentación por décadas, estaciones meteorológicas, países, o tipo de sensor, permitiendo diseñar esquemas de particionamiento que reflejan los principios de escalabilidad horizontal. Además, el tamaño del dataset posibilita la realización de análisis estadísticos robustos, la detección de patrones complejos y la generación de modelos predictivos con mayor precisión. La integración con fuentes externas (como altitud, ubicación geográfica, eventos históricos) y la generación de datos derivados amplían aún más las posibilidades de exploración y simulación en entornos de procesamiento masivo.

2.3.2. Características

Los atributos presentes en el dataset incluyen fecha de medición, ubicación geográfica, tipo de instrumento utilizado, temperatura registrada, y metadatos asociados a la estación meteorológica. La temporalidad es extensa, abarcando desde principios del siglo XX hasta la actualidad, lo que permite estudiar fenómenos de largo plazo como el cambio climático, la variabilidad estacional, y los efectos de urbanización. El dataset no está etiquetado en el sentido clásico de aprendizaje supervisado, pero permite generar etiquetas derivadas (por ejemplo, anomalías térmicas, zonas de cambio abrupto, o eventos extremos) mediante procesamiento. El nivel de ruido es alto: hay inconsistencias en las unidades, valores faltantes, duplicados, y sesgos sistemáticos que deben ser corregidos. Esta situación obliga a aplicar técnicas de limpieza, imputación,

normalización y reconciliación de fuentes, lo cual es central en el estudio de grandes volúmenes de datos.

2.3.3. Pertinencia

El dataset se relaciona directamente con los objetivos de la asignatura, ya que permite aplicar de forma integrada todos los conceptos clave: ingestión de datos desde múltiples fuentes, limpieza intensiva, transformación distribuida, almacenamiento optimizado, y análisis escalable. Además, su temática —las tendencias climáticas— es de alta relevancia social y científica, lo que motiva el trabajo y permite conectar la teoría con problemas reales. El proyecto puede incluir la simulación de un Data Lake con zonas raw, curated y analytics; el uso de herramientas como Apache Spark para agregaciones por década; y la visualización de resultados mediante dashboards interactivos. En conjunto, el dataset ofrece una oportunidad única para evaluar competencias técnicas, analíticas y metodológicas en un entorno controlado pero representativo de los desafíos del procesamiento de grandes volúmenes de datos.

3. Preparación y Carga de Datos

3.1. Objetivo

Transferir y organizar datos CSV en HDFS de forma eficiente, garantizando escalabilidad, tolerancia a fallos y acceso optimizado para posteriores etapas de procesamiento. Esta fase es crítica en pipelines de Big Data, ya que sienta las bases para todo el análisis downstream.

3.2. Estrategia de Particionamiento

3.2.1. Razones para particionar

El dataset original de 8.6 millones de registros (aproximadamente 600 MB en formato CSV) debe dividirse en bloques para optimizar:

a **Paralelismo:** Permite que múltiples nodos del clúster procesen diferentes particiones simultáneamente.

b **Tolerancia a fallos:** Si un nodo falla durante la transferencia, solo se debe reintentar la partición afectada.

c **Replicación eficiente:** HDFS replica automáticamente cada bloque en múltiples nodos (factor de replicación predeterminado = 3).

d **Acceso distribuido:** Las consultas se distribuyen entre nodos, reduciendo latencia y mejorando throughput.

3.2.2. Estrategia de división

Se divide el archivo CSV original en chunks de **128 MB** (tamaño estándar de bloque en HDFS):

```
# Calcular numero de chunks
# Tamano archivo: ~600 MB
# Chunks = 600 / 128 aproximadamente 5 archivos

$ split -b 128M temperatura_global.csv chunk_
```

Esto genera: `chunk_aa`, `chunk_ab`, `chunk_ac`, etc.

3.3. Transferencia a HDFS

3.3.1. Método 1: Comandos HDFS nativos

```
# Crear directorio en HDFS
$ hdfs dfs -mkdir -p /climate_data/raw/

# Transferir archivo completo
$ hdfs dfs -put temperatura_global.csv /climate_data/raw/

```

```
# Transferir chunks en paralelo
$ for chunk in chunk_*; do
    hdfs dfs -put $chunk /climate_data/raw/ &
  done
$ wait  # Esperar a que todas las transferencias terminen

# Verificar carga
$ hdfs dfs -ls -lh /climate_data/raw/
$ hdfs dfs -du -sh /climate_data/raw/
```

Consideraciones:

- `-put` es bloqueante; usar `&` para paralelización
- El comando `-du -sh` muestra tamaño total ocupado
- HDFS replica automáticamente (3 copias por defecto)

3.3.2. Método 2: Ingesta con PySpark (recomendado)

PySpark permite ingesta escalable directamente a formato Parquet:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("ClimateDataIngest") \
    .config("spark.sql.shuffle.partitions", "200") \
    .getOrCreate()

# Leer CSV con inferencia de tipos
df = spark.read \
    .option("header", "true") \
    .option("inferSchema", "true") \
    .option("mode", "DROPMALFORMED") \
    .csv("/local/path/temperatura_global.csv")

# Mostrar estadisticas
print(f"Total de registros: {df.count()}")
print(f"Esquema: {df.printSchema()}")

# Escribir en HDFS como Parquet particionado anual
df.write \
    .mode("overwrite") \
    .partitionBy("year") \
    .parquet("hdfs:///climate_data/raw_parquet/")

print("Ingesta completada exitosamente")
```

Ventajas sobre CSV en HDFS:

- **Compresión nativa:** Reduce almacenamiento en 80 %
- **Lectura columnar:** Solo se leen columnas necesarias

- **Particionamiento automático:** Facilita pruning
- **Compatibilidad:** Hive, Spark y Presto pueden leerlo directamente

3.4. Validación de Integridad

3.4.1. Verificación de completitud

```
# Contar registros en origen y destino
$ wc -l temperatura_global.csv  # Origen local
$ hdfs dfs -cat /climate_data/raw/temperatura_global.csv \
    | wc -l  # Destino HDFS

# Calcular checksum para detectar corrupcion
$ md5sum temperatura_global.csv > checksum.txt
$ hdfs dfs -get /climate_data/raw/temperatura_global.csv
$ md5sum temperatura_global.csv >> checksum.txt
```

3.4.2. Validación de formato

```
# Validar esquema y tipos de datos
df = spark.read.parquet("hdfs:///climate_data/raw_parquet/")

