

PROYECTO INTEGRADOR

SIATA, ESTUDIO DE PRECIPITACIONES EN MEDELLÍN

Autores:

Emanuel Castañeda Cardona

Yeniffer Andrea Córdoba González

Pablo Gómez Mutis

Francisco Javier Villadiego Yanes

Luz Adriana Yepes Arias

2 de diciembre 2024

Universidad EAFIT

ÍNDICE

1. INTRODUCCIÓN

1.1. Marco teórico

1.2. Comprensión del negocio

 1.2.1 Dominio del estudio

1.3. Entendimiento de los datos

 1.3.1. Estructura de los datos y variables de estudio

 1.3.2. Criterios de calidad de los datos

1.4. Metodología

2. ÁLGEBRA DE LOS DATOS

2.1. Limpieza y filtrado de datos a través de métricas y distancias adecuadas

2.2. Identificación y remoción de outliers

2.3. Aplicación de valores y vectores propios

2.4. Normas y determinantes para calcular varianzas globales

2.5. Productos internos para identificación de asociaciones lineales entre variables.

2.7. Número condición de las matrices de covarianzas

3. ESTADÍSTICA DE LOS DATOS

3.1. Descripción de datos multivariante

3.2. Modelos Metodología

3.3. Modelos de clasificación supervisados: Predicción de precipitaciones

3.4. Modelos de clasificación no supervisados

4. CICLO DE VIDA DE LOS DATOS Y PROCESAMIENTO ANALÍTICO

4.1. Descripción del sistema y enfoque propuesto

4.2. Arquitectura

5. CONCLUSIONES

6. REFERENCIAS BIBLIOGRÁFICAS

1. INTRODUCCIÓN

Los sistemas de alerta temprana (SAT) son herramientas críticas en la gestión de riesgos que permiten la identificación, monitoreo y alerta ante amenazas inminentes para proteger la vida humana y los bienes materiales. De acuerdo con la Oficina de las Naciones Unidas para la Reducción del Riesgo de Desastres (UNDRR), un SAT efectivo depende de la recolección y el análisis continuo de datos, además de la comunicación oportuna de la información a la ciudadanía. El Sistema de Alerta Temprana del Valle de Aburrá (SIATA) es un proyecto estratégico del Área Metropolitana que integra tecnologías avanzadas de monitoreo y modelación hidrometeorológica para predecir eventos naturales y antrópicos que pueden afectar las condiciones ambientales del Valle de Aburrá, o puedan generar riesgos a la población; su función es clave para la gestión de riesgos en la región, pues permite la toma de decisiones basada en datos en tiempo real y análisis predictivo, promoviendo así la participación ciudadana, la apropiación social de la ciencia, la tecnología y el conocimiento del riesgo por parte de la ciudadanía.

En el contexto de riesgos ambientales, el Big Data es esencial para procesar información en tiempo real, lo que facilita la detección temprana de eventos críticos y la generación de alertas. Sin embargo, su manejo también plantea desafíos relacionados con la calidad y accesibilidad de los datos. Para el SIATA, la implementación de un sistema de Big Data permitirá procesar grandes volúmenes de datos ambientales y meteorológicos de manera eficiente, mejorando así la precisión de los modelos predictivos.

El Big Data se caracteriza por el manejo de grandes volúmenes de datos, los cuales suelen ser diversos, generarse a gran velocidad, y requerir procesos exhaustivos de validación, limpieza y calificación, para garantizar su valor en aplicaciones analíticas. Gartner define el Big Data como "activos de información de gran volumen, velocidad y variedad que demandan formas de procesamiento innovadoras para mejorar el entendimiento y la toma de decisiones".

Para el análisis de riesgos ambientales los modelos predictivos son fundamentales, ya que permiten anticipar eventos futuros basándose en datos históricos. Es por ello que existen diferentes herramientas estadísticas que permiten desarrollar modelos predictivos para, posteriormente, identificar patrones complejos en los datos y establecer relaciones causales entre variables ambientales.

En el SIATA, el uso de algoritmos de machine learning mejorará la predicción de variables ambientales, lo que facilitará la toma de decisiones proactivas en la gestión de riesgos. Estos modelos podrán analizar patrones asociados a fenómenos climáticos, facilitando así la generación de alertas tempranas que protejan a la población y el medio ambiente.

1.1. MARCO TEÓRICO

Las crecidas torrenciales son un tipo de inundación caracterizada por un caudal máximo relativamente alto y de corta duración, asociado generalmente a la incidencia de una alta intensidad de precipitación sobre un área (NOAA, 2012). Estos eventos tienen ocurrencia en una escala temporal relativamente corta, por lo que los tiempos para tomar acciones de respuesta por parte de los organismos de socorro y las comunidades en condición de vulnerabilidad por su grado de exposición, son bastante limitados, lo que podría constituir una amenaza súbita por la materialización del riesgo que oca-sionaría pérdidas mortales y desastres.

Medellín ha sido una ciudad altamente afectada por la frecuente ocurrencia de even-tos extremos de precipitación, que han generado múltiples despliegues de alertas por inundaciones y desbordamientos del río Aburrá - Medellín y sus quebradas afluentes. Esta situación recurrentemente ha desencadenado múltiples emergencias como des-plomes de árboles, accidentes de tránsito y problemas de movilidad en zonas reporta-das por encharcamientos e inundaciones de deprimidos viales, deslizamientos y dete-rioros estructurales (UNGRD).

En respuesta a estos eventos, las autoridades locales han implementado los Sistemas de Alertas Tempranas, con el fin de instrumentar el río y las quebradas afluentes de los 10 municipios del Valle de Aburrá, con una amplia red de estaciones hidrometeoroló-gicas que permiten monitorear permanentemente las condiciones de precipitación y los niveles del río en las zonas urbanas y corregimientos de la ciudad.

Por consiguiente, este proyecto propone el análisis predictivo de variables hidrometeo-rológicas que tienen fuerte influencia en el comportamiento del nivel y el cauce del río Medellín, mediante modelos estadísticos que permiten determinar tendencias futuras y patrones de cambio que podrían traducirse en el incremento del riesgo por la ocu-rrencia de eventos extremos de precipitación.

1.2. COMPRENSIÓN DEL NEGOCIO

1.2.1. Entendimiento del problema

Debido a la gran densidad de información que es generada y almacenada continuamente desde el proyecto SIATA, se evidencian diferentes necesidades orientadas a mejorar la accesibilidad a los datos mediante el uso de plataformas de Big Data, la implementación de técnicas de revisión y limpieza de los datos actuales, y el desarrollo de estrategias de visualización y metodologías de análisis predictivo, usando algoritmos de machine learning y herramientas de aprendizaje estadístico para predecir tendencias futuras a partir de los datos históricos, entre otras aplicaciones. Esto, con el objetivo de lograr la entrega eficaz de información a las instituciones y la ciudadanía como actores claves y responsables en la gestión integral de riesgos.

Comprendiendo el problema y las necesidades identificadas, la metodología implementada para dar solución a la pregunta objetivo busca establecer predicciones en las condiciones meteorológicas de Medellín y a partir del modelado de variables ambientales es posible identificar la probabilidad de evidenciar un aumento o disminución en el nivel del río Aburrá - Medellín.

1.2.2. Dominio del estudio

El área de estudio comprende la delimitación del distrito especial de Medellín en tres zonas específicas y con diferentes coberturas asociadas al uso del suelo, cuyas coordenadas corresponden a 6°14'41" sobre la latitud Norte y 75°34'29" en longitud Oeste (figura 1). La región Medellín Oriente comprende el corregimiento de Santa Elena, caracterizada por una mayor cobertura vegetal entre bosque y pasto, mientras que la región definida como Medellín Centro se caracteriza por una heterogeneidad entre cobertura vegetal y urbana en la zona de la cuenca, y comprende el casco urbano del distrito en sus 16 comunas. Por otro lado, la región Medellín Occidente abarca las zonas urbanas y rurales de los corregimientos San Sebastián de Palmitas, Altavista, San Antonio de Prado y San Cristóbal, las cuales son mayormente cubiertas por vegetación boscosa y pastizales. Además, se complementa el dominio con la subcuenca del río Aburrá - Medellín, delimitada en el punto de medición localizado en el Puente de la 33 con la Avenida Regional, en la comuna 10 – La candelaria (figura 2).

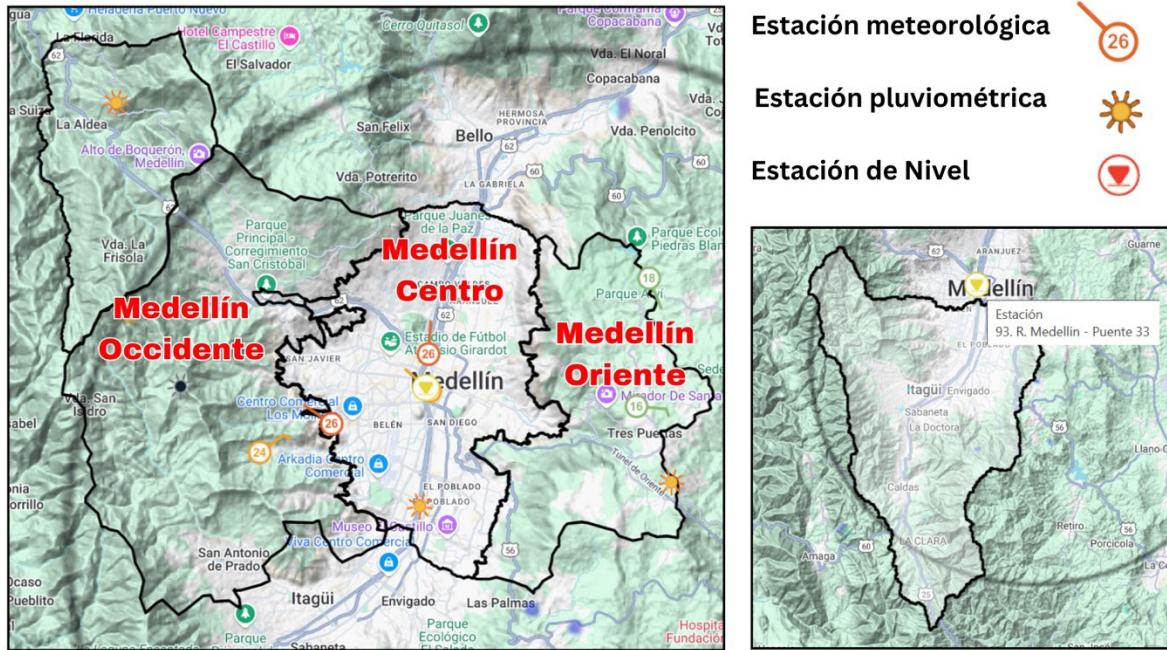


Figura 1. Área de estudio y ubicación estaciones de monitoreo



Figura 2. Punto de medición estación de nivel Puente 33. Fuente: SIATA

1.3. ENTENDIMIENTO DE LOS DATOS

Los datos suministrados por SIATA provienen de información histórica de 17 estaciones pluviométricas, 13 estaciones meteorológicas y 19 estaciones de nivel, las cuales se encuentran distribuidas a lo largo de los 10 municipios del Valle de Aburrá. Estos datos han sido revisados bajo el criterio de calidad exhaustiva hasta el mes de julio de 2024. Además, algunas de las estaciones cuentan con registros desde los años 2013 y 2018, por lo tanto, se sugiere la implementación de técnicas de limpieza y filtrado para que exista una homogeneidad en el dominio temporal y espacial del estudio.

SIATA cuenta con una amplia red de monitoreo compuesta por sensores que hacen parte de las diferentes redes de monitoreo hidrometeorológico. Para la finalidad de este estudio, se usaron datos de los sensores en tierra de las redes de estaciones pluviométricas, meteorológicas y nivel. Por lo tanto, se resumen las características de dichos datos y la metainformación asociada a las variables objeto de estudio.

Preliminarmente, se eligieron las estaciones de acuerdo a su ubicación, con el objetivo de obtener información representativa de cada uno de los municipios del Valle de Aburrá. Posteriormente, se realizó un proceso más riguroso de selección en concordancia con el objetivo de este proyecto, la metodología a implementar y las consideraciones en los criterios de calidad, para garantizar representatividad, uniformidad y homogeneidad en los datos. Por consiguiente, luego de aplicar estrategias de filtrado, limpieza y transformación de los datos crudos, se escogieron 4 estaciones pluviométricas, 6 estaciones meteorológicas y una estación de nivel, todas localizadas en jurisdicción del distrito de Medellín. El período de estudio seleccionado está comprendido entre las 14:45:00 del 2022-06-29 y las 10:30:00 del 2024-07-19, en intervalos de 15 minutos ($\Delta t = 15$). A continuación, se describen las consideraciones acerca de la funcionalidad de los diferentes tipos de sensores, su distribución espacial y la información de las mediciones de las variables que se usaron en este estudio.

Estaciones pluviométricas. Están compuestas por un sensor in situ que mide volúmetricamente la cantidad de precipitación que cae en un punto específico (figura 3), durante un instante de tiempo determinado. Esta red pluviométrica es una de las más densas y con una amplia distribución en los 10 municipios del Valle de Aburrá (figura 4), genera datos cada minuto y partir de sus mediciones, es posible calcular la cantidad de la lluvia acumulada (mm) en diferentes escalas temporales (horaria, diaria, semanal y mensual) y la intensidad de la lluvia (mm/h).



Figura 3. Estación pluviométrica. Fuente: SIATA

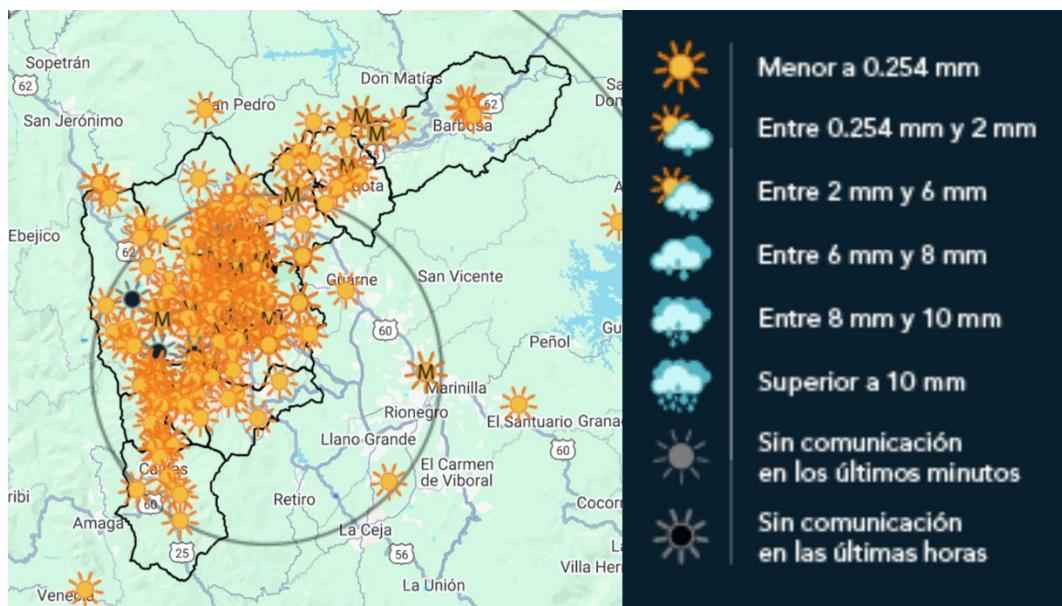


Figura 4. Distribución espacial de estaciones pluviométricas

Estaciones meteorológicas. Son estaciones in situ (figura 5) que monitorean de manera simultánea el comportamiento de variables meteorológicas como temperatura del aire en superficie, humedad relativa, presión atmosférica, precipitación, radiación, velocidad y dirección superficial del viento. La red meteorológica de SIATA cuenta con 44 estaciones de monitoreo localizadas en zonas estratégicas de la jurisdicción metropolitana (figura 6).

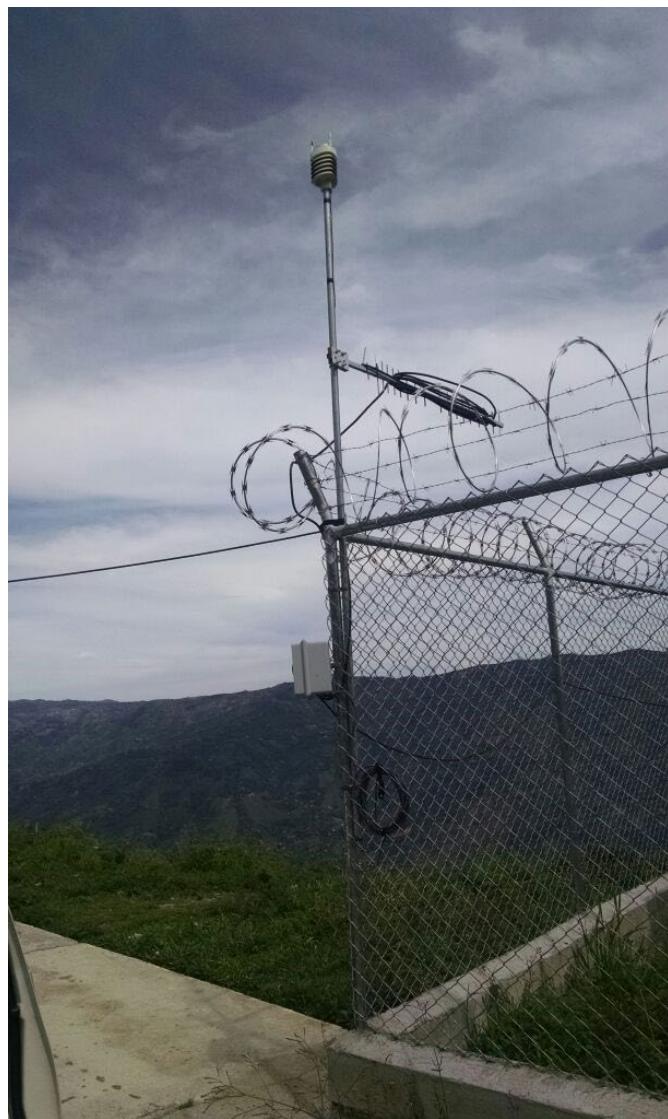


Figura 5. Estación meteorológica. Fuente: SIATA

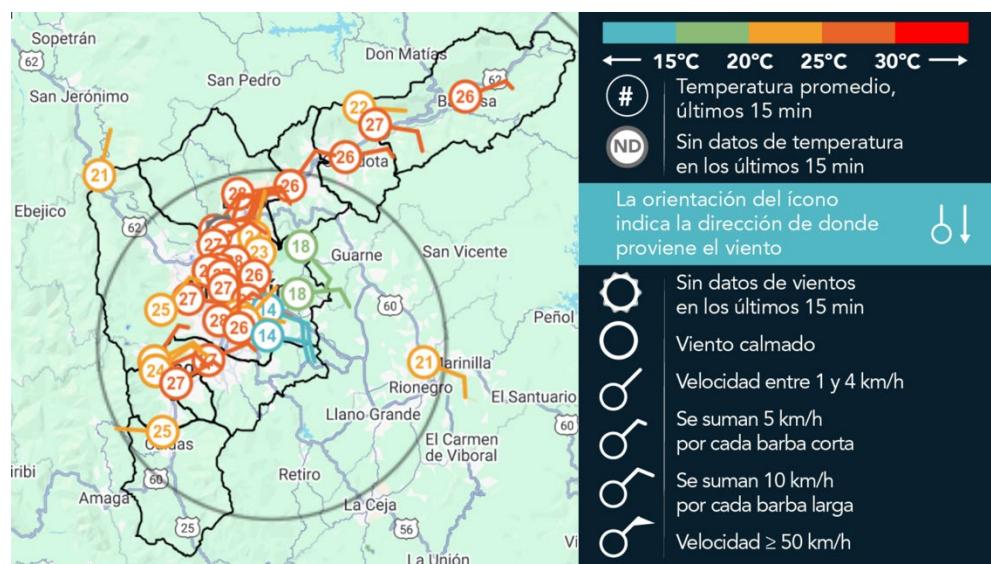


Figura 6. Distribución espacial de estaciones meteorológicas. Fuente: SIATA

Estaciones de nivel. Miden los cambios en el nivel del río Aburrá-Medellín (figura 8) y sus principales afluentes y quebradas del territorio; sin embargo, para fines prácticos, se eligió únicamente la estación ubicada sobre el cauce del río Aburrá, en el punto de monitoreo localizado en el Puente de la 33 con la Avenida Regional. La medición de estos sensores indica los niveles de riesgo asociados a una escala de colores, que van desde un nivel seguro hasta el que indica una posible inundación mayor o extensiva (figura 7).

