Informe de Proyecto

ClimaXtreme: análisis climático y modelado de eventos extremos

Eveliz Espinaco Milián
Richard Alejandro Matos Arderí

Universidad de La Habana,
Facultad de Matemática y Computación

6 de diciembre de 2025

Resumen: Este informe presenta ClimaXtreme, una herramienta para el análisis climático y el modelado de eventos extremos sobre datos a gran escala con Hadoop y PySpark. Se implementa una canalización distribuida para la ingesta, limpieza y enriquecimiento de datos; agregaciones temporales y espaciales; y análisis estadístico, incluyendo teoría de valores extremos. Se desarrollan modelos predictivos y visualizaciones para detectar tendencias, anomalías y riesgos, priorizando calidad de datos, trazabilidad y reproducibilidad en entornos Big Data.

Índice

1 Dataset Seleccionado

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

1.1 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

1.2 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 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.

2.1 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.

2.2 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 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.

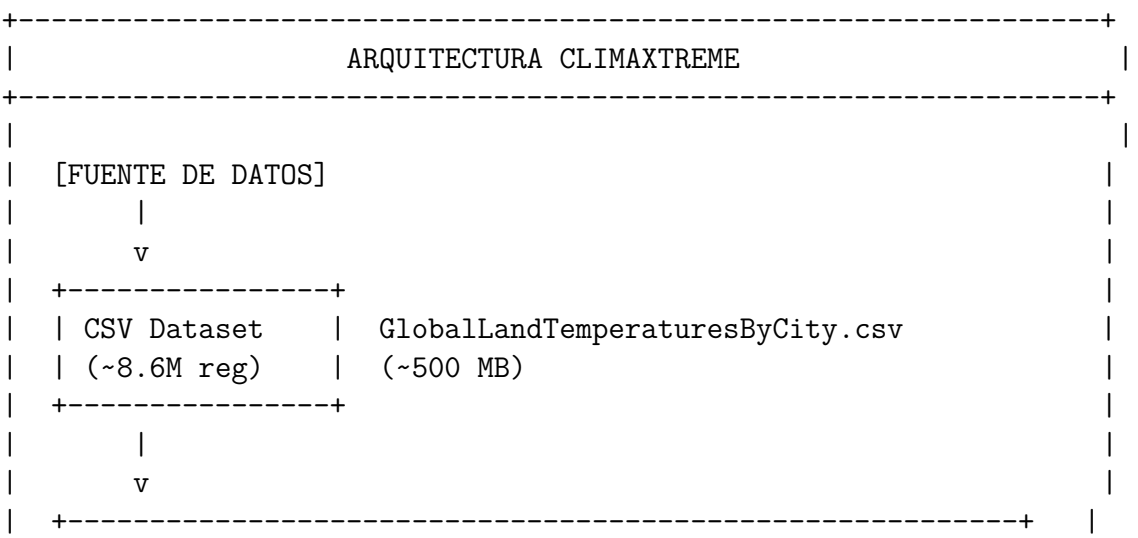
3 Arquitectura del Sistema

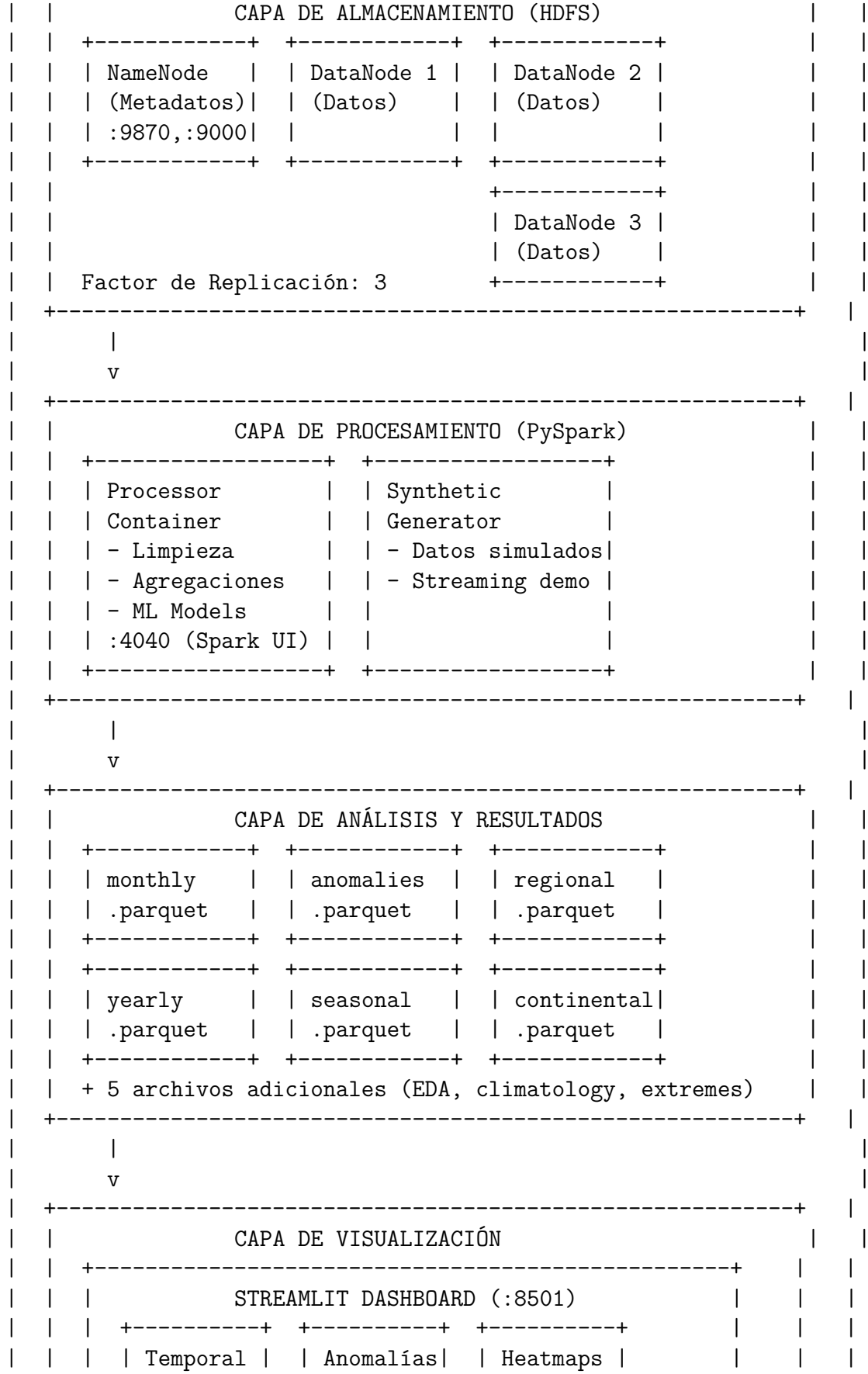
Esta sección describe la arquitectura completa del sistema ClimaXtreme, incluyendo los componentes de infraestructura, procesamiento y visualización.