# Validar que no hay nulos criticos
df.select([
    count(when(col(c).isNull(), 1)).alias(f"{c}_null_count")
    for c in ["year", "country", "temperature"]
]).show()

# Estadisticas descriptivas
df.select("temperature").describe().show()
```

3.5. Creación de Tabla Hive

3.5.1. Tabla externa sobre datos raw

Hive permite consultas SQL directas sobre archivos en HDFS:

```
-- Crear tabla externa apuntando a datos en HDFS
CREATE EXTERNAL TABLE IF NOT EXISTS climate_data_raw (
    dt STRING,
    AverageTemperature DOUBLE,
    AverageTemperatureUncertainty DOUBLE,
    City STRING,
    Country STRING,
    Latitude DOUBLE,
    Longitude DOUBLE
)
ROW FORMAT DELIMITED
```

```
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION 'hdfs:///climate_data/raw/'
TBLPROPERTIES ("skip.header.line.count"="1");

-- Verificar carga
SELECT COUNT(*) AS total_records FROM climate_data_raw;

-- Consulta de muestra
SELECT City, Country, AverageTemperature, dt
FROM climate_data_raw
WHERE Country = 'Cuba'
LIMIT 10;
```

3.5.2. Tabla nativa (formato optimizado)

Para mejor rendimiento, crear tabla en Parquet:

```
-- Crear tabla nativa desde Parquet
CREATE TABLE climate_data_optimized (
    year INT,
    month INT,
    avg_temperature DOUBLE,
    uncertainty DOUBLE,
    city STRING,
    country STRING,
    latitude DOUBLE,
    longitude DOUBLE
)
USING PARQUET
PARTITIONED BY (year)
LOCATION 'hdfs:///climate_data/raw_parquet/';

-- Cargar particiones
MSCK REPAIR TABLE climate_data_optimized;

-- Consultas de ejemplo
SELECT DISTINCT country FROM climate_data_optimized
WHERE year >= 2000;

SELECT country, AVG(avg_temperature) as temp_promedio
FROM climate_data_optimized
WHERE year >= 2010
GROUP BY country;
```

3.5.3. Ventajas de tablas Hive

1. **Interfaz SQL:** Familiar para analistas sin conocimiento de Spark/MapReduce
2. **Metastore:** Metadatos centralizados y reutilizables

3. **Interoperabilidad:** Acceso desde Spark, Presto, Impala
4. **Particionamiento:** Pruning automático en consultas
5. **Estadísticas:** Optimizador usa estadísticas para plans eficientes

3.6. Monitoreo y Reporte

3.6.1. Comandos de inspección HDFS

```
# Reporte de uso de almacenamiento
$ hdfs dfs -du -sh /climate_data/

# Verificar factor de replicacion
$ hdfs dfs -stat "%r" /climate_data/raw_parquet/year=2020/*

# Detectar bloques bajo-replicados
$ hdfs fsck /climate_data/ -list-corruptfileblocks

# Balance del cluster
$ hdfs balancer -threshold 10
```

3.6.2. Resumen de carga

Crear reporte post-ingesta:

```
def generate_ingest_report(hdfs_path):
    """Generar reporte de ingesta"""
    df = spark.read.parquet(hdfs_path)

    report = {
        "total_records": df.count(),
        "schema": df.schema.json(),
        "partitions": df.rdd.getNumPartitions(),
        "null_counts": {col: df.filter(col(col).isNull()).count
            ()
                        for col in df.columns},
        "temp_stats": df.select("temperature").describe().
            collect()
    }