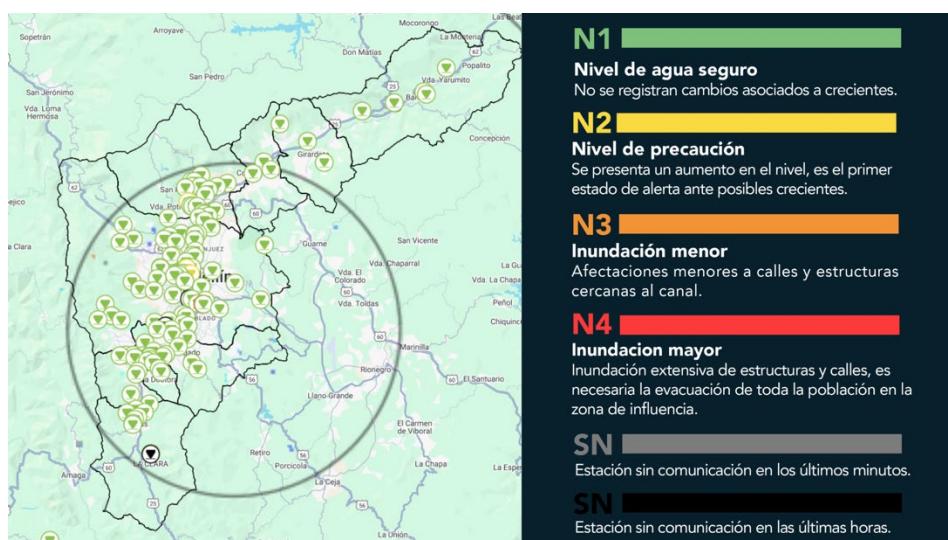


Figura 7. Distribución espacial de estaciones de nivel. Fuente: SIATA



Figura 8. Estación de nivel R. Aburrá-Medellín. Fuente: SIATA

1.3.1. Estructura de los datos y variables de estudio

A continuación, se presenta la estructura de los datos y las variables de estudio. Considerando, además, que se subdividió el dominio de estudio en 3 subzonas, las estaciones seleccionadas a partir de su ubicación y período de registro son 4 estaciones pluviométricas, 6 estaciones meteorológicas y una estación de nivel (Tabla 1).

Tabla 1. Tipo de estación y agrupamiento por subzona

# CÓDIGO ESTACIÓN	NOMBRE	TIPO DE ESTACIÓN	MUNICIPIO	SUBZONA
619	Cabaña Musical	Pluviometrica	Medellin Santa Elena	Oriente
311	Casa SIATA	Pluviometrica	Medellin	Centro
25	Escuela Rural Astilleros	Pluviometrica	Medellin San Antonio de Prado	Occidente
4	I.E Hector Rogelio Montoya	Pluviometrica	Medellin Palmitas	Occidente
93	Río Medellín Puente 33	Nivel	Medellin Perpetuo Socorro	Centro
419	Thies SENA Medellin	Meteorologica	Medellin	Centro
355	Escuela Piedras Gordas	Meteorologica	Medellin Santa Elena	Oriente
249	Escuela CEDEPRO	Meteorologica	Medellin Altavista	Occidente
207	Vivero EPM Piedras Blancas	Meteorologica	Medellín Oriente	Oriente
202	AMVA	Meteorologica	Medellin	Centro
197	Universidad de Medellín	Meteorologica	Medellín Occidente	Occidente

Estaciones pluviométricas. Las variables **p1** y **p2** (figura 9), que proveen estas estaciones, corresponden a los registros de precipitación para los pluviómetros 1 y 2, minuto a minuto. Sin embargo, en el proceso de filtrado, limpieza y homogeneización de los datos se creó una nueva columna llamada **p_sum** (figura 10), que agrupa los datos de ambas variables, asumiendo la validez de la medición cuando ambas son iguales y cuando, al ser diferentes, al menos una de las dos es mayor que cero. Dada la naturaleza de los datos y considerando que la precipitación se acumula en un intervalo de tiempo, siguiendo los objetivos del análisis, los datos de **p_sum** se acumularon en 15 minutos, obteniendo datos de precipitación acumulada en las unidades de **mm/min**.

```
[10] 1 pluviograficas.head()
```

	codigo	fecha_hora	p1	p2	calidad
0	25	2013-09-12 12:51:43	0.0	0.0	1.0
1	25	2013-09-12 12:52:43	0.0	0.0	1.0
2	25	2013-09-12 12:53:43	0.0	0.0	1.0
3	25	2013-09-12 12:54:43	0.0	0.0	1.0
4	25	2013-09-12 12:55:43	0.0	0.0	1.0

Figura 9. Datos estaciones pluviométricas sin preprocesamiento

	fecha_hora	codigo_y	p_sum
0	2022-06-29 14:45:00	25	0.0
1	2022-06-29 15:00:00	25	0.0
2	2022-06-29 15:15:00	25	0.0
3	2022-06-29 15:30:00	25	0.0
4	2022-06-29 15:45:00	25	0.0

Figura 10. Datos estaciones pluviométricas limpios y procesados

Estaciones meteorológicas. Las variables **h**, **t**, **pr**, **vv**, **vv_max**, **dv**, **dv_max**, **p**, son los registros minuto a minuto de humedad relativa (%), temperatura superficial del aire (°C), presión atmosférica (hPa), velocidad del viento (m/s), velocidad máxima del viento (m/s), dirección del viento en grados, dirección del viento máximo en grados, precipitación (mm), respectivamente (figura 11). Estas variables también fueron preprocesadas para obtener valores promediados en intervalos de 15 minutos y consecuentemente, se crearon variables adicionales con métricas calculadas, tales como la desviación estándar y el promedio (figuras 12 y 13).

```
1 meteorologicas.head()
```

	codigo	fecha_hora	h	t	pr	vv	vv_max	dv	dv_max	p	calidad
0	419	2019-12-26 15:59:00	37.0	28.80	843.73	8.54	11.1	96.0	2.0	0.0	2
1	419	2019-12-26 16:00:00	37.0	28.73	843.73	7.08	9.7	102.0	288.0	0.0	2
2	419	2019-12-26 16:01:00	37.0	28.76	843.80	4.94	8.1	93.0	206.0	0.0	2
3	419	2019-12-26 16:02:00	37.0	28.71	843.80	6.70	8.4	90.0	83.0	0.0	2
4	419	2019-12-26 16:03:00	37.0	28.70	843.83	6.66	8.7	89.0	59.0	0.0	2

Figura 11. Datos estaciones meteorológicas sin preprocessamiento

	codigo_x	fecha_hora	h_promedio	h_std	h_sum	t_promedio	t_std	t_sum	pr_promedio
0	197	2022-06-29 14:45:00	60.322667	0.593506	904.84	24.714667	0.105076	370.72	842.491333
1	197	2022-06-29 15:00:00	59.664667	1.366768	894.97	24.759333	0.070353	371.39	842.331333
2	197	2022-06-29 15:15:00	59.628000	0.709368	894.42	24.617333	0.112787	369.26	842.247333
3	197	2022-06-29 15:30:00	59.266667	0.651050	889.00	24.583333	0.106413	368.75	842.081333
4	197	2022-06-29 15:45:00	61.250000	0.883338	918.75	24.980000	0.165357	374.70	841.872000

Figura 12. Datos estaciones meteorológicas limpios y procesados

	fecha_hora	pr_sum	vv_promedio	vv_std	vv_sum	vv_max_promedio	vv_max_std	vv_max_sum
0	2022-06-29 14:45:00	12637.37	1.749333	0.483447	26.24	3.026667	0.509154	45.4
1	2022-06-29 15:00:00	12634.97	1.248667	0.530093	18.73	2.206667	0.743031	33.1
2	2022-06-29 15:15:00	12633.71	1.646000	0.515805	24.69	2.653333	0.720978	39.8
3	2022-06-29 15:30:00	12631.22	1.686667	0.447416	25.30	2.626667	0.527076	39.4
4	2022-06-29 15:45:00	12628.08	0.842000	0.406381	12.63	1.640000	0.642317	24.6

Figura 13. Datos estaciones meteorológicas limpios y procesados

Estaciones de nivel. La variable nivel provee datos en cm de la profundidad del cauce y altura de la lámina de agua del río Medellín en el punto de monitoreo seleccionado; por lo tanto, para facilitar el análisis de acuerdo con los objetivos del proyecto, se dividió por 100 en el preprocesamiento para considerar unidades de m (figuras 14 y 15).

```
1 nivel.head()
```

	codigo	fecha_hora	nivel	calidad
0	342	2018-12-04 13:35:00	395.450012	1
1	342	2018-12-04 13:36:00	395.390015	1
2	342	2018-12-04 13:37:00	395.899994	1
3	342	2018-12-04 13:38:00	395.709991	1
4	342	2018-12-04 13:39:00	395.769989	1

Figura 14. Datos estaciones de nivel sin preprocesamiento

	fecha_hora	codigo	nivel_metros_promedio
0	2022-06-29 14:45:00	93	0.597520
1	2022-06-29 15:00:00	93	0.596887
2	2022-06-29 15:15:00	93	0.590020
3	2022-06-29 15:30:00	93	0.584540
4	2022-06-29 15:45:00	93	0.581833

Figura 15. Datos estaciones de nivel limpios y procesados

1.3.2. Criterios de calidad de los datos

Para el preprocesamiento de los datos se tuvieron en cuenta los criterios de calidad dados por las generalidades de la información suministrada por SIATA, considerando que los archivos que contienen los registros pluviométricos están en formato de texto plano y contienen: en la primera columna la fecha y la hora del registro, y en la segunda, el registro de precipitación en unidades de mm del pluviómetro 1 (o sensor de lluvias, en caso de ser sensor multiparamétrico), y en la siguiente columna, se tiene el registro de precipitación en unidades de mm del pluviómetro 2 (en caso de que la estación cuente con dos pluviómetros).

En los archivos, el registro –999 corresponde con un dato faltante y en la última columna se consigna el índice de calidad del registro. La tabla a continuación resume el significado de los indicadores de calidad de estos datos (tabla 2), teniendo en cuenta que para el proceso de limpieza del dataset se consideraron únicamente los datos con un índice de calidad de 1 y 2, correspondientes a una Calidad confiable del dato en tiempo real y una Calidad confiable del dato no obtenido en tiempo real.

Los índices de calidad usados en la red meteorológica son acumulativos, ya que en una medida se tiene información de todas las variables y se conforman de la siguiente manera:

- El primer dígito indica la procedencia del dato, puede ser “1” si el dato fue obtenido en tiempo real o “2” si el dato ingresó por importación. Si el dato es bueno el índice solo va a contener este dígito.
- El segundo dígito es “5”, que indica calidad dudosa.

- Luego se agregan dígitos que correspondan de acuerdo con las variables que tengan medición dudosa, si solo es una se adicionaría sólo un dígito.
- Cuando más de una variable tiene calidad dudosa se ponen los dígitos que corresponden a la calidad dudosa de cada una de estas, es decir, es posible encontrar dígitos como “1534” (calidad dudosa de temperatura y humedad relativa) o “1515” (calidad dudosa de precipitación y presión atmosférica).
- Los índices luego del “5” siempre se ponen de menor a mayor y pueden ser máximo 3 dígitos.
- Un índice de calidad de “157” indica que tanto el dato de la dirección del viento promedio como el de la dirección del viento máximo son dudosos.
- Si 4 ó más variables tienen datos dudosos, el índice de calidad en este caso sería “151”.

Tabla 2. Criterios de calidad de la red pluviométrica

Caso-Descripción	Índice
Calidad confiable del dato en tiempo real	1
Calidad confiable del dato no obtenido en tiempo real	2
Calidad dudosa del dato en tiempo real en ambos pluviómetros	151
Calidad dudosa en dato del pluviómetro 1 en tiempo real	1511
Calidad dudosa en dato del pluviómetro 2 en tiempo real	1512
Calidad dudosa en dato no obtenido en tiempo real en ambos pluviómetros	251
Calidad dudosa en dato no obtenido en tiempo real del pluviómetro 1	2511
Calidad dudosa en dato no obtenido en tiempo real del pluviómetro 2	2512
Calidad dudosa en dato del pluviómetro 1 en tiempo real comparado con radar meteorológico	15110
Calidad dudosa en dato del pluviómetro 2 en tiempo real comparado con radar meteorológico	15120
Calidad dudosa en datos de ambos pluviómetros en tiempo real comparado con radar meteorológico	1510

Tabla 3. Criterios de calidad de la red meteorológica

Caso-Descripción	Índice
Calidad confiable del dato en tiempo real	1
Calidad confiable del dato no obtenido en tiempo real	2
Calidad dudosa en dato de precipitación en tiempo real	1511
Calidad dudosa en dato de temperatura en tiempo real	153
Calidad dudosa en dato de humedad relativa en tiempo real	154
Calidad dudosa en dato de presión atmosférica en tiempo real	155
Calidad dudosa en dato de magnitud de viento en tiempo real	156
Calidad dudosa en dato de magnitud de viento promedio en tiempo real	1561
Calidad dudosa en dato de magnitud de viento máximo en tiempo real	1562
Calidad dudosa en dato de dirección del viento en tiempo real	157
Calidad dudosa en dato de dirección del viento promedio en tiempo real	1571
Calidad dudosa en dato de dirección del viento máximo en tiempo real	1572
Calidad dudosa en datos de todas las variables	151

1.4 METODOLOGÍA

1.4.1. Comprensión del negocio

- **Objetivo del proyecto.** Mejorar el acceso, la gestión y el análisis predictivo de datos hidrometeorológicos para la generación de alertas tempranas y la toma de decisiones en el Valle de Aburrá.
- **Requisitos del sistema.** Implementar un sistema híbrido batch-streaming, diseñar modelos predictivos enfocados en precipitaciones y niveles del río, desarrollar un catálogo de datos abiertos, mejorar la accesibilidad mediante interfaces eficientes.
- **Factores críticos de éxito.** Precisión de los modelos predictivos, alta disponibilidad y calidad de datos, capacidad de los usuarios (ciudadanos e instituciones) para interpretar la información.

1.4.2. Comprensión de los datos

- **Fuentes de datos.** Estaciones pluviométricas, meteorológicas y de nivel, datos históricos del SIATA.
- **Descripción inicial.** Variables clave: Precipitación acumulada, humedad, temperatura, presión atmosférica, nivel del río. Intervalos de tiempo: Datos registrados cada 15 minutos.

- **Evaluación de la calidad.**
 - **Compleción:** Verificar datos faltantes e imputar valores usando regresión lineal y normalización.
 - **Consistencia:** Validar que los registros cumplan con los estándares del SIATA.
 - **Duplicación:** Identificar y eliminar duplicados en los datos almacenados.

1.4.3. Preparación de los datos

- **Limpieza.** Eliminación de datos atípicos utilizando métricas como IQR y distancias multivariantes (Mahalanobis, Euclídea), homogeneización temporal y espacial de los datos.
- **Transformación.** Conversión de datos minuto a minuto a intervalos de 15 minutos, normalización y creación de nuevas variables relevantes (promedios y desviaciones estándar).
- **Integración.** Consolidación de datasets de distintas estaciones en un único conjunto estructurado por subzonas.

1.4.4. Modelado

- **Modelos planteados.**
 - **Clasificación supervisada:** Modelos como CatBoost y Random Forest para predecir lluvia basados en la región oriental.
 - **Regresión continua:** LightGBM y Random Forest Regression para estimar el nivel promedio del río.
- **Selección de características.** Variables predictoras seleccionadas a partir de análisis de correlación (e.g., precipitación acumulada, humedad).
- **Entrenamiento y validación.**
 - División de datos: 70% entrenamiento, 30% prueba.
 - Balanceo de datos con SMOTE para evitar sesgos en las predicciones.

1.4.5. Evaluación

- **Métricas de rendimiento.** Para modelos de clasificación: Precisión, recall, F1-score, y área bajo la curva ROC. Para modelos de regresión: Error cuadrático medio (MSE) y coeficiente de determinación (R^2).
- **Validación cruzada:** Uso de k-folds para asegurar la robustez del modelo.
- **Iteraciones:** Optimización de hiperparámetros y evaluación comparativa entre modelos.

1.4.6. Despliegue

- **Integración tecnológica.** Uso de Amazon S3, EC2 y Athena para almacenamiento y consultas. Visualización a través de herramientas como Tableau o QuickSight.
- **Sistema de ingesta de datos.**
 - **Batch:** Procesamiento de datos históricos.
 - **Streaming:** Análisis en tiempo real para generar alertas inmediatas.
- **Accesibilidad.** Implementación de APIs para acceso a los datos procesados.