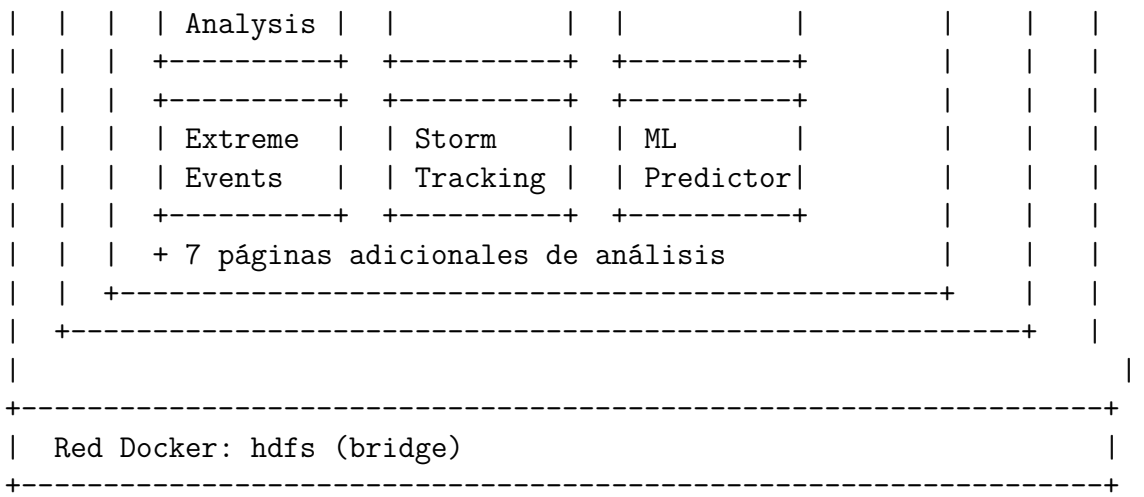
3.1 Visión General

El sistema implementa una arquitectura híbrida que combina procesamiento **batch** para el análisis histórico del dataset completo y capacidades de **streaming simulado** para demostrar análisis en tiempo real. La arquitectura se despliega completamente en contenedores Docker, garantizando reproducibilidad y portabilidad.

3.2 Diagrama de Arquitectura




CAPA DE ALMACENAMIENTO (HDFS)
NameNode (Metadatos) :9870,:9000
DataNode 1 (Datos)
DataNode 2 (Datos)
DataNode 3 (Datos)
Factor de Replicación: 3
CAPA DE PROCESAMIENTO (PySpark)
Processor Container - Limpieza - Agregaciones - ML Models :4040 (Spark UI)
Synthetic Generator - Datos simulados - Streaming demo
CAPA DE ANÁLISIS Y RESULTADOS
monthly .parquet
anomalies .parquet
regional .parquet
yearly .parquet
seasonal .parquet
continental .parquet
+ 5 archivos adicionales (EDA, climatology, extremes)
CAPA DE VISUALIZACIÓN
STREAMLIT DASHBOARD (:8501)
Temporal
Anomalías
Heatmaps



3.3 Componentes del Sistema

Capa de Almacenamiento: HDFS

El sistema de archivos distribuido Hadoop (HDFS) proporciona almacenamiento tolerante a fallos para los datos climáticos.

- **NameNode:** Nodo maestro que gestiona los metadatos del sistema de archivos, incluyendo la estructura de directorios y la ubicación de los bloques de datos. Expone la interfaz web en el puerto 9870 y la API HDFS en el puerto 9000.
- **DataNodes (x3):** Tres nodos de trabajo que almacenan los bloques de datos reales. Con un factor de replicación de 3, cada bloque se replica en los tres nodos, garantizando alta disponibilidad y tolerancia a fallos.
- **Estructura de directorios HDFS:**
  - `/data/raw/`: Datos originales en formato CSV
  - `/data/processed/`: Archivos Parquet procesados
  - `/data/synthetic/`: Datos sintéticos generados

Capa de Procesamiento: PySpark

Apache Spark se utiliza para el procesamiento distribuido de los datos climáticos.

- **Processor Container:** Contenedor principal que ejecuta los jobs de Spark para limpieza, transformación y agregación de datos. Incluye:
  - Limpieza de datos (eliminación de nulos, normalización de coordenadas)
  - Agregaciones temporales (mensual, anual, estacional)
  - Agregaciones espaciales (regional, continental)
  - Análisis estadístico (correlaciones, estadísticas descriptivas)
  - Detección de anomalías y eventos extremos

- **Synthetic Generator:** Servicio opcional para generar datos sintéticos que complementan el dataset original, permitiendo simulaciones de streaming y eventos extremos.
- **Spark UI (puerto 4040):** Interfaz web para monitorear la ejecución de jobs, stages, tasks y métricas de rendimiento.