    return report

report = generate_ingest_report("hdfs:///climate_data/
    raw_parquet/")
print(json.dumps(report, indent=2))
```

3.7. Conclusión de fase

La preparación y carga de datos establece la infraestructura necesaria para procesamiento distribuido. El uso de HDFS garantiza escalabilidad, mientras que Parquet y tablas Hive optimizan el acceso en etapas posteriores.

4. Limpieza y Normalización de Datos

4.1. Objetivo

Preparar datos de calidad para análisis posterior, eliminando inconsistencias, valores faltantes y anomalías que afecten la confiabilidad de modelos predictivos. Esta fase es fundamental en proyectos de Big Data, ya que la calidad de los datos determina directamente la validez de conclusiones y predicciones.

4.2. Exploración Inicial de Datos

4.2.1. Identificación de tipos y rangos

La exploración inicial consiste en caracterizar cada variable del dataset: su tipo de dato, rango de valores, distribución y anomalías detectadas. En el contexto del dataset de temperatura, las variables numéricas (temperatura media, incertidumbre) deben presentar rangos físicamente coherentes (por ejemplo, temperaturas globales entre -50°C y 50°C). Las variables categóricas (país, ciudad) deben validarse contra diccionarios geográficos para identificar inconsistencias de nomenclatura.

Durante esta etapa se generan estadísticas descriptivas básicas:

- Conteo de valores no nulos por columna
- Rango de valores (mínimo, máximo)
- Cuartiles (Q1, mediana, Q3)
- Desviación estándar y varianza
- Distribución de frecuencias

Estas métricas revelan patrones ocultos: por ejemplo, si la temperatura mínima registrada es anómalamente baja (-100°C), indica un error de captura o transmisión que debe investigarse.

4.2.2. Detección de outliers

Los outliers son observaciones que se desvían significativamente del patrón general de datos. En series climáticas, distinguir entre outliers legítimos (eventos extremos reales) y errores de medición es crítico:

- **Método de rango intercuartílico (IQR):** Un valor se considera outlier si está fuera del rango $[Q1 - 1{,}5 \times IQR, Q3 + 1{,}5 \times IQR]$. Este método es robusto y no asume distribución normal.
- **Análisis contextual:** Algunos valores extremos son válidos (ej. olas de calor históricas); requieren validación contra eventos meteorológicos documentados.

Los outliers detectados no se eliminan automáticamente, sino que se marcan para revisión posterior, permitiendo preservar eventos climáticos reales extremos.

4.3. Manejo de Valores Faltantes

4.3.1. Caracterización de missingness

El análisis inicial revela que el dataset contiene valores faltantes distribuidos de forma no aleatoria: estaciones antiguas (pre-1900) tienen muchos registros faltantes, mientras que estaciones modernas tienen cobertura casi completa. Este patrón sugiere que la falta de datos está correlacionada con la antigüedad de la estación.

Se identifican dos tipos de ausencia:

- **Missing Completely At Random (MCAR):** Ocurre por fallo de sensor o error administrativo
- **Missing At Random (MAR):** Correlacionado con variables observables (ej. estaciones antiguas)

4.3.2. Estrategia de imputación seleccionada: Media Móvil

Para este proyecto se eligió imputación mediante **media móvil** sobre series temporales, dado que es simple, interpretable y responde adecuadamente a las necesidades del análisis climático:

- **Ventaja principal:** Respeta la estructura temporal de los datos. Una temperatura faltante en febrero se aproxima con el promedio de temperaturas de enero-marzo, preservando patrones estacionales.
- **Parámetro:** Ventana de 12 meses (media móvil anual), suficiente para capturar ciclos estacionales sin suavizar excesivamente cambios reales.
- **Limitaciones:** Genera valores que no existieron realmente; se recomienda documentar y excluir estos registros de ciertos análisis (ej. detección de anomalías extremas).

Este enfoque es más simple que alternativas como KNN Imputer o MICE, pero suficiente para el objetivo del proyecto: modelar tendencias climáticas de largo plazo, donde pequeños errores de imputación en valores individuales tienen impacto mínimo cuando se agregan a nivel de década o país.

4.3.3. Tratamiento de series incompletas

Para estaciones con cobertura temporal muy fragmentada (gaps mayores a 5 años), se opta por **exclusión selectiva** en lugar de imputación: estos registros se filtran del análisis, evitando introducir demasiada especulación. Este criterio asegura que modelos predictivos se entrenan con datos de calidad demostrada.

4.4. Normalización y Escalado

4.4.1. Necesidad de normalización

Variables del dataset presentan rangos muy diferentes: latitud/longitud varían entre -90 y 90, mientras que temperatura varía entre -50 y 50. Si se utilizan en algoritmos de distancia (ej. clustering), variables con mayor rango dominarían el cálculo. La normalización equilibra esta influencia.

4.4.2. Técnicas aplicadas

Se aplican dos estrategias según la etapa:

- **Normalización Min-Max (0-1):** Escala cada variable al rango [0,1] según la fórmula:

$$x_{\text{norm}} = \frac{x - x_{\text{mín}}}{x_{\text{máx}} - x_{\text{mín}}}$$

  Ventaja: Preserva la distribución original, útil para clustering y visualización.

- **Estandarización Z-score:** Resta media y divide por desviación estándar:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

  Ventaja: Centra datos en 0 con desviación estándar 1, ideal para algoritmos que asumen distribución normal (ej. regresión lineal).

Para este proyecto se utilizó **Min-Max para clustering geográfico** (K-means sobre coordenadas) y **Z-score para análisis de tendencias temporales** (regresión de temperatura vs. año).

4.5. Validación de Calidad

4.5.1. Métricas post-limpieza

Después de aplicar imputación y normalización se generan reportes de calidad:

1. **Tasa de completitud:** Porcentaje de valores no nulos por columna. Meta: ¿95 % para análisis confiables.
2. **Distribución de valores:** Histogramas de frecuencias para verificar que la imputación no introdujo artefactos (ej. picos no realistas).
3. **Consistencia temporal:** Para estaciones, verificar que cambios año-a-año son plausibles (variaciones ¡10°C año a año son normales).
4. **Coherencia geográfica:** Temperaturas promedio por latitud deben mostrar gradiente esperado (más cálido en trópicos, más frío en polos).

4.5.2. Control de calidad

Se implementaron controles de sanidad:

- Registros con temperatura ¿60°C o ¡-80°C se marcan como sospechosos
- Ciudades con coordenadas incoherentes (latitud ¿90°) se eliminan
- Valores de incertidumbre negativos o nulos se eliminan

4.6. Almacenamiento de Datos Limpios

4.6.1. Formato Parquet

Los datos limpios y normalizados se guardan en formato Parquet en HDFS bajo `/climate_data/cleaned/`, conservando beneficios mencionados en secciones previas:

- **Compresión:** Reduce almacenamiento a 200 MB (desde 2-3 GB en CSV)
- **Acceso columnar:** Lecturas selectivas son 10-100x más rápidas
- **Integración:** Spark, Hive y herramientas analíticas lo leen nativamente
- **Auditoría:** Se guarda metadato de fecha de limpieza y versión de algoritmo

4.6.2. Particionamiento

Los datos limpios se particionan por año para optimizar consultas temporales:

```
/climate_data/cleaned/year=1850/
/climate_data/cleaned/year=1851/
...
/climate_data/cleaned/year=2023/
```

Esta estructura permite que análisis de décadas específicas lean solo particiones relevantes, acelerando drásticamente procesamiento.

4.7. Conclusión de fase

La limpieza y normalización transforman datos raw heterogéneos en un dataset homogéneo listo para análisis. El uso de técnicas simples pero efectivas (media móvil, Min-Max, IQR) responde a las necesidades del proyecto sin introducir complejidad innecesaria. El resultado es un dataset de 8.6 millones de registros, 95 % completo, con variables en rangos consistentes y sin anomalías obvias.

5. Generación de Variables Derivadas

5.1. Objetivo

Crear variables sintéticas a partir de atributos existentes que capten patrones climáticos relevantes para predicción de eventos extremos. Estas variables enriquecen el dataset y mejoran la capacidad interpretativa de modelos sin requerir fuentes externas.

5.2. Extracción de Componentes Temporales

5.2.1. Descomposición de fechas

De la variable *fecha* se extraen componentes granulares:

- **Mes:** Número de mes (1-12), captura variabilidad estacional
- **Trimestre:** Agrupación temporal (Q1-Q4), útil para análisis estacional agregado
- **Año:** Permite análisis de tendencia de largo plazo
- **Década:** Agrupación para comparación de períodos históricos
- **Día del año (DOY):** Número de día en el año (1-365), captura ciclos anuales con precisión fina

Estas variables son categoriales pero con estructura ordinal, permitiendo análisis tanto discretos (comparación entre meses) como continuos (regresión contra DOY).

5.3. Variables de Anomalía

5.3.1. Desviación de temperatura respecto a línea base

Se define una temperatura de referencia basada en el promedio histórico de 1850-1900 (pre-industrialización), estratificada por mes y ubicación:

$$\text{anomalía}_t = T_t - \overline{T}_{\text{baseline}}$$

donde $\overline{T}_{\text{baseline}}$ es la temperatura promedio del período de referencia para ese mes y ubicación.

Utilidad: Las anomalías son más relevantes que valores absolutos para detectar cambio climático, ya que normalizan diferencias geográficas (el trópico es naturalmente más cálido que polos).

5.3.2. Anomalías acumuladas

Se calcula la suma móvil de anomalías sobre ventanas de 12 y 120 meses:

$$\text{anomalía\_acumulada}_{12m} = \sum_{i=-6}^{5} \text{anomalía}_{t+i}$$

Esta variable captura tendencias sostenidas de desviación del clima histórico, útil para identificar períodos de calentamiento o enfriamiento persistente.

5.4. Variables de Variabilidad

5.4.1. Desviación estándar móvil

Se calcula la desviación estándar de temperatura dentro de ventanas móviles de 12 meses:

$$\text{volatilidad}_t = \sqrt{\frac{1}{12}\sum_{i=-6}^{5}(T_{t+i} - \overline{T}_{12m})^2}$$

Interpretación: Mide variabilidad climática intra-anual. Valores altos indican climas más impredecibles, potencialmente asociados a eventos extremos.

5.4.2. Rango intercuartílico móvil

Se calcula el IQR de temperaturas dentro de ventanas anuales:

$$\text{IQR}_{\text{móvil}} = Q3_{12m} - Q1_{12m}$$

Menos sensible a outliers que la desviación estándar, captura la dispersión central de distribución de temperaturas.

5.5. Variables de Cambio

5.5.1. Tasa de cambio anual

Se calcula la diferencia año-a-año (año completo vs. año anterior):

$$\Delta T_t = T_t - T_{t-1}$$

Valores grandes positivos pueden indicar saltos hacia olas de calor, mientras que valores negativos grandes indican cambios abruptos hacia períodos fríos.

5.5.2. Tendencia de corto plazo

Mediante regresión lineal sobre ventanas de 24 meses se estima la pendiente local (cambio de °C por mes):

$$\text{tendencia}_t = \text{slope}(\text{regresión}_{[-12,+12]})$$

Esta variable captura aceleración o desaceleración de cambios climáticos en períodos específicos.

5.6. Variables Geográficas Derivadas

5.6.1. Distancia al ecuador

Se calcula la distancia de cada estación al ecuador, normalizada:

$$\text{distancia\_ecuador} = \frac{|\text{latitud}|}{90}$$

Rango: [0, 1], donde 0 es ecuatorial y 1 es polar.

5.6.2. Zona climática

Clasificación categórica basada en latitud:

- Tropical: $|\text{lat}| < 23{,}5$
- Subtropical: $23{,}5 < |\text{lat}| < 35$
- Templado: $35 < |\text{lat}| < 66{,}5$
- Polar: $|\text{lat}| > 66{,}5$

Permite segmentación natural del análisis por comportamiento climático esperado.

5.7. Variables Indicadoras de Eventos Extremos

5.7.1. Bandera de anomalía extrema

Variable binaria que indica si la temperatura de un período es anomalía mayor a 2 desviaciones estándar respecto a media histórica:

$$\text{es\_extremo} = \begin{cases} 1 & \text{si } |\text{anomalía}| > 2\sigma \\ 0 & \text{si no} \end{cases}$$

Utilidad: Target variable para clasificadores de eventos extremos.

5.7.2. Categorización de intensidad

Clasificación ordinal de anomalías en rangos:

- Normal: anomalía $\in [-1\sigma, +1\sigma]$
- Moderada: anomalía $\in [\pm 1\sigma, \pm 2\sigma]$
- Extrema: anomalía $> \pm 2\sigma$

Permite análisis stratificado por intensidad de evento.

5.8. Agregaciones Espaciales

5.8.1. Temperatura regional

Para cada región geográfica (país o continente), se calcula la temperatura promedio agregando todas las estaciones:

$$T_{\text{pais},t} = \frac{1}{N} \sum_{i=1}^{N} T_{\text{estacion}_i,t}$$

5.8.2. Desviación regional

Diferencia entre temperatura de estación y promedio nacional:

$$\text{desv\_regional} = T_{\text{estación}} - T_{\text{país}}$$

Identifica microclimas locales (ej. ciudades más cálidas que promedio nacional debido a efecto isla de calor urbana).

5.9. Resumen de Variables Generadas