1.4.7. Monitoreo y mantenimiento

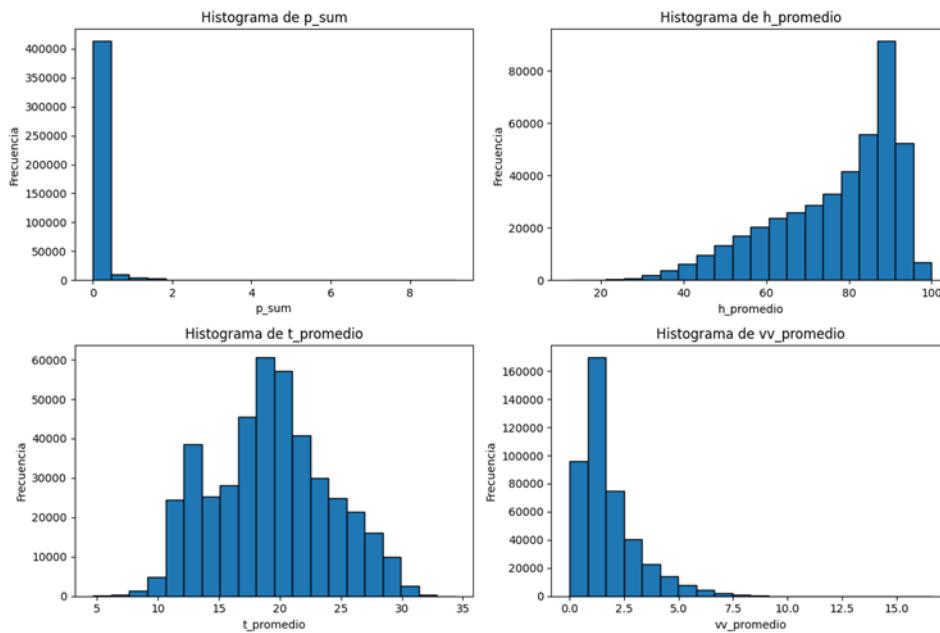
- **Monitoreo del modelo.** Actualización periódica con nuevos datos para mantener la precisión.
- **Calidad de datos.** Validación continua en la etapa de extracción.
- **Capacitación.** Entrenamiento de usuarios finales en la interpretación de los resultados.

2. ÁLGEBRA DE LOS DATOS

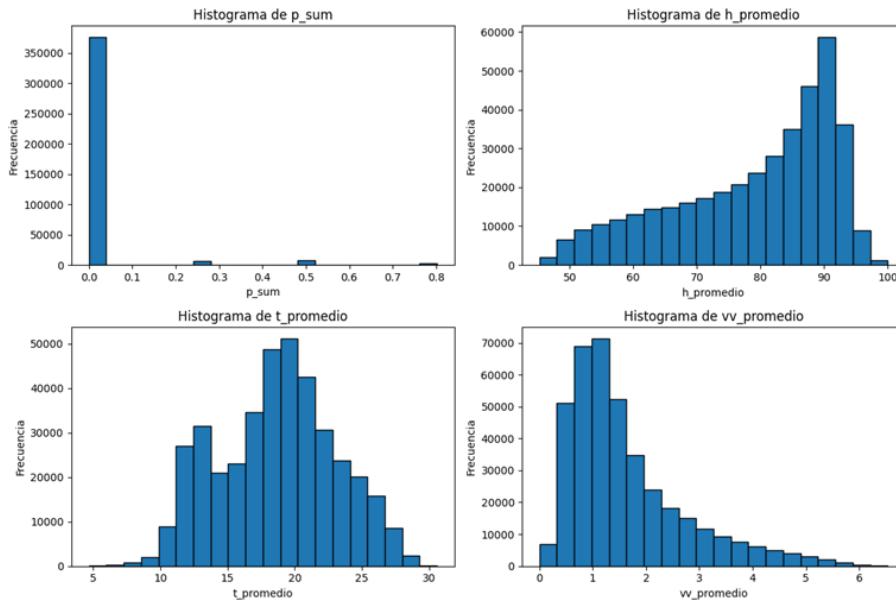
2.1. LIMPIEZA Y FILTRADO DE DATOS A TRAVÉS DE MÉTRICAS Y DISTANCIAS ADECUADAS

Del dataset original se tomaron las variables 'p_sum', 'h_promedio', 't_promedio', 'vv_promedio', para aplicar la técnicas y métricas para limpieza de datos, el total de observaciones de las variables son 432.480.

Se genero histograma para observar la distribución inicial de cada una de las variables.



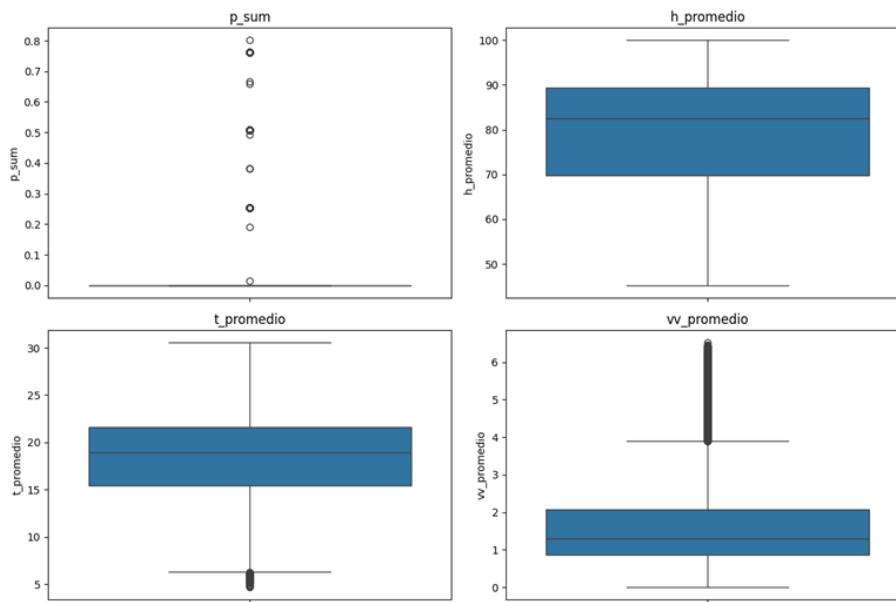
A este set de datos se aplicarán las distancias Euclidiana, Mahalanobis y Coseno; la primera la utilizaremos para agrupar puntos similares y observaciones (clustering), luego con Mahalanobis se establecerán puntos que tienen distancia considerable del centro del cluster, teniendo en cuenta la correlación entre variables y con la de Coseno se establecerá que tan similares son los vectores de los datos clusterizados. Al combinar estas tres técnicas, permite obtener mayor robustez y calidad de los datos, también permite mantener patrones significativos en los datos, de igual forma elimina puntos que no estén ajustados a la estructura estadística de los datos, toda esta técnica se implementó mediante clusterización de K_Means.



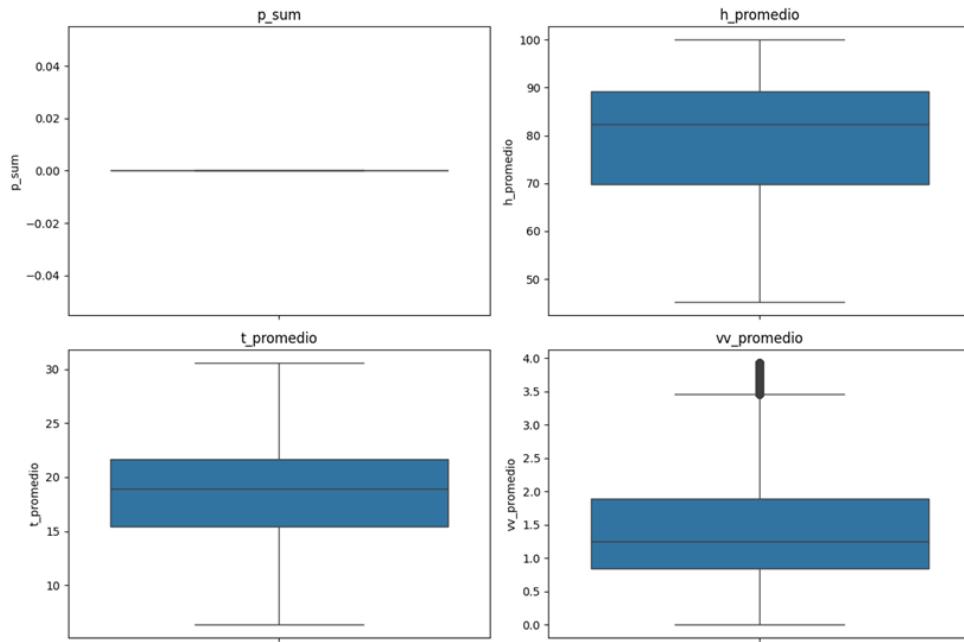
Luego de aplicar las técnicas se redujo el dataset de 432.480 a 393.017 observaciones, y obtenemos el siguiente grafico con el histograma de los datos

2.2. IDENTIFICACIÓN Y REMOCIÓN DE OUTLIERS

Seguidamente generamos boxplots con el fin de identificar datos atípicos



Como podemos observar en las variables p_sum, t_promedio y vv_promedio aún persisten datos atípicos por lo que se procedió a realizar la eliminación de estos, teniendo en cuenta un rango de $1.5 * \text{IQR}$ hacia arriba y abajo, obteniendo que el datasets quedara con 354.973 observaciones; podemos observar la eliminación de datos atípicos en el siguiente gráfico.

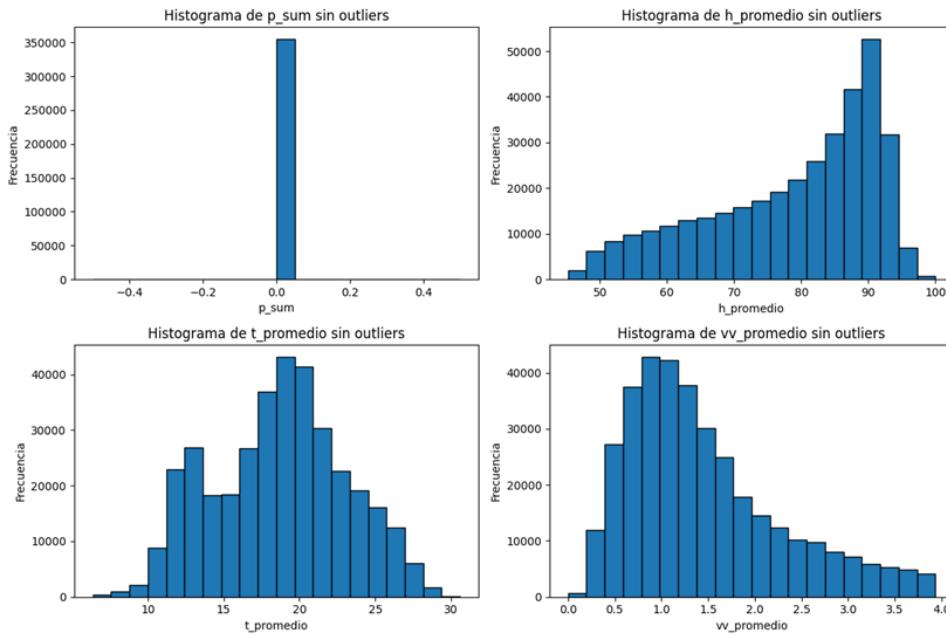


Y mostramos como quedó la distribución de los datos en el siguiente histograma

2.3. APLICACIÓN

DE
VA-
RES
VEC-
RES
PIOS

CA-
CIÓN
LO-
Y
TO-
PRO-



Calculamos la matriz de covarianza de los datos, con el fin encontrar los valores y vectores propios asociados.

Matriz de covarianza:

```
[[ 1.00000231  0.16824429 -0.11471115 -0.05155006]
 [ 0.16824429  1.00000231 -0.76958931 -0.31135749]
 [-0.11471115 -0.76958931  1.00000231  0.27143784]
 [-0.05155006 -0.31135749  0.27143784  1.00000231]]
```

Seguidamente calculamos los valores y vectores propios asociados a la matriz y seleccionamos los vectores propios que tiene mayor explicación de la varianza y realizamos la proyección de los datos al espacio definido.

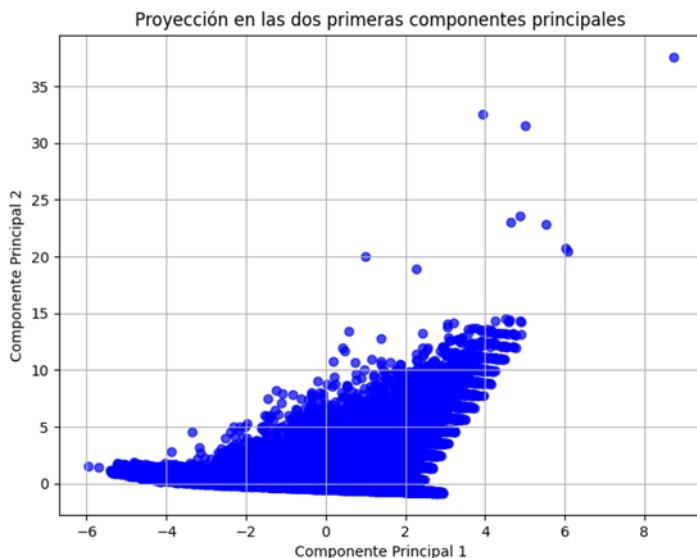
```
Valores propios:
[1.9908779  0.22728341  0.96608609  0.81576186]

Vectores propios:
[[ 0.20223842  0.05067989  0.95705195 -0.20145159]
 [ 0.64406148 -0.7181459 -0.04335393  0.25994562]
 [-0.62917578 -0.69273322  0.09838638 -0.33849467]
 [-0.38525612 -0.04264623  0.26924363  0.88162741]]
```

```
Vectores propios seleccionados:
[[ 0.20223842  0.95705195]
 [ 0.64406148 -0.04335393]
 [-0.62917578  0.09838638]
 [-0.38525612  0.26924363]]
```

Podemos observar en el grafico que los puntos se distribuyen tomando un patrón lineal ascendente, que indica correlación entre las dos componentes proyectadas; cabe anotar que se tienen puntos que están más dispersos o alejados, que pueden dar a entender la presencia de datos atípicos o con variaciones significativas en los datos originales. Además, podemos observar que los dos componentes proyectados en conjunto explican una alta proporción de la varianza de los datos.

```
Proporción de varianza explicada por cada componente:
[0.49771832 0.24152096 0.20393999 0.05682072]
```



2.4. NORMAS Y DETERMINANTES PARA CALCULAR VARIANZAS GLOBALES

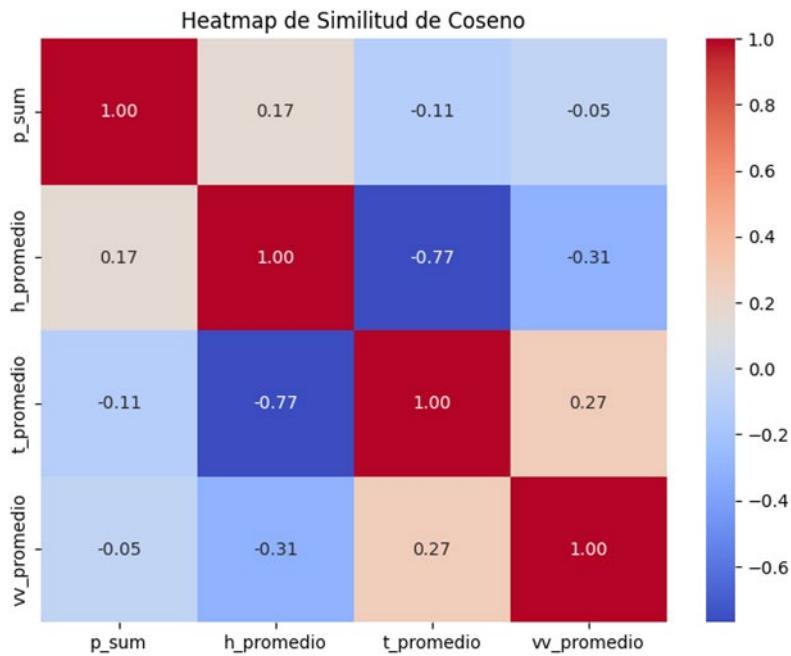
Al calcular las normas y determinantes nos permiten comprender la variabilidad de los datos. En este sentido calcularemos la norma de Frobenius y la norma 2, de igual forma calcularemos el determinante y finalmente la varianza global. En este sentido los resultados obtenidos son los siguientes

```
Norma Frobenius: 2.369397032435001
Norma 2 : 1.9908779019224114
Determinante de la matriz de covarianza: 0.356608405444004
Varianza global: 4.000009249004004
```

Para la norma Frobenius, que nos indica la magnitud total de la matriz de covarianza considerando todos los elementos, podemos interpretar que las variables estudiadas pueden tener una variabilidad moderada en el espacio multivariado. Por su parte la norma 2 que representa la varianza máxima explicada por cualquier componente principal, podemos decir que la mayor parte de la varianza está dominada por la dirección específica en el espacio de la primera componente principal. En tanto, el determinante nos indica que tanto están dispersos los datos en el espacio multivariado, el resultado nos sugiere que hay cierta correlación entre las variables, lo que significa que no todas varían de manera independiente; finalmente el resultado de la varianza global nos indica que las variables están contribuyendo de forma muy similar a la varianza total.

2.5. PRODUCTOS INTERNOS PARA LA IDENTIFICACIÓN DE ASOCIACIONES LINEALES ENTRE VARIABLES.

El cálculo de los productos internos nos permite medir la asociación lineal entre dos variables, en este caso calculamos los productos internos y verificamos las relaciones lineales con la similitud de coseno, obteniendo los siguientes resultados.



Observamos que existe una baja similitud negativa entre la variable p_sum respecto a t_promedio y vv_promedio, lo que sugiere que estas variables son independientes, en tanto respecto a h_promedio se tiene una similitud baja, pero positiva; es decir que, si una aumenta la otra también, pero en menor proporción. Ahora bien, entre h_promedio y t_promedio existe fuerte relación inversa, lo que denota que si una sube la otra disminuye.