Capa de Análisis: Archivos Parquet

El pipeline genera 11 archivos Parquet optimizados para consultas analíticas:

Archivo	**Descripción**	**Registros Aprox.**
monthly.parquet	Agregaciones mensuales por ciudad	8.6M
yearly.parquet	Agregaciones anuales por ciudad	350K
anomalies.parquet	Desviaciones de temperatura	350K
climatology.parquet	Valores climatológicos promedio	170K
seasonal.parquet	Agregaciones por estación	1M
extreme_thresholds.parquet	Umbrales P10, P90	170K
regional.parquet	Agregaciones por 16 regiones	2K
continental.parquet	Agregaciones por 7 continentes	300
correlation_matrix.parquet	Matriz de Pearson	25
descriptive_stats.parquet	11 métricas estadísticas	4
chi_square_tests.parquet	Tests de independencia	3

Cuadro 1: Archivos Parquet generados por el pipeline

Capa de Visualización: Streamlit Dashboard

Dashboard interactivo con 13 páginas de análisis:

1. **Temporal Analysis:** Tendencias de temperatura a lo largo del tiempo
2. **Anomalies:** Detección y visualización de anomalías térmicas
3. **Seasonal Analysis:** Patrones estacionales y climatología
4. **Extreme Events:** Identificación de eventos extremos
5. **Country Analysis:** Análisis por país
6. **Continental Analysis:** Análisis por continente con mapas interactivos
7. **Statistical Analysis:** EDA con correlaciones y estadísticas
8. **Climate Heatmaps:** Mapas de calor geográficos
9. **Storm Tracking:** Seguimiento de tormentas (datos sintéticos)
10. **Active Alerts:** Sistema de alertas en tiempo real

11. **Intensity Prediction:** Predicción de intensidad con ML
12. **Weather TimeSeries:** Series temporales meteorológicas
13. **Historical Comparison:** Comparación entre períodos históricos

3.4 Enfoque de Procesamiento

El sistema implementa un enfoque híbrido:

- **Batch Processing:** Procesamiento del dataset histórico completo (8.6M registros) mediante jobs de Spark. Este modo se utiliza para generar las agregaciones y análisis que alimentan el dashboard.
- **Streaming Simulado:** Módulo de demostración que simula la llegada de datos en tiempo real mediante cadenas de Markov y distribuciones estadísticas. Permite probar capacidades de alertas y visualizaciones en tiempo real.

3.5 Infraestructura Docker

Toda la infraestructura se despliega mediante Docker Compose, facilitando la reproducibilidad:

```
docker-compose.yml
 namenode (bde2020/hadoop-namenode)
 datanode1 (bde2020/hadoop-datanode)
 datanode2 (bde2020/hadoop-datanode)
 datanode3 (bde2020/hadoop-datanode)
 processor (custom: Dockerfile.processor)
 dashboard (custom: Dockerfile.processor)
 synthetic-generator (profile: synthetic)
```

Red: Todos los contenedores se comunican a través de una red bridge llamada `hdfs`.
Volúmenes: Se utilizan volúmenes Docker para persistir los datos de HDFS y los datos sintéticos.

4 Metodología de Procesamiento

Esta sección detalla las transformaciones y algoritmos aplicados a los datos climáticos.

4.1 Pipeline de Procesamiento

El pipeline de procesamiento sigue las siguientes etapas:

1. Ingesta de Datos
   - Lectura del archivo CSV desde HDFS usando Spark
   - Schema inferido automáticamente con validación de tipos
   - Particionamiento inicial basado en el tamaño del archivo

2. Limpieza de Datos

   Las transformaciones de limpieza incluyen:
   - **Eliminación de nulos:** Registros sin temperatura se eliminan
   - **Normalización de coordenadas:** Conversión de formato "57.05N" a valores decimales (57.05)
   - **Extracción temporal:** Derivación de año, mes, día de semana desde la fecha
   - **Validación de rangos:** Temperaturas fuera de [-90°C, 60°C] se marcan como outliers
   - **Deduplicación:** Eliminación de registros duplicados por ciudad y fecha