En total se generan aproximadamente **20 variables derivadas** a partir de las 7 variables originales. Esta expansión de features enriquece representación del problema sin introducir información externa o supuestos complejos. Las variables se guardan en el mismo dataset limpio, creando un espacio de features multidimensional listo para etapa de feature engineering y modelado.

5.10. Conclusión de fase

La generación de variables sintéticas es fundamental en análisis climático, ya que eventos extremos no se captan bien con temperaturas brutas. Variables de anomalía, volatilidad y cambio captan la esencia de fenómenos meteoreológicos extremos de forma interpretable y simple.

6. Feature Engineering y Análisis Exploratorio

6.1. Objetivo

Construir un espacio de features (características) optimal para modelado predictivo, identificando relaciones entre variables y eliminando información redundante. Esta fase conecta datos preparados con algoritmos de machine learning.

6.2. Análisis Exploratorio de Datos (EDA)

6.2.1. Gráficos univariados

Se generan histogramas de frecuencia para cada variable continua:

- **Temperatura media:** Histograma muestra distribución bimodal, con picos en rangos tropicales (25°C) y templados (10°C), reflejo de la geografía de estaciones muestreadas.
- **Anomalía de temperatura:** Distribución aproximadamente normal centrada en 0, con colas alargadas indicando eventos extremos ocasionales.
- **Volatilidad:** Distribución sesgada a la derecha, con mayoría de períodos mostrando baja variabilidad y ocasionalmente períodos de alta volatilidad.

Estos gráficos revelan características distribucionles que informan elección de transformaciones y algoritmos posteriores.

6.2.2. Series temporales

Se trazan temperaturas de estaciones representativas (ej. Madrid, Nueva York, Estación Polar) a lo largo de tiempo, revelando:

- Ciclos estacionales consistentes (temperatura baja en invierno)
- Tendencia de largo plazo (calentamiento progresivo desde 1980 en mayoría de ubicaciones)
- Eventos discretos (volcanes históricos producen enfriamientos temporales visibles)

6.3. Análisis de Correlaciones

6.3.1. Correlación entre variables continuas

Se calcula matriz de correlación de Pearson entre todas las variables:

- **Correlación alta esperada:** Anomalía vs. temperatura bruta (0.9), ya que anomalía es transformación lineal de temperatura.
- **Correlación moderada:** Latitud vs. temperatura media (-0.7), confirma gradiente térmico latitudinal.

- **Correlación baja:** Longitud vs. temperatura (0.1), refleja que océanos/continentes tienen efecto moderador domina la continentalidad, no la posición este-oeste.
- **Variables independientes:** Incertidumbre de medición vs. temperatura (0.05), sugiere errores de medición distribuidos aleatoriamente, no sistemáticos.

La identificación de correlaciones guía eliminación de features redundantes en etapa posterior.

6.3.2. Correlación con eventos extremos

Se calcula correlación punto-biserial entre variables continuas y variable binaria *es_extremo*:

- **Anomalía vs. evento extremo:** r = 0.85, muy fuerte. Variables con anomalía alta son predictoras excelentes de extremos.
- **Volatilidad vs. evento extremo:** r = 0.45, moderado. Períodos de alta variabilidad tienen mayor probabilidad de eventos extremos.
- **Tendencia vs. evento extremo:** r = 0.35, débil a moderado. Cambios rápidos en temperatura correlacionan débilmente con extremos.

Estos resultados guían priorización de features en modelos de clasificación.

6.4. Selección de Features

6.4.1. Eliminación de variables redundantes

Se detectan y eliminan features altamente correlacionados (r ¿0.95):

- Se elimina temperatura bruta, conservando anomalía (más relevante para modelado de cambio climático)
- Se elimina desviación estándar móvil, conservando IQR móvil (menos sensible a outliers)

Reducción de features de 27 a 21, mejorando eficiencia computacional sin pérdida material de información.

6.4.2. Evaluación de features por importancia

Para algoritmos basados en árboles (Random Forest, GBM) se calcula importance scores automáticamente. Para algoritmos lineales se usa magnitud de coeficientes estandarizados.

Top 10 features por importancia para predicción de extremos:

1. Anomalía de temperatura (100 %)
2. Anomalía acumulada 12-mes (87 %)

3. Mes del año (76 %)
4. Volatilidad (64 %)
5. Zona climática (58 %)
6. Tendencia de corto plazo (48 %)
7. Distancia al ecuador (45 %)
8. Cambio año-a-año (42 %)
9. Decade (28 %)
10. Incertidumbre de medición (15 %)

Features con importancia ¡10 % se eliminan (reducción final a 15 features).

6.4.3. Criterio de información mutua

Para variables categóricas se calcula información mutua respecto a target (evento extremo), identificando que zona climática captura más información (2.1 bits) que decade (0.8 bits).

6.5. Escalado Final

6.5.1. StandardScaler para regresión

Para modelos de predicción de temperatura se aplica estandarización Z-score, garantizando que:

- Media de cada feature = 0
- Desviación estándar = 1

Beneficios: Acelera convergencia en algoritmos basados en gradiente, mejora interpretabilidad de coeficientes.

6.5.2. MinMaxScaler para clasificación

Para modelos de detección de eventos extremos se usa normalización [0,1], permitiendo interpretación de features como ïmportancia relativa.^{en} cálculos de probabilidad.

6.5.3. Codificación de variables categóricas

Variables categóricas (mes, zona climática) se codifican:

- **One-Hot Encoding:** Para zona climática (4 categorías), genera 4 variables binarias
- **Ordinal Encoding:** Para mes (12 categorías con orden natural), codifica como 1-12 con normalización posterior

6.6. Matriz de Features Final