2.6. APLICACIÓN DE LA DESCOMPOSICIÓN EN VALORES SINGULARES

La descomposición en valores singulares nos permite reducir la dimensionalidad de los datos, indicando las direcciones de mayor variabilidad en los datos, al aplicar esta técnica al dataset obtenemos lo siguiente.

```

Valores singulares (importancia de cada componente):
[927.90779938 646.38374379 593.9695882 313.52081256]
Varianza explicada por cada componente:
[0.49771832 0.24152096 0.20393999 0.05682072]
Varianza total explicada: 1.0000000000000058
  
```

Podemos ver, de acuerdo a los resultados, que la mayor importancia relativa en torno a la información de los datos lo tiene el primer componente, el segundo y tercer componente la información que recoge cada uno de ellos es similar. en este sentido y como

se había visto en la parte de valores y vectores propios, entre el primer y segundo componente explican cerca del 75% de la varianza de los datos.

2.7. NÚMERO CONDICIÓN DE LAS MATRICES DE COVARIANZAS

El número de condición en la matriz de covarianza nos permite evaluar la calidad de los datos y la estabilidad de los cálculos, al realizar esta técnica a los datos obtenemos.

Número de condición de la matriz de covarianzas: 8.7594512143004

Resultado que puede indicar que las variables utilizadas no tienen una fuerte dependencia lineal o por decir que se tenga problemas de multicolinealidad entre ellas, pero se detalla que existe correlación entre ellas, baja o no tan fuerte.

En conclusión, luego de aplicar las técnicas de distancias y eliminación de datos atípicos se obtuvo una reducción del dataset de 432.480 a 354.973 observaciones. Adicionalmente cuando se realizaron cálculos para reducción de dimensionalidad, con los resultados de los valores propios y vectores propios, en análisis de las varianzas mediante las normas, el análisis de relaciones lineales y la descomposición de valores, se pudo establecer que se podía realizar dimensionalidad en dos componentes principales, los cuales en conjunto explican cerca del 75% de la varianza de los datos. Finalmente, la condición de la matriz de covarianza nos dice que no hay problemas de multicolinealidad entre las variables y que puede existir correlación baja entre ellas.

3. ESTADÍSTICA DE LOS DATOS

3.1. DESCRIPCIÓN DE LOS DATOS MULTIVARIANTES

El conjunto de datos con el que trabajaremos está dividido en tres principales categorías informativas: temperatura, nivel y pluviometría. El modelo predictivo que

realizaremos más tarde se concentra en las variables de naturaleza pluviométrica dado su objetivo de predecir las lluvias en distintas áreas de la ciudad. No obstante, antes de proceder con este modelo, haremos un primer análisis de todas las variables para entender sus estadísticas descriptivas y las correlaciones entre ellas.

Estadísticas descriptivas

El primer análisis de los datos que realizaremos consiste en obtener sus distintas estadísticas descriptivas.

TEMPERATURA

Estadísticas de temperatura:

	t_promedio	t_std	t_sum
count	432480.000000	432480.000000	432480.000000
mean	19.212334	0.122659	286.334174
std	4.785047	0.119411	74.980462
min	4.673333	0.000000	0.000000
25%	15.844667	0.048795	235.800000
50%	19.200000	0.088372	287.200000
75%	22.353333	0.162340	335.130000
max	34.374000	3.301616	515.610000

Temperatura promedio (t_promedio):

- Media: 19.21 °C, lo que indica un clima moderado.
- Rango: 4.67 °C (mínimo) a 34.37 °C (máximo), mostrando una variabilidad amplia.
- Distribución: El 50% de los datos (mediana) está alrededor de 19.20 °C, y la mayoría de los valores (entre el 25% y el 75%) se encuentra entre 15.84 °C y 22.35 °C.

Desviación estándar de temperatura (t_std):

- Media: 0.12, sugiere poca variabilidad en general.
- Máximo: 3.30, refleja eventos de alta fluctuación en algunos períodos.
- Mínimo: 0, indica intervalos sin variación.

Suma de temperatura (t_sum):

- Media: 286.33, acumulado típico por intervalo.
- Rango: 0 (mínimo) a 515.61 (máximo), refleja variaciones en el tiempo de observación.

PLUVIOMETRÍA

Estadísticas pluviométricas:

	pr_promedio	pr_std	pr_sum	p_promedio
count	432480.00000	432480.00000	432480.00000	432480.00000
mean	805.904208	0.634755	12011.451670	0.003079
std	87.828313	12.718727	1577.872601	0.015334
min	0.000000	0.000000	0.000000	0.000000
25%	766.691192	0.041404	11497.927500	0.000000
50%	830.898667	0.056061	12340.750000	0.000000
75%	848.553333	0.085060	12726.100000	0.000000
max	857.140000	528.845162	12857.100000	0.609600

	p_std	p_sum	p_min	p_max
count	432480.00000	432480.00000	4.324800e+05	432480.00000
mean	0.005720	0.046184	6.556582e-07	0.015209
std	0.023472	0.230007	2.542771e-04	0.060812
min	0.000000	0.000000	0.000000e+00	0.000000
25%	0.000000	0.000000	0.000000e+00	0.000000
50%	0.000000	0.000000	0.000000e+00	0.000000
75%	0.000000	0.000000	0.000000e+00	0.000000
max	0.524660	9.144000	1.270000e-01	1.778000

Promedio (pr_promedio y p_promedio):

- pr_promedio (precipitación acumulada): Media de 805.9; valores entre 0 y 857.14, indicando períodos secos y otros con fuertes lluvias.
- p_promedio: Valores bajos en promedio (0.003), con un máximo aislado de 0.6096.

Desviación estándar (pr_std y p_std):

- pr_std: Alta dispersión en algunos momentos (máximo 528.85), reflejando eventos extremos.
- p_std: Generalmente baja (0.0057), pero con máximos de 0.5246 en intervalos específicos.

Suma (pr_sum y p_sum):

- pr_sum: Media de 12,011; máximo de 12,857, sugiriendo acumulaciones altas en períodos con lluvia.
- p_sum: Media baja (0.046), indicando lluvias muy esporádicas.

Mínimos y máximos (p_min, p_max):

- Valores predominantemente cercanos a 0, salvo picos como 1.778 (lluvias intensas).

NIVEL DE AGUA

```

Estadísticas de nivel:
      nivel_metros_promedio  nivel_metros_std  nivel_metros_sum
count        432480.000000    432480.000000   432480.000000
mean         0.557829       0.006960       7.707596
std          0.170199       0.015917       3.180823
min          0.351660       0.000000       0.000000
25%          0.464924       0.001716       6.529775
50%          0.517940       0.004410       7.565600
75%          0.586128       0.007679       8.597500
max          3.785730       0.698045      48.245999

      nivel_metros_min  nivel_metros_max
count        432480.000000    432480.000000
mean         0.546971       0.569595
std          0.162832       0.178928
min          0.341200       0.357400
25%          0.455900       0.474400
50%          0.510800       0.525300
75%          0.577300       0.596100
max          3.423000       3.909200

```

Promedio (nivel_metros_promedio):

- Media: 0.558 metros, indicando niveles de agua generalmente bajos.
- Rango: 0.352 m (mínimo) a 3.786 m (máximo), con un 75% de los valores por debajo de 0.586 m.
- Distribución: Valores moderados, salvo algunos picos altos.

Desviación estándar (nivel_metros_std):

- Media: 0.007, sugiriendo estabilidad en la mayoría de las mediciones.
- Máximo: 0.698, lo que podría reflejar cambios drásticos en períodos aislados.

Suma (nivel_metros_sum):

- Promedio acumulado: 7.71 m por intervalo, con un máximo de 48.25 m.

Valores mínimos y máximos:

- Promedio del mínimo: 0.547 m; promedio del máximo: 0.570 m, mostrando poca variación en intervalos.

Adicionalmente, se realizaron histogramas y gráficas de dispersión de las variables más importantes para contar con una representación visual de cómo están distribuidos los datos.

Histogramas:

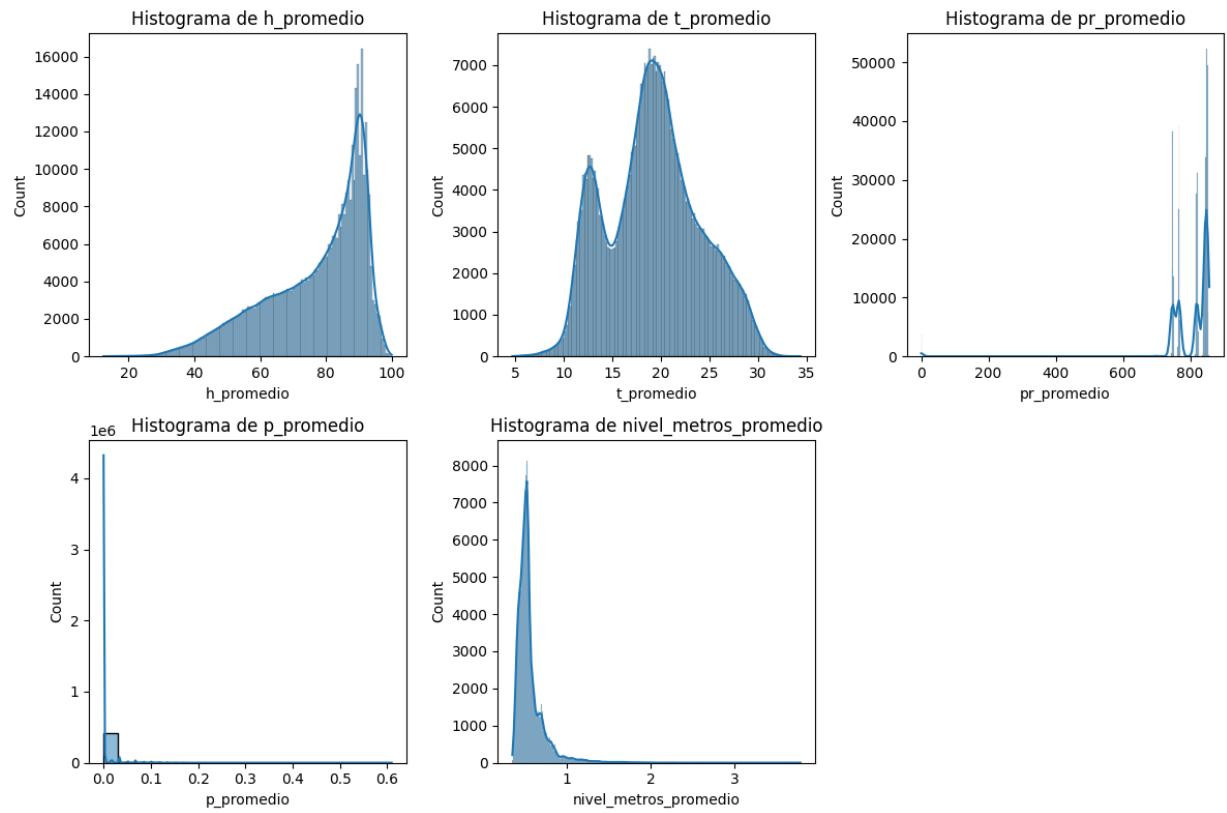
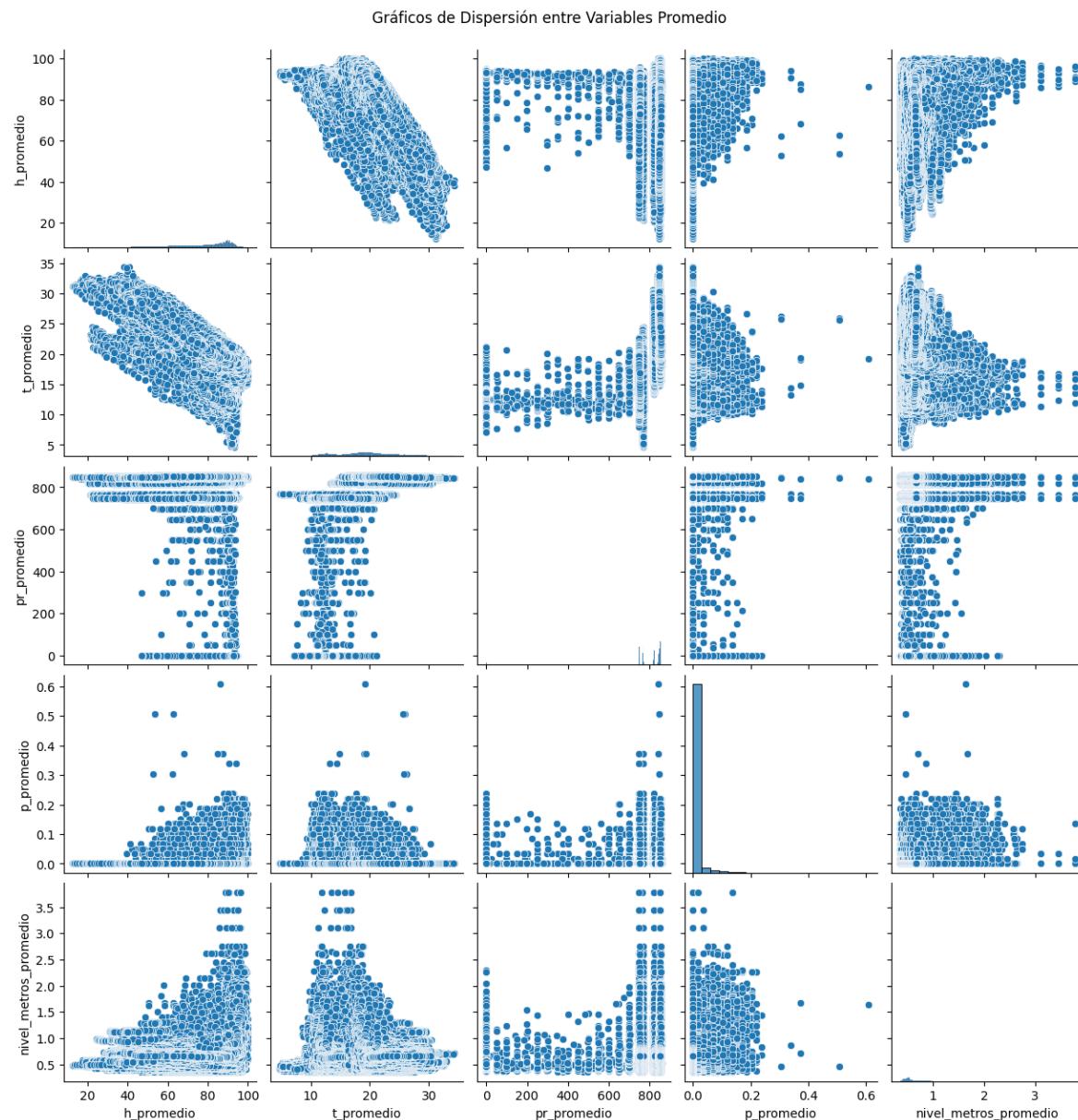


Gráfico de dispersión:



Análisis de correlación

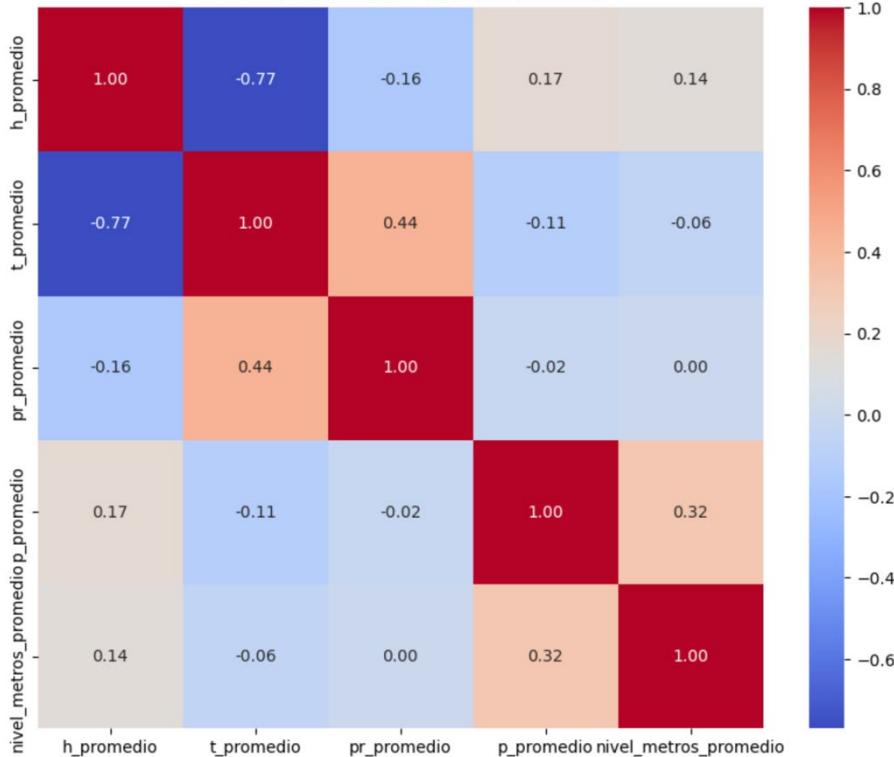
Se realizó un estudio de la correlación que hay entre las distintas variables. Para tal fin se utilizaron los registros “promedio” de las variables más importantes. Estos registros hacen una media del valor que adoptó cada variable en un lapso de quince minutos.

Matriz de correlación entre las variables promedio:

	h_promedio	t_promedio	pr_promedio	p_promedio
h_promedio	1.000000	-0.769588	-0.156962	0.168244
t_promedio	-0.769588	1.000000	0.437047	-0.114711
pr_promedio	-0.156962	0.437047	1.000000	-0.021117
p_promedio	0.168244	-0.114711	-0.021117	1.000000
nivel_metros_promedio	0.142116	-0.055073	0.001559	0.320585

	nivel_metros_promedio
h_promedio	0.142116
t_promedio	-0.055073
pr_promedio	0.001559
p_promedio	0.320585
nivel_metros_promedio	1.000000

Matriz de Correlación - Variables Promedio



Después de analizar los resultados, podemos concluir que la relación más destacada es la fuerte correlación negativa (-0.77) entre la humedad promedio (h_promedio) y la temperatura promedio (t_promedio). Es decir que, a medida que aumenta la temperatura, la humedad promedio tiende a disminuir significativamente.

Adicionalmente hay correlaciones moderadas entre:

·Temperatura promedio (t_promedio) y precipitación promedio (pr_promedio): lo que indica que en condiciones de temperaturas más altas hay mayor probabilidad de que la precipitación aumente moderadamente.