3. Agregaciones Temporales
   - **Mensual:** Promedio, mínimo, máximo y desviación estándar por ciudad/mes
   - **Anual:** Agregación de métricas mensuales a nivel anual
   - **Estacional:** Agrupación por estaciones (DJF, MAM, JJA, SON)
   - **Climatología:** Promedios históricos de referencia (1961-1990)

4. Agregaciones Espaciales
   - **Regional:** Clasificación en 16 regiones geográficas basada en latitud/longitud
   - **Continental:** Agregación por 7 continentes

5. Análisis Estadístico
   - **Matriz de correlación de Pearson:** Entre variables numéricas (año, temperatura promedio, mínima, máxima, rango)
   - **Estadísticas descriptivas:** Media, mediana, desviación estándar, cuartiles, asimetría, curtosis
   - **Pruebas Chi-cuadrado:** Tests de independencia entre variables categóricas (continente, estación, período vs. categoría de temperatura)

6. Detección de Anomalías

Las anomalías se calculan como desviaciones respecto a la climatología de referencia:

$$\text{Anomalía} = T_{observada} - T_{climatología} \tag{1}$$

Se clasifican según umbrales:

- $|\text{Anomalía}| < 1\sigma$: Normal
- $1\sigma \leq |\text{Anomalía}| < 2\sigma$: Moderada
- $|\text{Anomalía}| \geq 2\sigma$: Extrema

4.2 Modelos de Machine Learning

El sistema implementa una arquitectura de Machine Learning por capas, con modelos base individuales que se combinan en un ensemble para mejorar la precisión de las predicciones.

Modelos Base (BaselineModel)

La clase `BaselineModel` implementa cinco tipos de modelos de regresión:

Modelo	**Hiperparámetros**	**Características**
Linear Regression	-	Modelo base simple, rápido de entrenar
Ridge Regression	$\alpha = 1{,}0$	Regularización L2 para evitar overfitting
Lasso Regression	$\alpha = 1{,}0$	Regularización L1, selección automática de features
Random Forest	n_estimators=100, n_jobs=-1	Ensemble de árboles, robusto ante outliers
Gradient Boosting	n_estimators=100	Boosting secuencial, alta precisión

Cuadro 2: Modelos base implementados en el sistema

Features utilizadas:

- `year`: Año de la medición
- `month`: Mes (1-12)
- `year_normalized`: Año normalizado al rango [0,1]
- `month_sin, month_cos`: Codificación cíclica del mes
- Features adicionales derivadas del dataset

Modelo Ensemble (ClimatePredictor)

La clase `ClimatePredictor` combina múltiples modelos base usando `VotingRegressor` de scikit-learn:

```
VotingRegressor(estimators=[
    ('linear', LinearRegression()),
    ('ridge', Ridge(alpha=1.0)),
    ('random_forest', RandomForestRegressor(n_estimators=100))
])
```

Características del ensemble:

- Combinación por votación promediada de predicciones
- Time Series Cross-Validation con 5 splits para evaluación
- Cuantificación de incertidumbre mediante dispersión de predicciones individuales
- Persistencia de modelos con `joblib`

Predictor de Intensidad (IntensityPredictor)

Modelo especializado para predecir la intensidad de eventos extremos en una escala de 0 a 10.

Features del predictor de intensidad:

- `temperature_hourly`: Temperatura horaria (°C)
- `rain_mm`: Precipitación (mm)
- `wind_speed_kmh`: Velocidad del viento (km/h)
- `humidity_pct`: Humedad relativa (%)
- `pressure_hpa`: Presión atmosférica (hPa)
- `month, hour`: Variables temporales
- `latitude_numeric`: Latitud numérica
- Codificación one-hot de zona climática y tipo de evento

Configuración del modelo:

- Random Forest: max_depth=15, min_samples_split=5, min_samples_leaf=2
- Gradient Boosting: max_depth=8, learning_rate=0.1
- Ensemble: VotingRegressor con RF + GB

Métricas de Evaluación

Los modelos se evalúan con las siguientes métricas:

Métrica	**Fórmula**	**Interpretación**
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Error cuadrático medio (misma unidad que y)
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Error absoluto medio
R^2	$1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$	Varianza explicada (0 a 1)

Cuadro 3: Métricas de evaluación de modelos

Resultados Esperados de Modelos

[Tabla a completar con métricas reales tras ejecución:]

Modelo	**RMSE (°C)**	**MAE (°C)**	**R^2**
Linear Regression	-	-	-
Ridge Regression	-	-	-
Random Forest	-	-	-
Gradient Boosting	-	-	-
Ensemble	-	-	-

Cuadro 4: Comparativa de rendimiento de modelos (pendiente de ejecución)

4.3 Generación de Datos Sintéticos

Para las visualizaciones avanzadas (tormentas, alertas, streaming), se generan datos sintéticos:

Modelo de Temperatura Horaria

$$T_{horaria}(h) = T_{media} + A_{diurna} \cdot \sin\left(\frac{2\pi(h - h_{max})}{24}\right) + \epsilon \tag{2}$$

Donde:

- T_{media}: Temperatura media diaria del dataset original
- A_{diurna}: Amplitud diurna (función de latitud y estación)
- h_{max}: Hora de temperatura máxima (14:00)
- $\epsilon \sim N(0, \sigma^2)$: Ruido gaussiano

Modelo de Precipitación

Cadena de Markov de orden 1 para estados wet/dry, con cantidad de lluvia modelada mediante distribución Gamma.

Tracking de Tormentas

Simulación de trayectorias de tormentas con:

- Posición inicial aleatoria en zonas de formación
- Movimiento basado en patrones climáticos típicos
- Intensificación/debilitamiento según temperatura del océano

5 Análisis de Resultados

[Esta sección se completará con los resultados del análisis climático una vez ejecutado el pipeline completo. Incluirá:]

5.1 Tendencias de Temperatura Global

Análisis de la evolución de temperaturas desde 1750 hasta la actualidad, identificando períodos de calentamiento y enfriamiento.

5.2 Patrones Regionales

Comparación de tendencias entre las 16 regiones geográficas y los 7 continentes.

5.3 Eventos Extremos Detectados

Estadísticas sobre anomalías extremas detectadas, frecuencia por década y distribución geográfica.

5.4 Rendimiento de Modelos Predictivos

Tabla comparativa de métricas (RMSE, MAE, R^2) para cada modelo implementado.

6 Análisis de Rendimiento del Clúster

[Esta sección se completará con las métricas capturadas durante la ejecución del pipeline. Incluirá:]

6.1 Consumo de Recursos

Gráficos de CPU, RAM y disco durante la ejecución de jobs de Spark.

6.2 Tiempos de Ejecución

Tabla comparativa de tiempos por operación:

Operación	Tiempo (s)	Registros	Throughput
Lectura CSV	-	8.6M	-
Limpieza	-	-	-
Agregación Mensual	-	-	-
Agregación Anual	-	-	-
Detección Anomalías	-	-	-
Total Pipeline	-	-	-

Cuadro 5: Tiempos de ejecución del pipeline (pendiente de medición)

6.3 Optimizaciones Aplicadas

Descripción de optimizaciones implementadas y su impacto en el rendimiento.

6.4 Cuellos de Botella Identificados

Análisis de stages lentas, shuffles y posibles mejoras.

7 Conclusiones

[Esta sección se completará al finalizar el proyecto. Incluirá:]

- Resumen de logros alcanzados
- Lecciones aprendidas sobre procesamiento de grandes volúmenes de datos
- Limitaciones del sistema actual
- Trabajo futuro y posibles mejoras

8 Referencias

1. Berkeley Earth. (2017). Climate Change: Earth Surface Temperature Data. Kaggle. https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-tem
2. Apache Spark Documentation. https://spark.apache.org/docs/latest/
3. Hadoop HDFS Architecture Guide. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html
4. Streamlit Documentation. https://docs.streamlit.io/
5. Docker Documentation. https://docs.docker.com/