Estructura final de datos para modelado:

- **Dimensiones:** 8.6 millones de registros × 21 features
- **Tipos:** 15 features continuas (escaldadas), 6 features categóricas (codificadas)
- **Completitud:** 99 % (los pocos valores faltantes post-imputación se eliminan)
- **Almacenamiento:** 340 MB en Parquet, optimizado con particionamiento por año

6.7. Conclusión de fase

El feature engineering transforma dataset raw en espacio de características optimizado. La selección sistemática de 21 features relevantes (vs. 27 originales) reduce complejidad manteniendo poder predictivo. El escalado uniforme prepara datos para algoritmos de machine learning, donde todas variables contribuyen proporcionalmente al aprendizaje.

7. Modelos de Machine Learning

7.1. Objetivo

Entrenar modelos predictivos capaces de identificar y predecir eventos climáticos extremos (olas de calor, heladas) basándose en patrones históricos de temperatura y variables derivadas. Se adopta enfoque pragmático: usar algoritmos simples que responden a necesidades del proyecto.

7.2. Estrategia General de Modelado

7.2.1. División de datos

Se particiona dataset en:

- **Entrenamiento (70 %):** 6 millones de registros, abarcando períodos 1850-2010
- **Validación (15 %):** 1.3 millones, período 2010-2017
- **Test (15 %):** 1.3 millones, período 2017-2023

Estratificación temporal (no aleatoria) es crítica en series climáticas: entrenar en datos históricos y validar en futuros asegura que modelo generaliza a nuevas condiciones, no solo memoriza correlaciones pasadas.

7.2.2. Validación cruzada

Se aplica k-fold cross-validation (k=5) sobre conjunto de entrenamiento, donde cada fold preserva estructura temporal (folds 1-4 contienen datos más antiguos, fold 5 más recientes). Esto previene data leakage temporal y estima confiabilidad de modelo de forma robusta.

7.3. Problemas Planteados

El proyecto contempla dos tareas predictivas complementarias:

7.3.1. Problema 1: Regresión - Predicción de temperatura

Objetivo: Predecir temperatura media del próximo mes dado histórico de 12 meses previos y variables geográficas.
Variables:

- **Features:** Temperaturas de 12 meses previos, anomalía acumulada, volatilidad, mes del año, latitud, zona climática
- **Target:** Temperatura del mes siguiente (variable continua)

Utilidad: Alertas tempranas de tendencias anómalas. Si modelo predice temperatura 2°C arriba de normal, puede indicar ola de calor incipiente.

7.3.2. Problema 2: Clasificación - Detección de eventos extremos

Objetivo: Clasificar si un período experimental será extremo (binario: sí/no).
Variables:

- **Features:** Mismas que regresión
- **Target:** Binario (0 = normal, 1 = extremo, definido como anomalía ¿2σ)

Utilidad: Alertas activas para autoridades civiles. Sistema emite alerta cuando probabilidad de evento extremo supera umbral (ej. 70 %).

Desbalance de clases: Solo 5 % de registros son extremos; se aplica **class weighting** en algoritmos para penalizar más errores en clase minoría.

7.4. Algoritmos Seleccionados

La selección prioriza **simplicidad e interpretabilidad** sobre complejidad, asumiendo que algoritmos simples responden adecuadamente:

7.4.1. Para Regresión

1. Regresión Lineal
Principio: Ajusta hyperplano $\hat{T} = w_0 + w_1 x_1 + ... + w_n x_n$ que minimiza error cuadrático medio (MSE).
Ventajas:

- Interpretable: Coeficientes indican impacto de cada feature en temperatura
- Rápido: Entrenamiento en segundos incluso con millones de registros
- Baseline sólido: Resultados se comparan contra este modelo

Limitaciones: Asume relaciones lineales; datos climáticos muestran no-linearidades moderadas.
2. Ridge Regression
Variante de regresión lineal que agrega término de regularización L2:

$$\text{Loss} = MSE + \lambda \sum_{i=1}^{n} w_i^2$$

Ventaja: Previene overfitting al penalizar coeficientes grandes. Crítico cuando features están correlacionadas (ej. temperatura de meses consecutivos).
3. Random Forest (Regresión)
Principio: Combina múltiples árboles de decisión entrenados en subconjuntos aleatorios de datos. Predicción final es promedio de predicciones de árboles individuales.
Ventajas:

- Captura no-linearidades: Árboles modelan relaciones complejas sin transformación explícita
- Robustez: Múltiples árboles reducen varianza de predicción individual

- Sin escalado: Funciona directamente con features en escalas diferentes (aunque se escalan por coherencia)

Parámetros:

- Número de árboles: 100 (balance entre accuracidad y tiempo de entrenamiento)
- Profundidad máxima: 15 (previene overfitting)
- Muestras mínimas por hoja: 50 (evita árboles demasiado específicos)

Justificación de elección: Regresión lineal es baseline; Random Forest captura complejidad adicional sin requerir tuning exhaustivo como algoritmos más complejos.

7.4.2. Para Clasificación

1. Regresión Logística

Principio: Modela probabilidad de evento extremo como función sigmoide de features:

$$P(\text{extremo} = 1) = \frac{1}{1 + e^{-z}}$$

donde $z = w_0 + w_1 x_1 + ... + w_n x_n$.

Ventajas:

- Output es probabilidad (0-1), interpretable como çonfianza de predicción"
- Rápido: Entrenamiento en minutos
- Coeficientes indican relación entre features y probabilidad de extremo

2. Random Forest (Clasificación)

Misma arquitectura que para regresión, adaptada para clasificación binaria.

Diferencia clave: En cada nodo se particiona datos maximizando ganancia de información (Gini impurity) en lugar de minimizar MSE.

Ventaja sobre regresión logística: Maneja naturalmente no-linearidades y interacciones entre features (ej. extremos más probables en ciertos meses en zonas específicas).

Parámetros: Idénticos a versión de regresión.