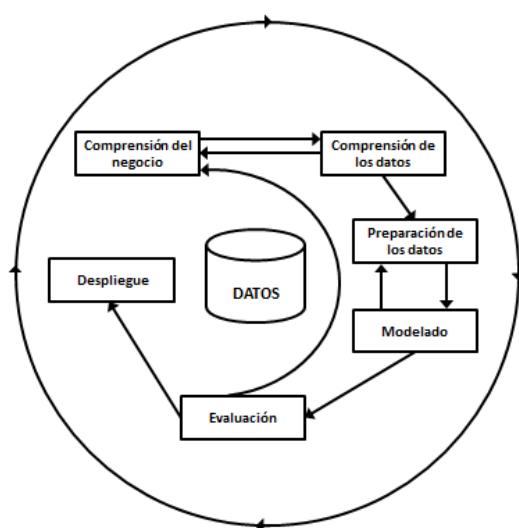
·Presión promedio (p_promedio) y altitud promedio (nivel_metros_promedio): a medida que aumenta el nivel de altitud, también lo hace ligeramente la presión promedio.

Las demás correlaciones son débiles o insignificantes, lo que implica que las variables no tienen relaciones lineales importantes entre ellas. También, encontramos que hay variables como pr_promedio que parecen estar débilmente correlacionadas con la mayoría de las otras variables, lo que sugiere un comportamiento más independiente.

3.2. MODELOS METODOLOGÍA:

Para la realización de los modelos seguimos los pasos de la metodología CRISP. De este modo, nos aseguramos de entender las necesidades del negocio de SIATA y de crear un producto acorde a ellas y a la información que nos brindan los datos con los que trabajamos.

Realizamos un modelo predictivo que nos permite predecir si habrán lluvias en la parte central de Medellín—que es aquella más vulnerable a inundaciones—with base en si se registraron lluvias en la parte oriental de la ciudad en momentos anteriores. Esto permite crear una red de alertas pluviométricas en la ciudad que facilitarán el establecimiento de sistemas de alarmas para los ciudadanos en casos de lluvias.



Comprensión de los datos

El dataset con el que trabajamos cuenta con bastantes variables de distintas tipologías. Sin embargo, para este modelo nos concentraremos en aquellas de naturaleza pluviométrica.

La información que nos brindan estos datos es un reportaje muy preciso y exhaustivo de cuáles son las precipitaciones detectadas en las estaciones esparcidas por la ciudad cada quince minutos. Recordemos que la precipitación es una medición numérica que, basándonos en los criterios del SIATA, si supera el umbral de 0.7 (0.07 en este caso dada la preparación previa del dataset) quiere decir que efectivamente está lloviendo en la zona. Con base en esto podemos dividir los registros de un modo muy simple: está lloviendo y *no* está lloviendo. De este modo, el modelo se activará cuando detecte lluvias en la zona oriental y procederá a realizar la predicción en las demás zonas.

La variable que utilizaremos para la realización de este modelo es **p_sum**, que realiza una sumatoria de las precipitaciones en los últimos quince minutos. Con estos datos acumulados por intervalos de tiempo podemos saber si las precipitaciones en una dada zona alcanzan el umbral para ser consideradas lluvia.

Comprensión del negocio

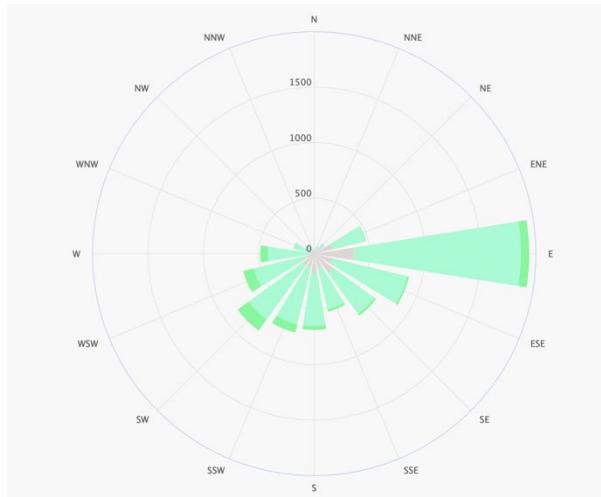
Para el SIATA, es muy importante gestionar eficientemente un sistema de alerta que le sea de ayuda a los ciudadanos y los notifique al detectarse riesgos inminentes. La ciudad de Medellín ha sufrido varias inundaciones recientemente, como aquella que se dio en el barrio Tricentenario hace pocas semanas. Lo que buscamos con este modelo es establecer un mecanismo de predicción que sea capaz de alertar la parte central de la ciudad de posibles lluvias si se detectaron lluvias en la parte oriental.

Nos concentramos en alertar la parte central de la ciudad por los siguientes motivos:

-Es la región de la ciudad con mayor población.

-Es la región de la ciudad más vulnerable a inundaciones: acorde a Control Urbanístico, en la parte céntrica de Medellín se encuentran los barrios más vulnerables frente a inundaciones (Robledo, Tricentenario, Olaya Herrera, Moravia, El Picacho, Ayurá).

La predicción se hace a partiendo desde la región oriental puesto que, según los datos proporcionados por la rosa de los vientos de la ciudad, los vientos de oriente a occidente son aquellos más frecuentes en el área metropolitana. Esto quiere decir que lo más normal es que las lluvias sean arrastradas por el viento desde el oriente hasta el centro y luego hasta el occidente.



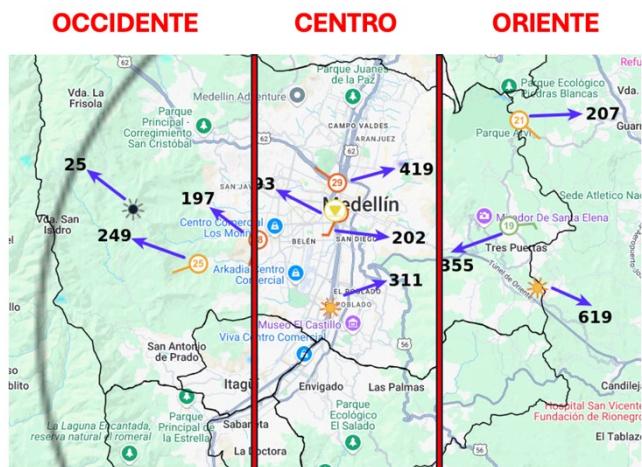
La intención a largo plazo de este modelo sería aquella de ser implementado en muchas otras direcciones e incluyendo otras regiones de Antioquia. De este modo no se limitaría solo a las lluvias que pasan sobre Medellín desde el oriente, sino que tendría la capacidad de crear una red interconectada de alertas de precipitaciones entre los distintos municipios del departamento.

Preparación de los datos

En este paso es muy importante recordar que el dataset con el que se empezó a trabajar en este modelo contaba ya con una preparación previa (aquella que se explica mucho más a detalle al inicio de este documento). Algunos de los puntos que vale la pena recordar son:

- Las variables de distintas tipologías (pluviometría, temperatura, nivel), que anteriormente se encontraban separadas en tres distintos sets de archivos, fueron juntadas en uno solo.
- Los datos antes eran registrados minuto a minuto, lo que hacía que nuestro dataset inicial tuviera millones de registros; pero realizamos una promedio de ellos en intervalos de 15 minutos para que fuera más eficiente trabajar con ellos.

Además de esto, se hizo también una preparación especialmente para este modelo de predicción. Esta consistió en dividir la ciudad en tres regiones: oriente, centro y occidente. A partir de esta división, seleccionamos dos estación meteorológicas por zona, para luego crear tres diccionarios en los que se encontraban los datos de cada una. De este modo podremos trabajar en el modelo utilizando los tres grupos de interés con mucha más facilidad.



Valores únicos en el grupo oriente: [207 355]

Valores únicos en el grupo occidente: [197 249]

Valores únicos en el grupo centro: [202 419]

Adicionalmente, se creó y agregó una columna binaria al dataset llamada “Lluvia”. Esta columna tiene en consideración el umbral establecido por SIATA para saber si está efectivamente lloviendo (precipitación > 0.7), y le asigna a todos los datos la característica de si está o no está lloviendo (“Sí” y “NO”). Esta columna nos es de mucha utilidad porque permite filtrar muy fácilmente los datos al momento de estructurar el modelo, de este modo el modelo activa su mecanismo al detectar una casilla con lluvias.

p_max	codigo_mapeado_2	codigo	nivel_metros_promedio	nivel_metros_std	nivel_metros_sum	nivel_metros_min	nivel_metros_max	lluvia
0.0	93.0	93.0	0.597520	0.006205	8.962799	0.5892	0.6065	No
0.0	93.0	93.0	0.596887	0.007073	8.953300	0.5848	0.6141	No
0.0	93.0	93.0	0.590020	0.006234	8.850299	0.5794	0.5984	No
0.0	93.0	93.0	0.584540	0.005627	8.768099	0.5735	0.5935	No
0.0	93.0	93.0	0.581833	0.004773	8.727501	0.5740	0.5913	No

Modelado

En esta etapa, buscamos construir modelos de clasificación supervisados, modelos predictivos y modelos no supervisados utilizando los datos previamente preparados. Los pasos realizados para configurar los modelos son:

Definición de los objetivos de los modelos:

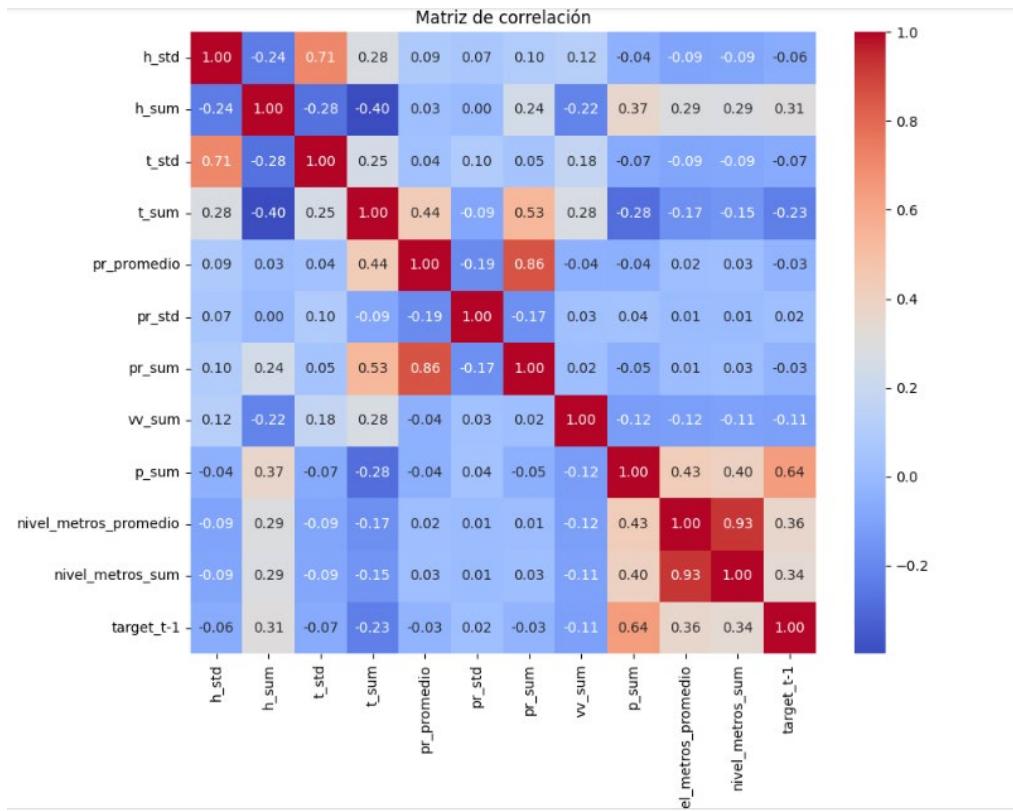
Los objetivos principales de los modelos son: predecir el nivel promedio de agua en las estaciones meteorológicas de Medellín en función de las variables relacionadas con la pluviometría, un modelo de clasificación de supervisado que nos permita saber si llovió o no y finalmente, un modelo de clasificación no supervisado. Para esto, se utilizaron las variables más relevantes del conjunto de datos previamente analizado.

Cada uno de los modelos de clasificación se ejecutó dos veces: una sin balancear los datos y otra después de aplicar el balanceo con SMOTE, utilizando una proporción de 50/50. Esta decisión se tomó para garantizar que no se pierda ninguna predicción de lluvia, ya que el objetivo de nuestro modelo es la detección de alertas tempranas, prestando igual atención a ambas clases.

1. Selección de variables predictoras

Con base en el análisis exploratorio y de correlación, se seleccionaron las variables más representativas que explican cada una de las variables a modelar:

Modelo de clasificación supervisada: Predicción de Lluvia



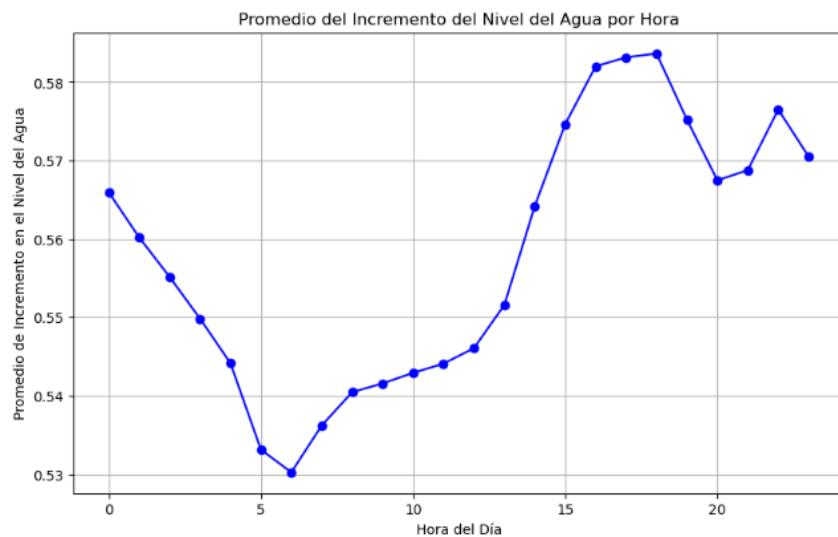
- **p_sum (0.64):** La presión acumulada es una variable clave para determinar el estado de lluvia del período anterior, ya que refleja condiciones atmosféricas directamente relacionadas con las precipitaciones.
- **h_sum (0.31):** La suma de la humedad presenta una correlación moderada con target_t-1, es decir, que un mayor nivel de humedad acumulada podría estar vinculado con la continuidad de condiciones de lluvia.
- **t_sum (-0.23):** Las temperaturas acumuladas más altas pueden reducir la probabilidad de que el estado de lluvia previo influya en las condiciones actuales.
- **pr_sum (0.11):** La precipitación acumulada tiene una relación positiva débil con target_t-1. Esto indica que, aunque no es determinante, un aumento en la precipitación podría estar ligeramente asociado con el estado de lluvia anterior.
- **vv_sum (-0.11):** La suma de la velocidad del viento muestra una relación débil negativa con target_t-1, lo que indica que niveles altos de viento acumulado no están directamente asociados con las condiciones de lluvia previas.
- **h_std (-0.06):** La variabilidad de la humedad presenta una correlación muy débil negativa, indicando que no tiene un impacto significativo en la predicción del estado de lluvia previo.

- **t_std (-0.07):** La desviación estándar de la temperatura tiene una relación negativa débil, lo que sugiere que la variabilidad de la temperatura tampoco es relevante para determinar el estado de lluvia previo.
- **pr_promedio (0.02):** El promedio de precipitación muestra una correlación casi nula, indicando que no contribuye significativamente a la predicción basada en target_t-1.
- **pr_std (0.00):** La variabilidad de la precipitación no tiene correlación con target_t-1, lo que indica que su impacto en el estado previo de lluvia es inexistente.

Observamos que las variables seleccionadas deberían proporcionar suficiente información para modelar el estado de lluvia (target_t-1) con mayor precisión, ya que presentan las correlaciones más significativas. En particular, la presión acumulada (p_sum) y la humedad total (h_sum) destacan como factores clave.

Modelo de Predicción: Nivel del agua del río:

Para determinar las variables a incluir en el modelo, realizamos un análisis exploratorio utilizando gráficos como los que se muestran a continuación.



El primer gráfico muestra cómo, a medida que la precipitación acumulada (p_sum) aumenta, también se incrementa el nivel del agua, especialmente en las horas de mayor intensidad de lluvia.

Adicionalmente, incorporamos variables de tiempo como la hora del día, el día de la semana y representaciones trigonométricas de la periodicidad diaria (hora_sin, hora_cos). Estas variables permitieron capturar mejor las dinámicas temporales y

estacionales, mejorando la capacidad del modelo para identificar patrones complejos y realizar predicciones más precisas del nivel del río.

2. División de los datos

Los datos fueron divididos en dos subconjuntos:

- Entrenamiento (80%): Para ajustar el modelo.
- Prueba (20%): Para evaluar el desempeño del modelo en datos no vistos.

3. Elección de modelos

En el proceso de selección de modelos, consideramos distintas alternativas para abordar el problema de predicción y clasificación, eligiendo aquellos algoritmos que mejor se ajustan a las características del conjunto de datos y a los objetivos planteados. A continuación, se describen las opciones evaluadas y las razones para su consideración:

Modelo de clasificación supervisada: Predicción de Lluvia

- Random Forest Classifier: Este modelo se basa en múltiples árboles de decisión entrenados en subconjuntos aleatorios del conjunto de datos, lo que le da una alta capacidad de generalización.
- Regresión Logística: Considerada para predecir la presencia o ausencia de lluvia (Si o No) debido a su simplicidad y capacidad para problemas de clasificación binaria. Su interpretabilidad es una ventaja, aunque es limitada para capturar relaciones no lineales.
- CatBoosts: Este modelo, basado en boosting de gradiente, es especialmente eficaz para manejar datos categóricos de manera eficiente sin necesidad de transformaciones extensivas. Se evaluó debido a su capacidad para capturar relaciones complejas entre las variables y su desempeño competitivo tanto en tareas de clasificación como de regresión, incluso con conjuntos de datos con estructuras mixtas o desbalanceadas.

Modelo de Predicción: Nivel del agua del río

- Random Forest Regression: Utilizamos este modelo para predecir el nivel del río, debido a que se destaca por su precisión en tareas de regresión y su capacidad para manejar relaciones complejas entre las variables como en este caso.

4. Entrenamiento del modelo:

El modelo fue entrenado utilizando el conjunto de datos de entrenamiento, optimizando los parámetros iniciales y realizando un balanceo a las métricas considerando la naturaleza desbalanceada de los datos.

5. Evaluación

Para la evaluación de los modelos, revisamos las métricas: precisión, recall, F1-Score, accuracy, y curva ROC en la variable de lluvia. Comparamos estas métricas y de acuerdo con esto seleccionamos el mejor modelo. Adicionalmente, se realizó una validación cruzada, haciendo una separación adecuada entre el conjunto de entrenamiento, de validación y test.

3.3. MODELO DE CLASIFICACIÓN SUPERVISADA: PREDICCIÓN DE PRECIPITACIONES EN MEDELLÍN

Implementamos un modelo para predecir la probabilidad de lluvia utilizando datos meteorológicos históricos. Nuestro objetivo fue clasificar los días entre "lluvia" y "no lluvia", empleando los algoritmos Random Forest, Regresión Logística y CatBoost.

Para el análisis, dividimos los datos en tres conjuntos: entrenamiento (70%), validación (30% del entrenamiento) y prueba (20% del total). Esto nos permitió ajustar los modelos, evaluar su desempeño en un entorno controlado y comprobar su capacidad de generalización con datos nuevos. Las métricas utilizadas incluyeron precisión, recall, F1-score y el área bajo la curva (AUC).

Adicionalmente, para mejorar el análisis y capturar patrones temporales, creamos una nueva variable llamada target_t-1. Esta variable representa el estado de lluvia del período anterior para cada ubicación geográfica (codigo_x), lo que permitió al modelo incorporar información histórica inmediata en sus predicciones. Agrupamos los datos por codigo_x y utilizamos la función shift(-1) para generar este rezago temporal. Esto nos permitió relacionar el comportamiento de la lluvia en un momento específico.

Esta estrategia nos permitió integrar un componente temporal al análisis, capturando las dependencias entre períodos consecutivos. Además, eliminamos las filas con valores nulos generados por el rezago para asegurar la integridad de los datos. Esto fue especialmente útil para entrenar modelos que pueden aprovechar las secuencias temporales, como Random Forest y CatBoost.

Logistic Regression - Test Set Metrics				
	precision	recall	f1-score	support
0.0	0.84	0.88	0.86	3190
1.0	0.68	0.59	0.63	1344
accuracy			0.80	4534
macro avg	0.76	0.74	0.75	4534
weighted avg	0.79	0.80	0.79	4534
Confusion Matrix:				
	[[2820 370]			
	[547 797]]			

```
Random Forest - Test Set Metrics
precision    recall   f1-score  support
```

0.0	0.87	0.84	0.85	3190
1.0	0.64	0.69	0.67	1344

accuracy			0.79	4534
macro avg	0.75	0.76	0.76	4534
weighted avg	0.80	0.79	0.80	4534

```
CatBoost - Test Set Metrics
```

precision	recall	f1-score	support
-----------	--------	----------	---------

0.0	0.87	0.84	0.86	3190
1.0	0.65	0.71	0.68	1344

accuracy			0.80	4534
macro avg	0.76	0.78	0.77	4534
weighted avg	0.81	0.80	0.80	4534

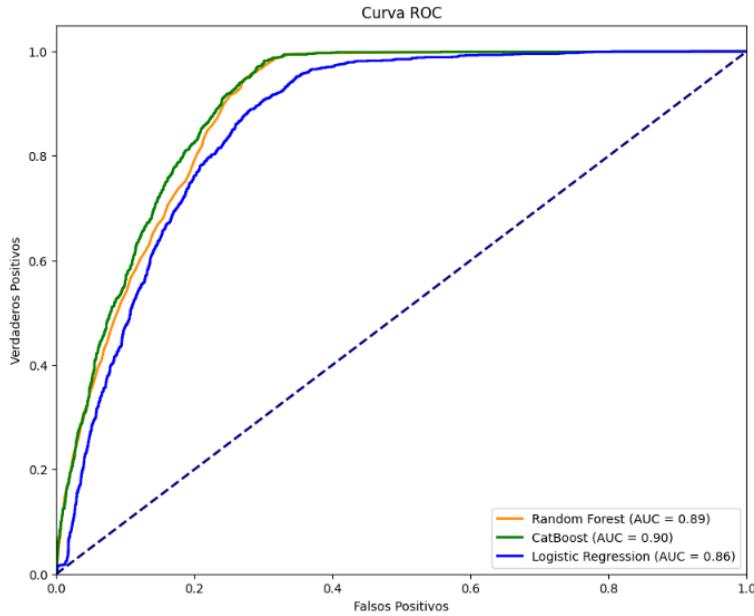
```
Confusion Matrix:
```

```
[[2681  509]
 [ 386 958]]
```

Entre los modelos probados, CatBoost es la mejor opción, logrando un AUC promedio de 0.90 en validación, como se muestra en la imagen y un desempeño consistente en el conjunto de prueba. Su capacidad para identificar días con lluvia fue notable, con un recall del 74% en validación y 71% en prueba. Además, presentó la menor cantidad de falsos negativos, lo que lo hace ideal para escenarios donde no predecir la lluvia puede ser crítico.

Random Forest mostró un desempeño equilibrado, con un AUC competitivo (0.89) y un recall del 71%, aunque con más falsos negativos que CatBoost. Por otro lado, la Regresión Logística, aunque adecuada, tuvo limitaciones en la identificación de lluvia, con un recall menor y más errores de predicción.

En conclusión, recomendamos el uso de CatBoost como el modelo principal, complementado con una posible optimización de sus parámetros para maximizar su efectividad en la práctica.



3.4. MODELO DE PREDICCIÓN: NIVEL DEL RÍO

Construimos un modelo para predecir los niveles futuros del río basándonos en patrones históricos y datos climáticos clave. El objetivo era anticipar los cambios en el nivel del agua y reducir riesgos asociados a fluctuaciones inesperadas. Para ello, procesamos los datos aplicando técnicas como rezagos temporales, acumulados y transformaciones trigonométricas para capturar patrones diarios y la influencia directa de las precipitaciones.

Entrenamos el modelo dividiendo los datos en un 80% para entrenamiento y un 20% para prueba, asegurando así una evaluación confiable con datos no vistos. Las métricas de evaluación demostraron un desempeño sólido: el error absoluto medio (MAE) fue de 0.03, lo que significa que, en promedio, las predicciones del nivel del río se desviaron solo 0.03 unidades de los valores reales. El error cuadrático medio (RMSE), de 0.09, reflejó una dispersión mínima en los errores, mientras que el coeficiente de determinación (R^2) alcanzó un 71%, mostrando que el modelo logra explicar la mayor parte de la variación en los niveles futuros del río.

.

3.4. MODELOS DE CLASIFICACIÓN NO SUPERVISADOS

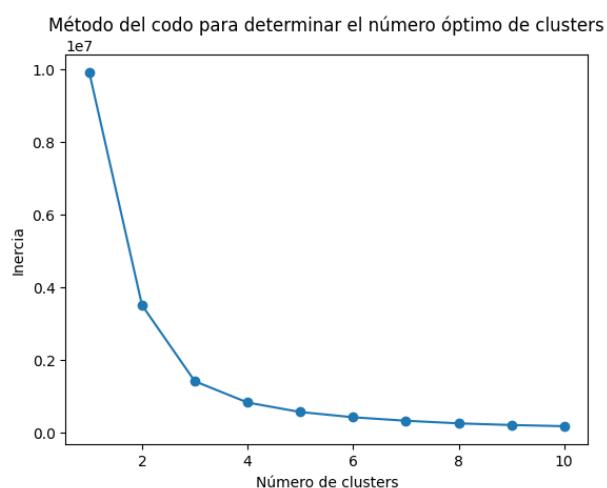
Para esta sección de nuestro proyecto trabajaremos con las variables de **temperatura** de nuestro dataset. El objetivo que tenemos es realizar una clasificación de las temperaturas registradas en las distintas zonas.

K-MEANS

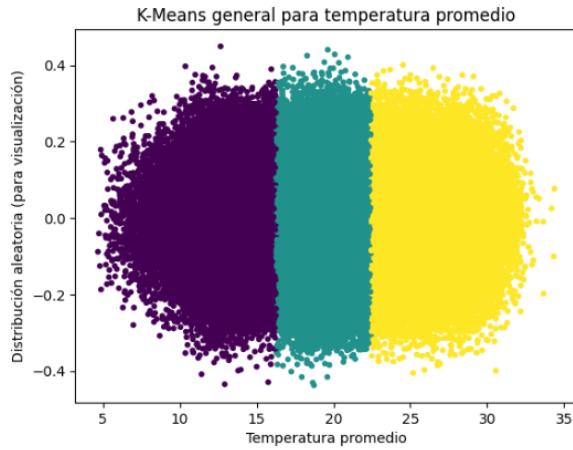
Como el modelo se concentra en una sola variable, la temperatura promedio, no es necesario realizar una reducción de dimensionalidad previa. Esto se debe a que el algoritmo de K-Means operará directamente sobre esta variable, agrupando los valores en clusters según su proximidad.

El clustering realizado es **no supervisado**, ya que no depende de etiquetas predefinidas o segmentación jerárquica. El algoritmo de K-Means identifica clusters basados únicamente en la proximidad de los valores de temperatura. Aunque las estaciones meteorológicas proporcionan el contexto geográfico de los datos, K-Means agrupa las temperaturas por similitud, sin tomar en cuenta explícitamente la ubicación de las estaciones.

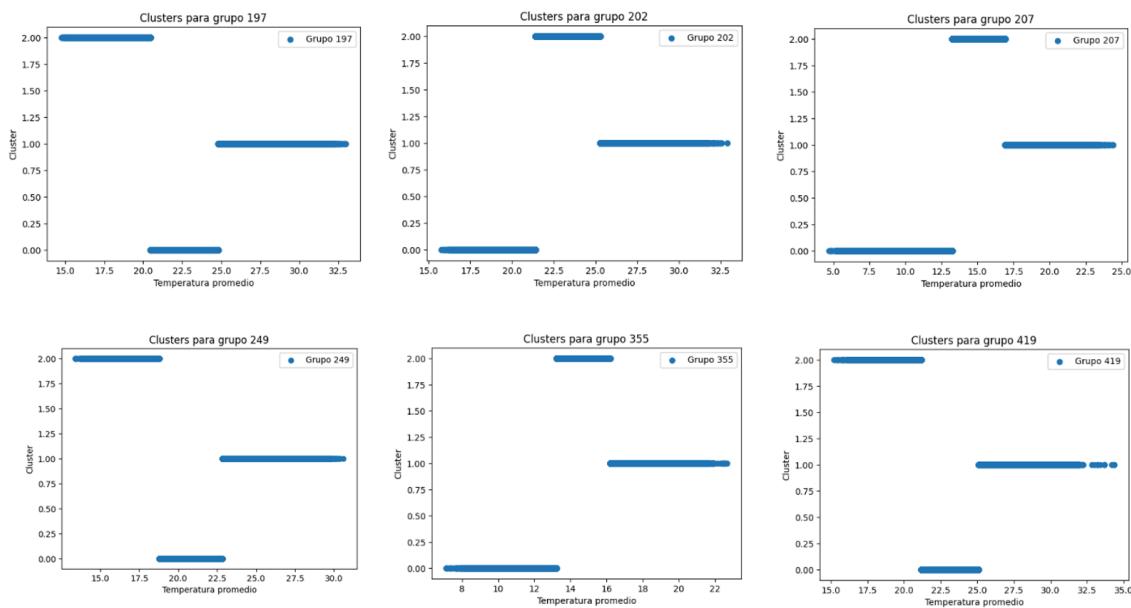
Para determinar la cantidad ideal de clusters, utilizamos el **método del codo**, el cual nos muestra el punto en que la disminución de la inercia (error cuadrático) al agregar más clusters se vuelve marginal. En nuestro caso, un número adecuado de clusters es 3, ya que después de este punto la mejora es mínima. Una vez definidos los clusters, asignamos etiquetas como 'baja', 'moderada' y 'alta' temperatura, lo cual es comprensible para los usuarios y facilita la interpretación de los resultados.



K-MEANS general de la variable t_promedio:



K-MEANS por estaciones meteorológicas:



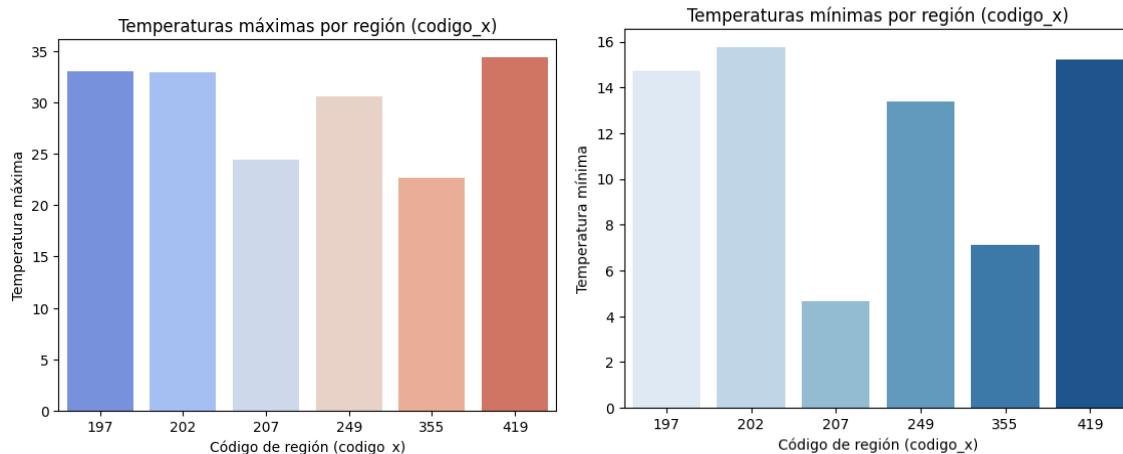
A simple vista, podemos notar un comportamiento bastante parecido entre todos los grupos. Esto tiene mucho sentido dado que no hay una distancia geográfica lo suficientemente grande entre ellos como para ver diferencias considerables de temperaturas. Otras conclusiones visibles en estos clusters son:

-Algunas estaciones (355 y 207) registran temperaturas por lo general más bajas que el resto de las estaciones.

-En todos los grupos, los clusters que reúnen las temperaturas más altas son aquellos con más registros. Por lo que podríamos decir que entre temperaturas bajas, moderadas y altas; las altas son las más frecuentes.

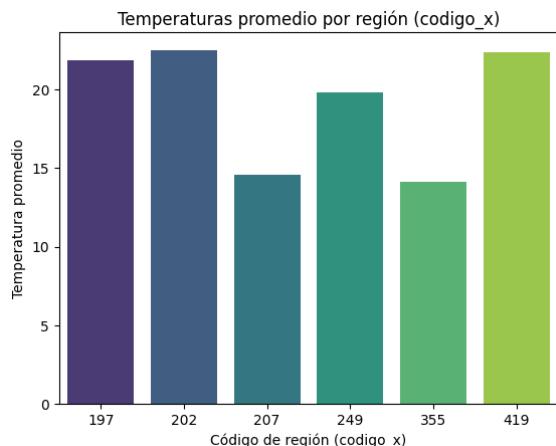
Temperaturas mínimas y máximas registradas por zona:

Estos gráficos muestran cuál fue la temperatura más alta y la temperatura más baja registrada en cada una de las estaciones. No obstante, esto no es un indicador muy recomendable para sacar conclusiones puesto que pueden ser casos excepcionales que no reflejan correctamente las temperaturas más frecuentes en las respectivas zonas.



Temperatura promedio por zona:

La temperatura promedio es un indicador mucho más adecuado pues tiene en cuenta todas las temperaturas registradas y no solo los casos extremos. Al ver este gráfico podemos sacar las siguientes conclusiones:



- Las estaciones 207 (cercana al Parque Ecológico Piedras Blancas) y 355 (cercana a Santa Elena) son aquellas con las temperaturas más bajas de toda la ciudad. Registrando en promedio temperaturas que rondan en los 15 °C. Esto quiere decir que la región **oriental** de Medellín es aquella más fría.
- El resto de la ciudad (**centro** y **occidente**) presentan temperaturas muy similares y más templadas que el oriente. Las temperaturas rondan entre los 20 y los 24 °C.

más templadas que el oriente. Las temperaturas rondan entre los 20 y los 24 °C.

Para evaluar el rendimiento de nuestro modelo K-MEANS, calculamos su **silhouette score**:

Silhouette Score MiniBatch K-Means: 0.5896789192851108

El silhouette score obtenido es de aproximadamente 0.6. Esto quiere decir que los clusters están razonablemente bien definidos, pero es posible que haya solapamiento, es decir que algunos elementos se superpongan entre sí. En nuestro caso este solapamiento es aceptable dada la naturaleza de los datos con los que trabajamos.

Se prefirió afrontar este problema con K-MEANS sobre DBSCAN por los siguientes motivos:

· El K-MEANS es un método muy eficiente computacionalmente, incluso para datasets grandes como con el que estamos trabajando. Además, es de muy fácil interpretación, y esto es un punto muy importante porque nuestro objetivo es darle información clara a la ciudadanía.

· El DBSCAN puede no funcionar muy bien cuando hay presencia de solapamiento entre los clusters, como lo es el caso dada la naturaleza de los datos.

· Los datos que estamos utilizando no presentan outliers muy considerables y, cosa que hace que la complejidad adicional del DBSCAN no sea necesaria.

4. CICLO DE VIDA DE LOS DATOS Y PROCESAMIENTO ANALÍTICO

4.1. DESCRIPCIÓN DEL SISTEMA Y ENFOQUE PROPUESTO

El sistema del SIATA recolecta información en tiempo real desde diferentes sectores. Esta información es preprocesada para evaluar su calidad antes de ser enviada a la base de datos del SIATA. Sin embargo, los datos disponibles en las páginas web se obtienen a través de un flujo streaming, lo que implica que no cuentan con una capa de validación exhaustiva para garantizar una alta calidad de los datos, no obstante, en SIATA hace una evaluación exhaustiva para determinar la calidad de los datos después de ser publicados.

Por este motivo, se decide aplicar diversas metodologías para el procesamiento de la información, priorizando un enfoque batch que recopile y procese los datos en lotes antes de almacenarlos y analizarlos. Esto permite alcanzar un nivel de calidad más riguroso, considerando la falta de acceso directo a la fuente original de datos con validación exhaustiva.

Paralelamente, se incorpora el enfoque streaming, que gestiona la captura y el análisis de datos en tiempo real, proporcionando insights inmediatos y permitiendo la toma de decisiones rápida cuando sea necesario. Esta combinación garantiza que tanto los datos históricos como los datos en tiempo real sean procesados de manera efectiva, maximizando la utilidad y la calidad de la información.