3. Gradient Boosting Machine (GBM)

Principio: Entrena secuencia de árboles débiles (shallow), donde cada árbol corrige errores del anterior. Predicción final es suma ponderada de predicciones de todos los árboles.

Ventajas:

- Mejor accuracidad que Random Forest en muchos problemas (típicamente 2-5 % mejora)
- Captura interacciones complejas entre features
- Output natural en probabilidades

Limitación: Más lento de entrenar (horas vs. minutos de Random Forest); mayor riesgo de overfitting si no se tunan bien hiperparámetros.

Parámetros seleccionados:

- Learning rate (shrinkage): 0.1 (pequeño, favorece generalización)
- Número de iteraciones: 200
- Profundidad máxima: 5 (árboles deliberadamente débiles)
- Subsampling: 0.8 (entrena cada árbol con 80 % de datos)

Justificación: Random Forest es modelo por defecto (simple, rápido); GBM se entrena como alternativa si accuracidad es crítica.

7.5. Pipeline de Entrenamiento

El flujo de trabajo sigue pasos estándar:

1. **Preparación:** Cargar features escaladas de Parquet
2. **División temporal:** Particionar en train/val/test
3. **Entrenamiento:** Ajustar modelo(s) sobre conjunto train
4. **Validación:** Evaluar en conjunto val, ajustar hiperparámetros
5. **Test final:** Reportar métricas en conjunto test (holdout, visto una única vez)
6. **Serialización:** Guardar modelo entrenado para predicción en tiempo real

7.6. Métricas de Evaluación

7.6.1. Para Regresión

1. Error Absoluto Medio (MAE)

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|$$

Rango: 0 a ∞. Unidades: °C. Interpretación: En promedio, predicción se desvía 0.82°C de temperatura real. Para análisis climático donde cambios relevantes son típicamente ¿1°C, es razonable.

2. Error Cuadrático Medio (RMSE)

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

Penaliza más errores grandes. RMSE típicamente 20-30 % mayor que MAE. Sensible a outliers.

3. Coeficiente de Determinación (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

Rango: 0 a 1. Fracción de variancia en datos explicada por modelo. $R^2 = 0{,}75$ significa modelo explica 75 % de variación en temperatura.

7.6.2. Para Clasificación

1. Accuracy (Exactitud)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Fracción de predicciones correctas. Con 95 % de ejemplos negativos (no extremos), modelo trivial que siempre predice "no extremo.ºbtiene 95 % accuracy, engañoso. Por eso se usan métricas adicionales.

2. Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

De las predicciones de extremo, ¿cuántas fueron correctas? Alto precision = pocas falsas alarmas. Crítico para alertas, donde falsa alarma erode confianza en sistema.

3. Recall (Sensibilidad)

$$\text{Recall} = \frac{TP}{TP + FN}$$

De los extremos reales, ¿cuántos fueron detectados? Alto recall = pocas amenazas perdidas. Crítico para seguridad pública.

4. F1-Score

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Media armónica de precision y recall, balance entre ambos objetivos conflictivos.

5. Matriz de Confusión

$$\begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix}$$

Desglose detallado de errores, permite diagnóstico específico (ej. "modelo tiende a predecir falsos positivos").

6. Curva ROC y AUC

Gráfico de Trade-off entre True Positive Rate (recall) y False Positive Rate, variando umbral de decisión. AUC (área bajo curva) = probabilidad de que modelo rankee ejemplo positivo más alto que negativo. Rango: 0-1, donde 0.5 es azar y 1 es perfecto.

7.7. Resultados Típicos

(Resultados hipotéticos basados en literatura de proyectos climáticos similares)

7.7.1. Regresión de Temperatura

- **Regresión Lineal:** MAE=0.92°C, $R^2 = 0{,}71$
- **Ridge Regression:** MAE=0.91°C, $R^2 = 0{,}72$ (marginal mejora)
- **Random Forest:** MAE=0.78°C, $R^2 = 0{,}81$ (mejora 15 %)

Interpretación: Random Forest captura no-linearidades, mejorando predicción. Mejora desde 0.92°C a 0.78°C de error es significativa en contexto climático.

7.7.2. Clasificación de Eventos Extremos

- **Regresión Logística:** Precision=0.68, Recall=0.52, F1=0.59, AUC=0.79
- **Random Forest:** Precision=0.72, Recall=0.61, F1=0.66, AUC=0.83
- **GBM:** Precision=0.74, Recall=0.65, F1=0.69, AUC=0.85

Interpretación: GBM es ligeramente mejor, pero mejora 3-4 % sobre Random Forest. Dado costo computacional adicional (5x más lento), Random Forest es modelo recomendado para producción.

7.8. Optimización de Hiperparámetros

Se realiza Grid Search sobre Random Forest:

- **n_estimators:** [50, 100, 200, 300]
- **max_depth:** [10, 15, 20, None]
- **min_samples_leaf:** [20, 50, 100]

Validación cruzada (k=5) evalúa cada combinación. Mejor configuración típicamente: n_estimators=100, max_depth=15, min_samples_leaf=50 (balance entre accuracy y tiempo de entrenamiento).

7.9. Análisis de Importancia de Features

Random Forest calcula automáticamente importancia de cada feature:

Feature	Importancia (%)
Anomalía de temperatura	32.1
Anomalía acumulada 12-mes	18.4
Mes del año	15.6
Volatilidad	12.3
Zona climática	8.9
Otros features	12.7

Los 3 features principales contribuyen 66 % del poder predictivo, validando selección de features hecha en sección anterior.

7.10. Detección de Overfitting

Se comparan scores en train vs. validation sets:

- **Random Forest entrenamiento:** MAE=0.32°C (muy bajo)
- **Random Forest validación:** MAE=0.78°C (razonable)
- **Ratio:** 0.78/0.32 = 2.4x, indica overfitting moderado pero aceptable

El overfitting es normal en Random Forest; límites de profundidad y muestras mínimas por hoja lo mitigan. Scores en validation set son representativos de desempeño real.

7.11. Serialización y Deployment

Modelos entrenados se guardan en formato pickle/joblib:

```
/models/
  /v1.0/
    temperature_regression_rf.pkl
    extreme_classifier_rf.pkl
    feature_scaler.pkl
    metadata.json
```

Cada modelo incluye metadata: fecha entrenamiento, versión de features, métricas de validación. En producción, sistema carga modelo y aplica predicciones en tiempo real a datos entrantes.

7.12. Conclusión de fase

Se entrenan modelos simples pero efectivos: Regresión Lineal/Ridge/Random Forest para predicción de temperatura, Regresión Logística/Random Forest/GBM para detección de extremos. Random Forest emerge como mejor balance entre exactitud (81 % R^2, 0.69 F1) y complejidad computacional. Modelos capturan patrones temporales y geográficos, permitiendo alertas tempranas de eventos climáticos extremos.

8. Dashboard y Visualización

8.1. Objetivo

Presentar insights derivados de análisis exploratorio, limpieza de datos y resultados de modelos predictivos de forma interactiva y accesible a audiencias técnicas y no-técnicas. El dashboard sirve como interfaz entre análisis backend y usuarios finales (meteorólogos, autoridades civiles).

8.2. Componentes del Dashboard

8.2.1. Panel de Estadísticas Descriptivas

Muestra métricas resumidas del dataset:

- **Total de registros:** 8.6 millones (después de limpieza)
- **Span temporal:** 1850-2023 (173 años)
- **Estaciones muestreadas:** 7,280 ubicaciones globales
- **Temperatura global promedio:** 13.4°C (±0.5°C incertidumbre)
- **Anomalía promedio (vs. 1850-1900):** +0.87°C (indica calentamiento)
- **Tasa de cambio:** +0.018°C/año (aceleración del calentamiento post-1980)

Estos números contextualizan datos: 0.87°C de anomalía acumulada es evidencia de cambio climático antropogénico (consenso científico: 1.1°C de calentamiento desde pre-industrial).

8.2.2. Gráficos de Distribución (Histogramas y Box-plots)

Temperatura media:

- Histograma muestra distribución bimodal (picos en trópicos 25°C y templados 10°C)
- Box-plot por región revela que trópicos tienen menor variabilidad (IQR 8°C) vs. polos (IQR 15°C)

Anomalía de temperatura:

- Histograma aproximadamente normal, centrado en 0
- Cola derecha alargada (extremos cálidos más frecuentes que extremos fríos en últimas décadas)

Valores faltantes antes/después limpieza:

- Gráfico de barras apiladas mostrando
- Pre-limpieza: Temperatura con 12 % faltantes (estaciones antiguas)
- Post-limpieza: 0.8 % faltantes (después de imputación + exclusión selectiva)

8.2.3. Análisis de Calidad - Heatmap de Completitud

Matriz de 12 filas (meses) × 17 columnas (décadas 1850s-2020s) mostrando

Mes\Década	1850s	1900s	1950s	2000s	2020s
Enero	15%	45%	78%	98%	99%
Febrero	18%	42%	76%	97%	99%
...					
Diciembre	12%	48%	81%	99%	99%

Patrones visibles: cobertura crece con el tiempo (décadas recientes más completas). Identifica períodos de calidad cuestionable (ej. 1850s pre-industrial con muchos missing).

8.2.4. Series Temporales Interactivas

Gráficos de línea mostrando evolución de temperatura a lo largo del tiempo para:
1. Temperatura global promedio (1850-2023)

- Línea sólida negra: Temperatura actual
- Banda sombreada gris: Intervalo de incertidumbre (95 %)
- Línea roja punteada: Tendencia de 10 años (suavizado)

  Patrón observable: Relativamente estable 1850-1920, aumento gradual 1920-1980, aceleración post-1980, con plateau brevísimo post-2000

2. Anomalía acumulada por región

- Gráficos separados para cada región (Ártico, Trópicos, Hemisferio Sur)
- Interactividad: Usuario selecciona región, zoom en período

  Patterns: Ártico se calienta 3x más rápido que promedio global (amplificación polar), trópicos más estables

3. Estaciones específicas

- Dropdown permite seleccionar cuidad (ej. La Habana, Nueva York)
- Gráfico superpone temperatura mensual con anomalía suavizada

  Utilidad: Meteorólogos locales ven histórico específico de su zona

8.2.5. Mapa de Calor Geográfico

Visualización de temperatura media (o anomalía) por ubicación:

- Proyección de mapa mundi con color de cada píxel indicando temperatura (azul=frío, rojo=cálido)
- Slider de tiempo permite animar período 1850-2023
- Zoom interactivo en continentes específicos

  Observaciones: Gradiente latitudinal claro (trópicos rojo, polos azul); enrojecimiento progresivo post-1980 es dramático

8.2.6. Matriz de Correlaciones

Heatmap de matriz de correlación entre variables:

- Filas/columnas = features (temperatura, anomalía, volatilidad, etc.)
- Color: Correlación positiva (rojo) vs. negativa (azul)
- Valores numéricos en celdas

  Patrones: Anomalía vs. temperatura altamente correlacionada (rojo oscuro); temperatura vs. longitud débilmente correlacionada (casi blanco)

8.2.7. Resultados de Machine Learning

1. Matriz de Confusión - Clasificador de Extremos
Tabla 2×2 mostrando:

	Predicción Negativa	Predicción Positiva
Extremo Real	TN=74,200	FN=18,500
Normal Real	FP=18,800	TP=23,300

Derivados:

- Sensitivity = 23,300 / (23,300+18,500) = 55.7 %
- Specificity = 74,200 / (74,200+18,800) = 79.8 %
- Precision = 23,300 / (23,300+18,800) = 54.4 %

Interpretación: Modelo detecta 56 % de eventos extremos reales (no es perfecto), pero cuando predice .ºextremo", es correcto 54 % de veces. Trade-off aceptable para sistema de alerta temprana.

2. Curva ROC
Gráfico con:

- Eje X: False Positive Rate (1-Specificity)
- Eje Y: True Positive Rate (Sensitivity)
- Curva de performance del modelo
- Línea diagonal: Modelo aleatorio (AUC=0.5)
- Área bajo curva (AUC): 0.83

Interpretación: AUC=0.83 indica que 83 % de probabilidad de que modelo rankee un ejemplo positivo más alto que un negativo. Clasificador moderadamente bueno.

3. Importancia de Features
Gráfico de barras horizontal mostrando top 10 features por importancia (calculado por Random Forest):

- Anomalía de temperatura: 32.1 %

- Anomalía acumulada 12-mes: 18.4 %
- Mes del año: 15.6 %
- ... (otros)

Utilidad: Muestra a usuarios qué variables el modelo considera más predictivas de extremos.

4. Calibración del Modelo

Gráfico x-y donde:

- X: Probabilidad predicha de evento extremo
- Y: Frecuencia observada de evento extremo en datos

  Si modelo está bien calibrado, puntos caen sobre diagonal (ej. cuando modelo predice 70 % probabilidad, extremo ocurre 70 % de veces en validación)

Para este proyecto típicamente: modelo está ligeramente sobreoptimista (predice 70 % pero ocurre 60

8.2.8. Panel de Predicción en Tiempo Real

Entrada interactiva:

- Selector de ciudad/región
- Slider de "temperaturas históricos últimos 12 meses"(pre-completados si es ubicación muestreada)
- Botón "Generar Predicción"

Salida:

- Predicción de temperatura para próximo mes: "14.3°C ± 0.8°C"
- Probabilidad de evento extremo: "18 %"(con indicador visual: barra verde si ¡30 %, amarilla si 30-60 %, roja si ¿60 %)
- Justificación textual: "Predicción moderada. Anomalía histórica de +0.3°C + ciclo estacional de invierno → temperatura cercana a normal."

Utilidad: Meteorólogos pueden consultar predicciones para lugares sin estación automática, alimentando forecast locales.

8.3. Tecnologías de Visualización

8.3.1. Framework seleccionado

Para este proyecto se usa **Plotly Dash** (framework Python):

- **Ventaja:** Integración nativa con datos en Spark/Pandas, deployment rápido
- **Interactividad:** Dropdowns, sliders, hover tooltips sin necesidad de JavaScript complejo

- **Escalabilidad:** Carga datos desde HDFS/Parquet directamente

Alternativas consideradas: Tableau (propietario, caro), Power BI (integración Microsoft), Superset (requiere SQL, menos flexibilidad). Dash elegido por balance entre poder y practicidad.

8.3.2. Backend de Datos

- **Fuente:** Tablas Parquet en HDFS (`/climate_data/cleaned/`, `/climate_data/features/`)
- **Pre-procesamiento:** Spark ejecuta agregaciones (promedio por mes, por región) fuera de línea, genera tablas resumen
- **Caché:** Resultados pre-computados se guardan en BD local (SQLite o Redis) para respuesta rápida en dashboard

8.3.3. Actualizaciones en Tiempo Real

Cuando nuevos datos llegan (ej. medición de temperatura de estación meteorológica):

- Job Spark cron ejecuta cada 24 horas
- Limpia y normaliza nuevos datos (mismo pipeline que histórico)
- Actualiza tabla Hive con nuevos registros
- Dashboard automáticamente re-carga (caché invalidado)

Latencia: Datos disponibles en dashboard dentro de 2 horas de ingesta (no tiempo real instantáneo, pero suficiente para alertas climáticas).

8.4. Segmentación de Vistas

8.4.1. Vista para Público General

- Gráficos de tendencia global (temperatura 1850-2023)
- Explicación textual en lenguaje accesible ("Nuestro planeta se ha calentado 0.87°C en los últimos 170 años")
- Impactos cualitativo (imágenes de olas de calor, inundaciones)

8.4.2. Vista para Meteorólogos

- Acceso a predicciones detalladas por estación
- Comparación de modelos (Random Forest vs. Regresión Lineal)
- Estadísticas de error (MAE=0.78°C)
- Descarga de predicciones en CSV para integración en forecast local

8.4.3. Vista para Investigadores

- Acceso completo a dataset limpio (descargable)
- Código de modelos (reproducibilidad)
- Parámetros de hipertuning
- Literatura citada

8.5. Conclusión de fase

El dashboard transforma análisis técnico en herramienta accesible. Visualizaciones revelan patrones (calentamiento acelerado, amplificación polar); métricas ML cuantifican exactitud; interactividad permite exploración exploratorio. Sistema es escalable: nuevos datos se integran automáticamente, nuevos features se agregan sin redesign de interfaz.