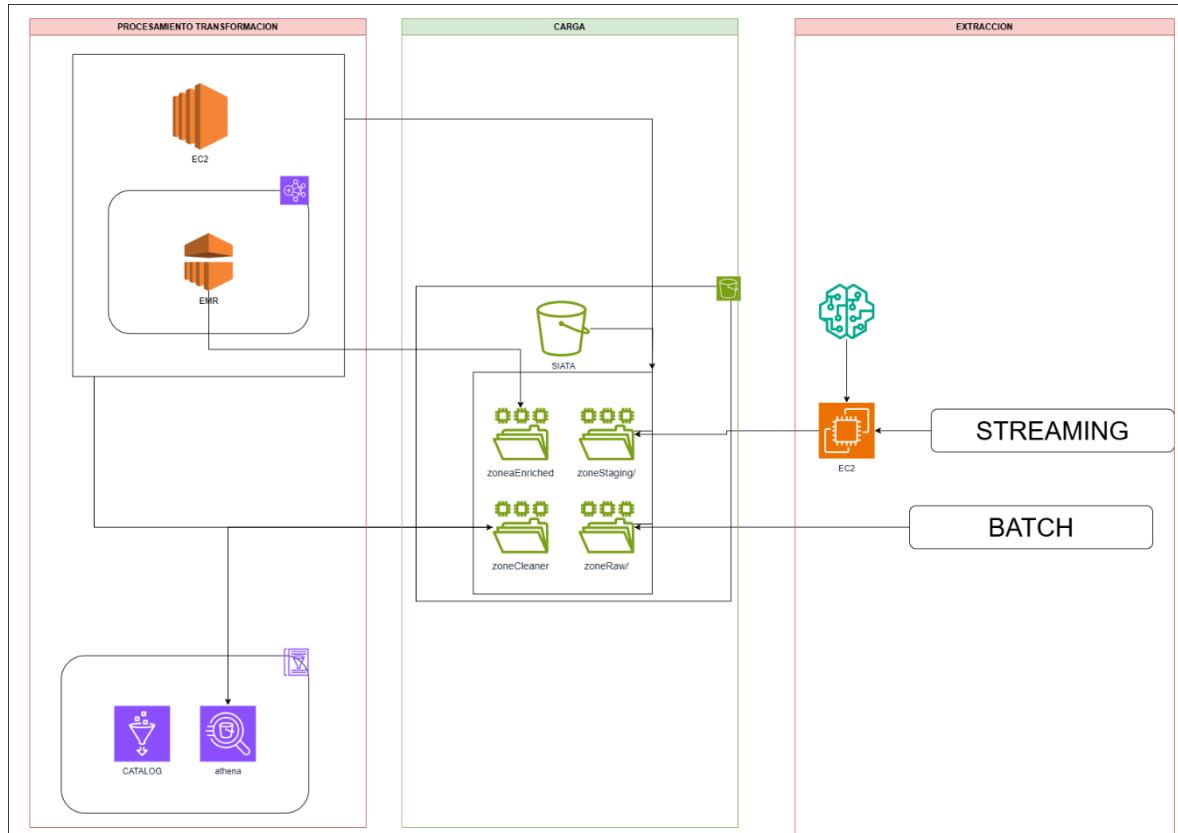
4.2. ARQUITECTURA

La arquitectura para usar es la lambda donde construye sobre un data lakehouse, combinando las capacidades de los almacenes de datos tradicionales y los data lakes. Esto proporciona un acceso rápido a los datos y compatibilidad, lo que facilita a los usuarios de negocio visualizar métricas de manera eficiente mediante herramientas de ETL.

El enfoque incluye el uso de data lakes para almacenar grandes volúmenes de datos, ya sean estructurados (como en nuestro caso) o no estructurados, en su forma nativa. Los data lakes serán fundamentales para la creación de modelos estadísticos y

analíticos, aunque presentan el desafío de un posible deterioro en la calidad de los datos almacenados.

Sin embargo, al optar por un data lakehouse, se unifica la estructura de datos, permitiendo al equipo acceder a los datos de manera eficiente y aprovecharlos para múltiples propósitos. Esta solución admite cualquier tipo de dato, permitiendo su almacenamiento inicial y posterior reestructuración según las necesidades analíticas o de negocio.



Modos de ingestión de datos

Se optó por una arquitectura de ingestión híbrida debido a la necesidad de manejar diferentes tipos de datos y niveles de procesamiento. Además, los datos recolectados no provienen directamente de los sensores, sino que son proporcionados por el SIATA con un cierto desfase temporal. Por lo tanto, se implementará la ingestión de la siguiente manera:

- **Batch:** Se utilizará el procesamiento por lotes para agrupar y procesar datos en conjuntos específicos. Esto es importante porque permitirá seleccionar conjuntos de datos provenientes de nuestra base de datos, como zonas definidas, para su análisis y

procesamiento. Este enfoque ayudará a ejecutar procesos destinados al análisis y a la creación de modelos, siguiendo un esquema D-1 este tipo de procesamiento es ideal para manejar grandes volúmenes de datos acumulados en intervalos minútiales.

- **Streaming:** Se requiere procesar los datos a medida que llegan, permitiendo la ejecución inmediata de los modelos y el almacenamiento simultáneo de la información procesada. Esto es debido a que se debe realizar análisis en tiempo real y tomar decisiones de forma inmediata. Sin embargo, este enfoque demanda una infraestructura robusta.

ETL (Extract, Transform, Load):

Extracción

La extracción de datos se realizará utilizando Python para tomar la información del SIATA cuando usen nuestro modelo y guardarla en S3. Para los datos batch, se integrarán procesos que consuman archivos con un desfase temporal y los almacenen directamente en la capa raw data lake en Amazon S3. Los datos streaming serán capturados mediante un endpoint en EC2, permitiendo la recepción en tiempo real y su almacenamiento inicial sin procesar en S3. Este enfoque asegura que tanto los datos históricos como los datos en tiempo real estén disponibles para los siguientes procesos.

Transformación

En esta etapa, los datos extraídos se someten a procesos de limpieza, validación y estandarización, utilizando librerías como Pandas y PySpark. Se aplican reglas de calidad basadas en completitud, conformidad, consistencia, precisión, duplicidad e integridad para garantizar que los datos estén alineados con los estándares de negocio. La transformación incluye la conversión de formatos csv y el particionamiento por fecha, zona o tipo de datos, optimizando su uso posterior en análisis y modelos estadísticos.

Calidad general

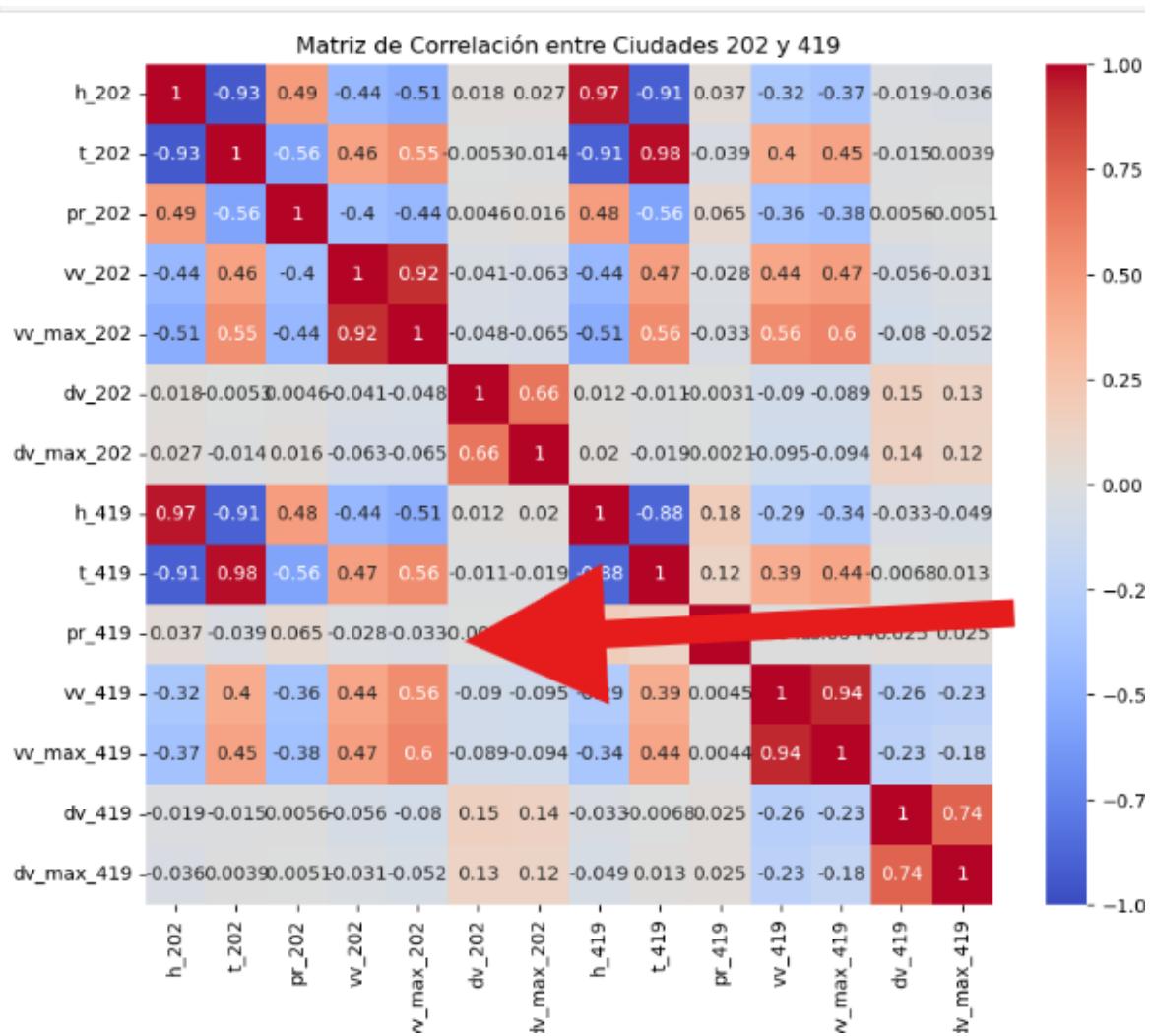
```

]: calidad_dudosa
[Buena]
14017327
[Humedad, Precipitación, Presión Atmosférica]
219949
[Temperatura, Precipitación, Presión Atmosférica]
57574
[Precipitación, Presión Atmosférica]
49962
[Precipitación, Presión Atmosférica, Viento Magnitud, Viento Dirección]
46531
[Humedad, Presión Atmosférica]
13901
[Precipitación, Presión Atmosférica, Viento Magnitud]
2918
[Presión Atmosférica]
2196
[Temperatura, Humedad, Precipitación, Presión Atmosférica]
1922
[Presión Atmosférica, Viento Magnitud, Viento Dirección]
914
[Humedad, Precipitación, Presión Atmosférica, Viento Magnitud]
174
[Presión Atmosférica, Viento Magnitud]
166
[Humedad, Precipitación, Presión Atmosférica, Viento Magnitud, Viento Dirección]
72
[Temperatura, Humedad, Presión Atmosférica]
38
[Temperatura, Precipitación, Presión Atmosférica, Viento Magnitud]
32
[Temperatura, Precipitación, Presión Atmosférica, Viento Magnitud, Viento Dirección]
5
[Humedad, Presión Atmosférica, Viento Magnitud, Viento Dirección]
4
[Precipitación, Presión Atmosférica, Viento Dirección]
3
Name: count, dtype: int64

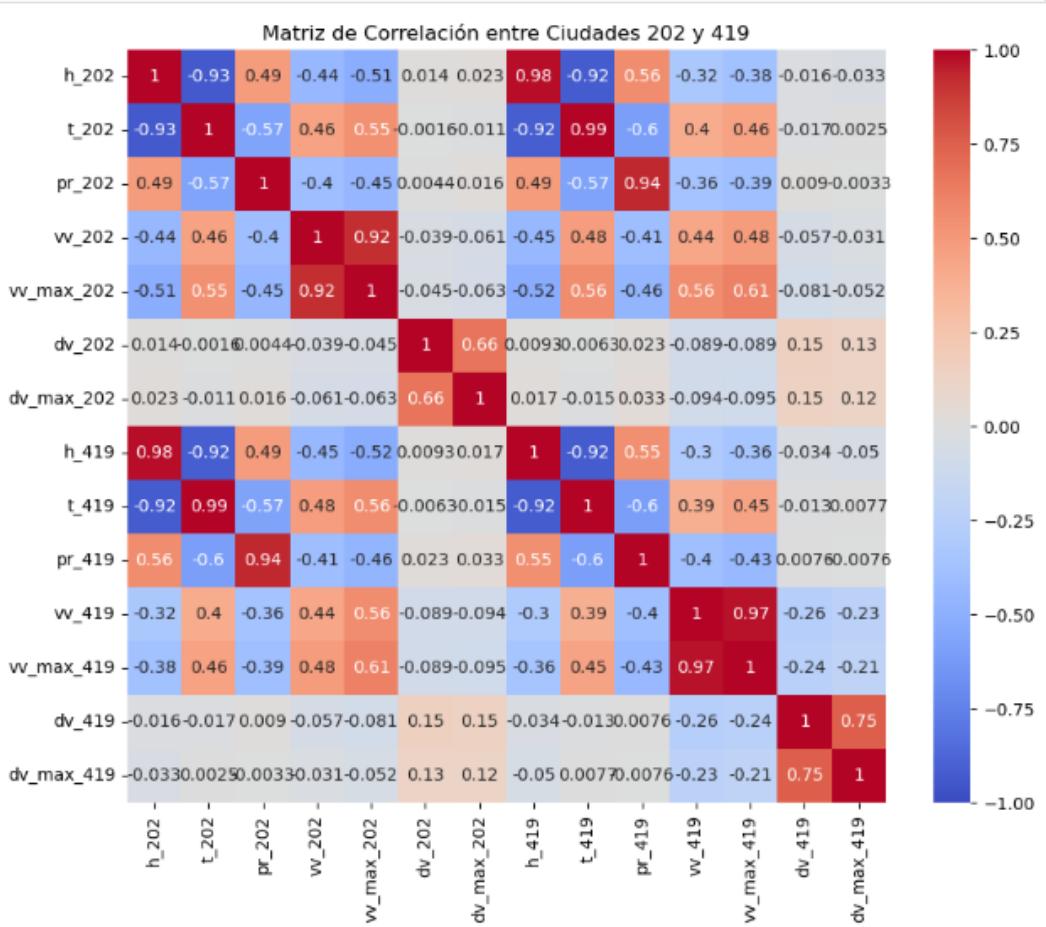
```

Compleitud

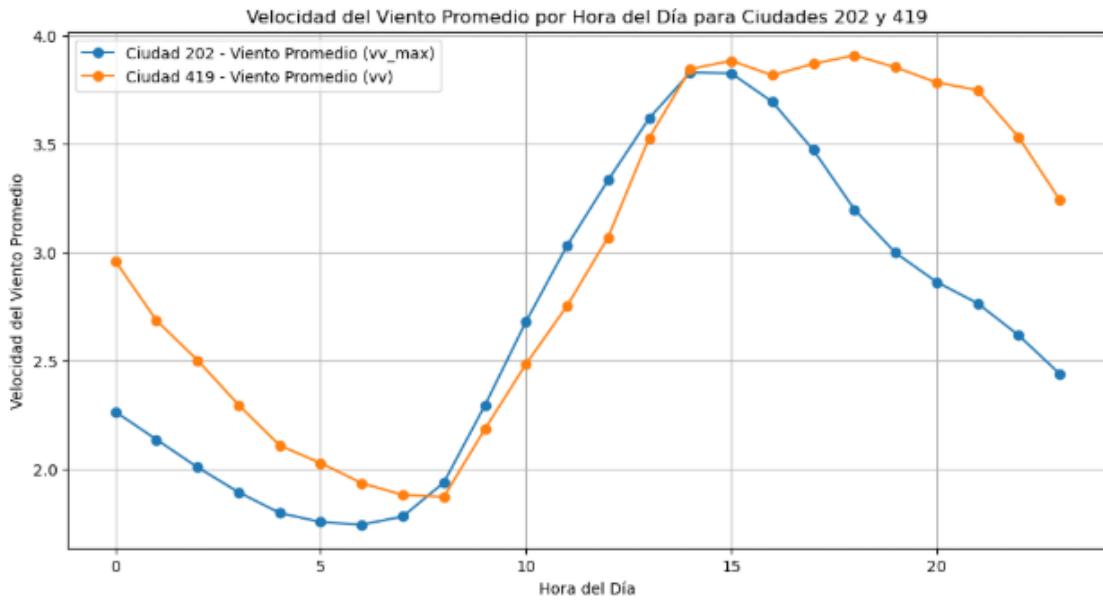
	codigo	fecha_hora	diferencia_tiempo
2225455	197	2024-06-26 11:03:00	37227.0
2914787	202	2021-04-13 14:18:00	4087.0
6444429	207	2023-12-12 14:14:00	121167.0
8501946	249	2023-07-31 12:04:00	27496.0
10621907	355	2023-06-08 16:13:00	21910.0
12981691	419	2023-07-28 15:30:00	31966.0



Se determina que se puede llenar la información de puntos por medio de una regresión lineal. Por ejemplo el punto 202 y 419 están muy correlacionadas, sin una buena selección de datos las relaciones no se obtendrían correctamente.



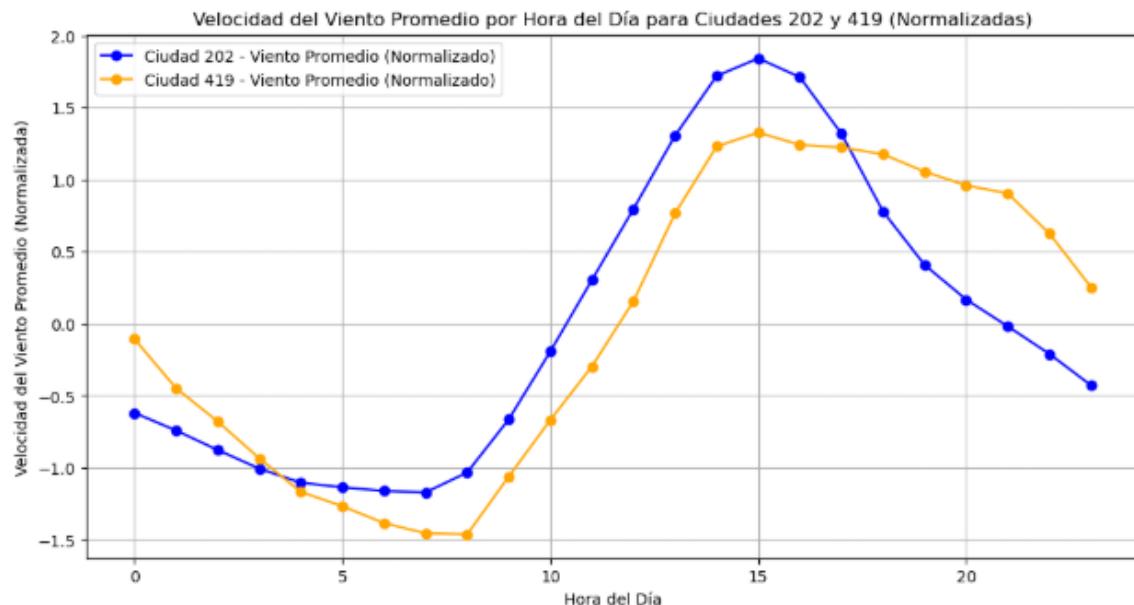
Se encontraron correlaciones Spearman de diferentes variables. En este caso se encontró que la velocidad promedio $vv_{max\ 202}$ es muy similar a vv de 419. Esto es debido a que están muy cerca.



Correlación de Spearman por hora entre vv_max (Ciudad 202) y vv (Ciudad 419): 0.853043
4782608695

Si

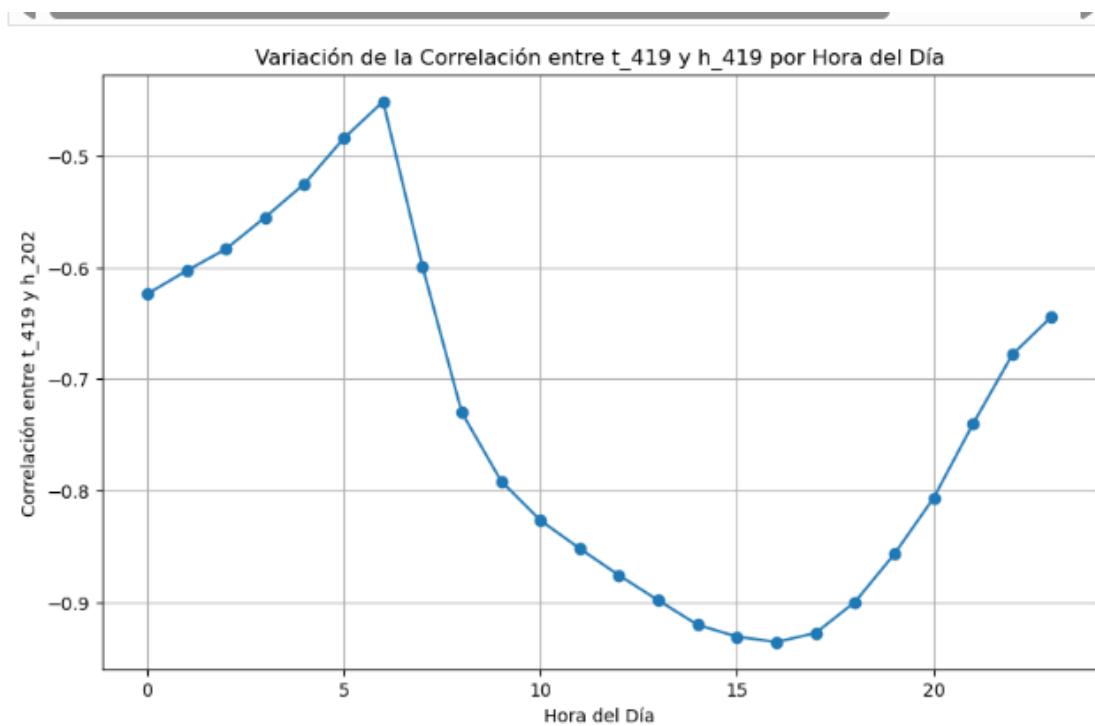
le aplicamos una normalización obtendremos un mejor puntaje de Spearman



Correlación de Spearman por hora entre vv (Ciudad 202) y vv (Ciudad 419): 0.9226086956
521738

Esto se aplica para cada uno de los puntos cercanos.

Se encontró que se puede jugar con la hora para tener mayor precisión al encontrar la temperatura:



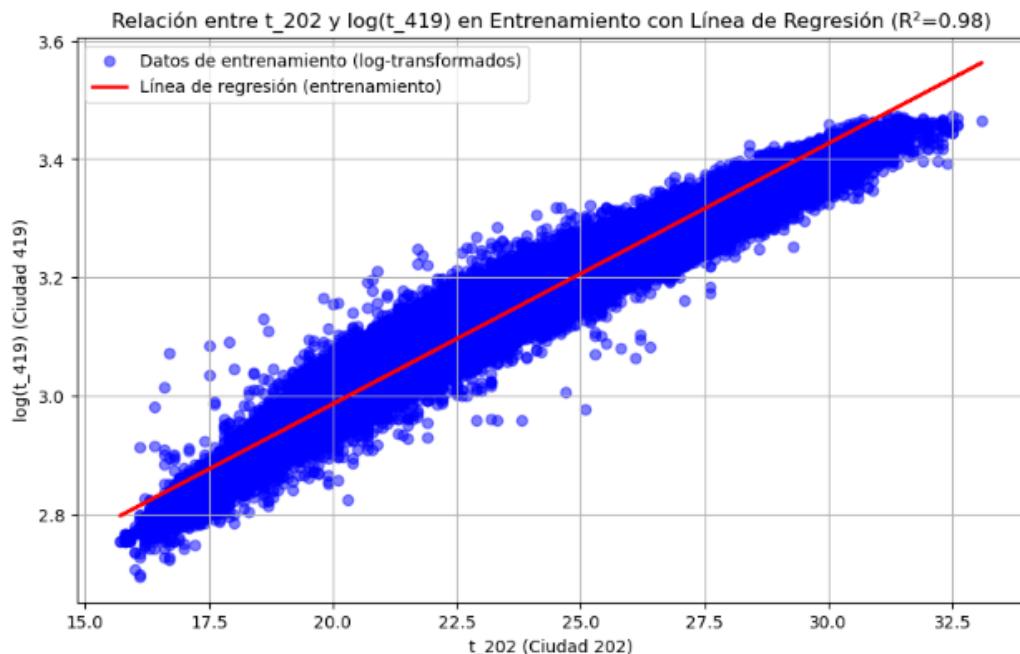
La correlación aumenta entre estas dos variables mientras se llega a la tarde.

Haciendo el entrenamiento de los modelos se obtiene:

```
Conjunto Entrenamiento
Error Absoluto Medio (MAE): 0.014987410229511852
Error Cuadrático Medio (MSE): 0.00040183978386739116
Raíz del Error Cuadrático Medio (RMSE): 0.02004594183039029
Precisión del modelo (R2): 0.9803295491859751
```

```
Conjunto Validación
Error Absoluto Medio (MAE): 0.014967961665289822
Error Cuadrático Medio (MSE): 0.0004002312296348058
Raíz del Error Cuadrático Medio (RMSE): 0.0200057799056874
Precisión del modelo (R2): 0.9804519353687532
```

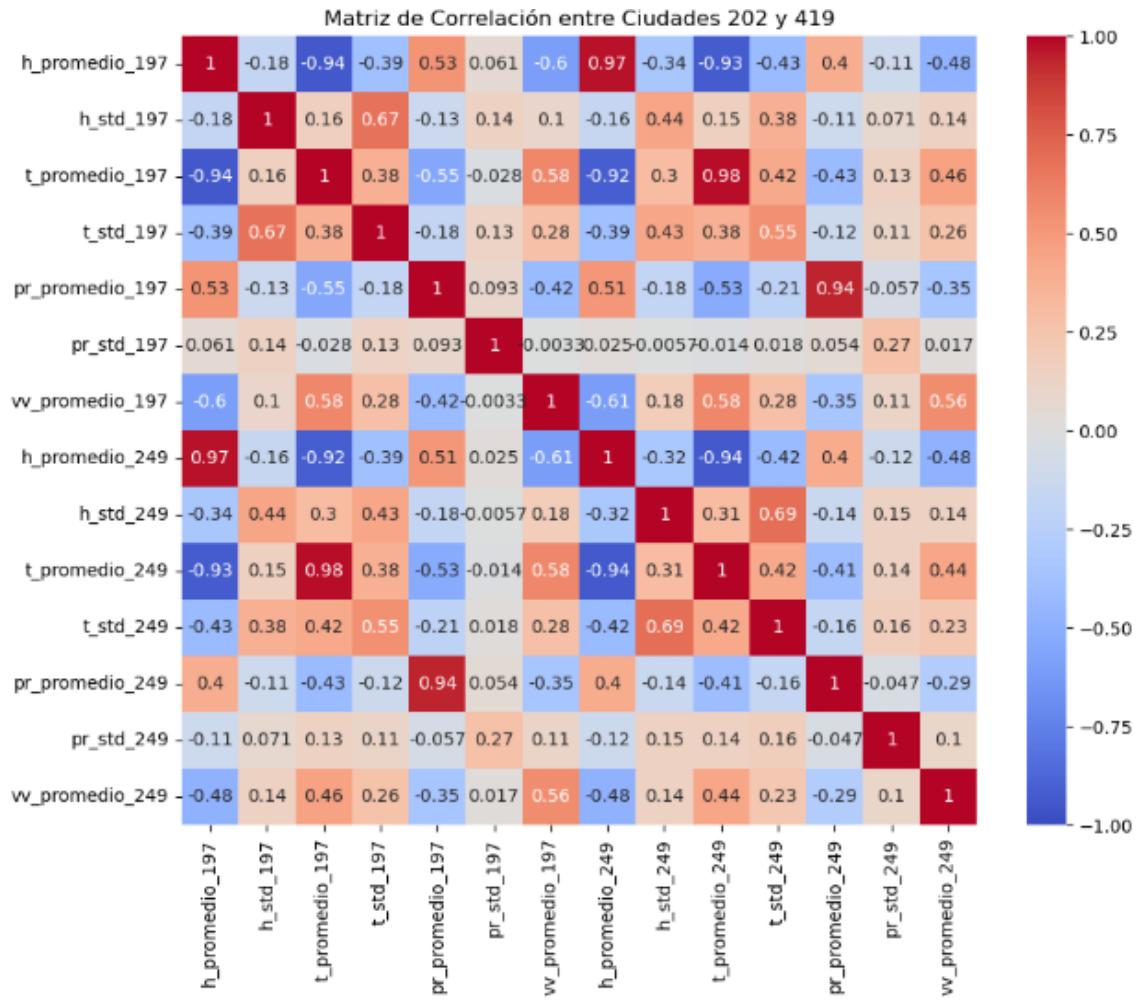
```
Conjunto Prueba
Error Absoluto Medio (MAE): 0.014989287805214068
Error Cuadrático Medio (MSE): 0.00040249293190659395
Raíz del Error Cuadrático Medio (RMSE): 0.02006222649425018
Precisión del modelo (R2): 0.9803366678814742
```



Se genera un script para generar métodos de imputación para completar los datos faltantes, en caso de que no se logre obtener la fecha para los dos puntos en caso que no encuentre información se realizará una imputación donde se va interpolar linealmente los datos.

Por último, se agrupan las fechas de minuto a 15 minutos para disminuir más el ruido generado por los datos de imputación y llevando a la Ley de los Grandes Números y creando nuevas variables.

Resultado:



Conformidad

```
root
|-- codigo: integer (nullable = true)
|-- fecha_hora: timestamp (nullable = true)
|-- h: double (nullable = true)
|-- t: double (nullable = true)
|-- pr: double (nullable = true)
|-- vv: double (nullable = true)
|-- vv_max: double (nullable = true)
|-- dv: double (nullable = true)
|-- dv_max: double (nullable = true)
|-- p: double (nullable = true)
|-- calidad: integer (nullable = true)
```

```
[*]: columnas_fecha = ['fecha_hora']

for columna in columnas_fecha:
    try:
        df = df.withColumn(columna, to_timestamp(col(columna), "yyyy-MM-dd HH:mm:ss"))
    except Exception as e:
        print(f"Error al convertir la columna {columna}: {e}")

df.show(truncate=False)
df.printSchema()
```

Precisión

```
from pyspark.sql.functions import col, min, max, lag, unix_timestamp
from pyspark.sql.window import Window

fecha_minima_por_codigo = df.groupBy("codigo").agg(min("fecha_hora").alias("fecha_minima"))
fecha_maxima_por_codigo = df.groupBy("codigo").agg(max("fecha_hora").alias("fecha_maxima"))

fecha_minima_global = fecha_maxima_por_codigo.agg(min("fecha_maxima").alias("fecha_minima_global")).collect()[0]["fecha_minima_global"]
fecha_maxima_global = fecha_minima_por_codigo.agg(max("fecha_minima").alias("fecha_maxima_global")).collect()[0]["fecha_maxima_global"]

print(f"Fecha mínima entre las fechas máximas: {fecha_minima_global}")
print(f"Fecha máxima entre las fechas mínimas: {fecha_maxima_global}")

df = df.filter((col("fecha_hora") > fecha_minima_global) & (col("fecha_hora") < fecha_maxima_global))

df = df.orderBy("codigo", "fecha_hora")
```

Duplicidad

```
[34]: from pyspark.sql.functions import col, to_timestamp

columnas_seleccionadas = ['codigo', 'fecha_hora', 'h', 't', 'pr', 'vv', 'vv_max', 'dv', 'dv_max', 'p', 'calidad']
df = df_filas_con_problemas_calidad.select(*columnas_seleccionadas)

df = df.dropDuplicates()
```

Integridad

Se verifica que los código sean correspondientes a la regla de negocio realizando las uniones correspondientes.

```
[23]: import pandas as pd

estacion_nivel = pd.read_csv('zoneCleaner/estacion_nivel.csv')
meteologica = pd.read_csv('zoneCleaner/meteologica.csv')
pluvio = pd.read_csv('zoneCleaner/pluvio.csv')
```

```
*[30]: codigo_mapping = {
    355: 619,
    207: 619,
    202: 311,
    419: 311,
    249: 4,
    197: 25
}
meteologica['codigo_mapeado'] = meteologica['codigo'].map(codigo_mapping)
pluvio['codigo_mapeado'] = pluvio['codigo']
estacion_nivel['codigo_mapeado_2'] = 93

merged_1 = pd.merge(
    meteologica,
    pluvio,
    on=['codigo_mapeado', 'fecha_hora'],
    how='inner'
)
merged_1["codigo_mapeado_2"] = 93
final_merged = pd.merge(
    merged_1,
    estacion_nivel,
    on=['codigo_mapeado_2', 'fecha_hora'],
    how='inner'
)

final_merged
```

Carga

Una vez transformados, los datos se almacenan en la capa processed del data lake en S3, donde están listos para análisis ad-hoc mediante Athena o para su carga en Redshift para consultas más complejas. Este almacenamiento estructurado permite consultas rápidas y asegura que los datos sean accesibles para los usuarios de negocio y los sistemas de análisis. Además, los datos procesados pueden ser integrados

con herramientas de visualización como QuickSight para ofrecer insights de alto impacto

En la carga abra diferentes lagos donde se obtendrá y procesará la información estas son las siguientes.

Capa de almacenamiento con nombre raw aquí se hospedarán los datos sin procesar pedidos por el SIATA e ingresados por el servicio streaming separándolos por siata online y siata offline.

Capa de pruebas llamada Zona Enriched se procesará la información realizando un análisis exploratorio para determinar la información, como también una estandarización a nivel de negocio para reunir y distribuir la información en un solo estado, se crearán esquemas y datos para el mayor uso posible en el negocio de acuerdo a los datos estructurados.

Capa de datos espontáneos Zona Staging, en esta zona se ingresará la información entrante por medio de la api rest

Ambiente Tecnológico:

- **Amazon S3:**
 - Capa de almacenamiento principal para datos en bruto (**raw**) y procesados enriched
 - Los datos se organizan por tipo (batch o streaming) y contexto (SIATA online y offline).

 zoneEnriched/	Carpeta
 zoneRaw/	Carpeta
 zoneStaging/	Carpeta
- **EMR**
 - Clústeres para procesamiento distribuido de grandes volúmenes de datos.
 - Ideal para transformaciones complejas y paralelización con Spark.

Clusters (4) Info		View details	Terminate	Clone	Create cluster
		Filter clusters by creation date-time		< 1 > ⚙️	
<input type="checkbox"/>	<input type="checkbox"/>	Cluster ID	Cluster name	Status	Creation time (UTC-05:00)
<input type="checkbox"/>	<input checked="" type="checkbox"/>	j-30OUFIFYYEXBJ	ClusterSiata	⌚ Waiting Ready to run steps	November 29, 2024, 22:30 1 hour, 4

- **EC2:**
 - Servidores virtuales para ejecución de scripts ETL personalizados en Python.
 - Usado para procesamiento de datos batch y streaming, almacenamiento temporal y carga a S3.
- **Athena:**
 - Motor de consultas SQL sobre datos almacenados en S3.

Sistemas de almacenamiento:

Sistemas de archivos distribuidos:

S3

Herramientas y frameworks:

- Apache Spark,
- Pandas

Origen de los Datos

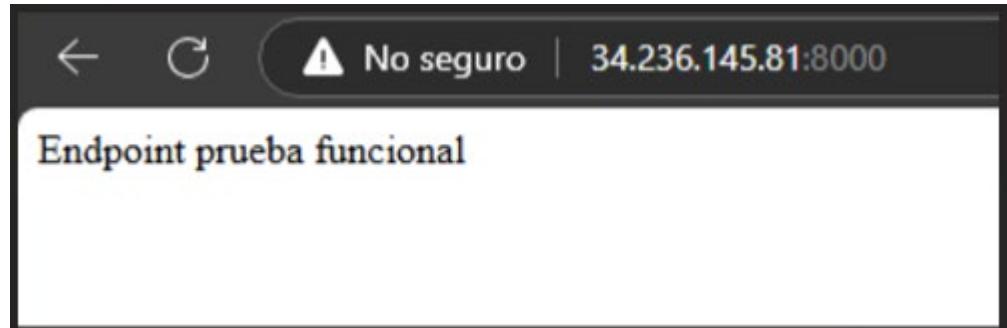
- Consumo del api
- Archivos estructurados CSV por medio de catalogo

6) Despliegue del Proyecto

Caso Hipotético de Implementación Real:

Prototipo desplegado en AWS utilizando herramientas como EC2

```
(venv) [ec2-user@ip-172-31-22-244 proyecto]$ flask run --host=0.0.0.0 --port=8000
 * Serving Flask app 'main.py'
 * Debug mode: off
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:8000
 * Running on http://172.31.22.244:8000
Press CTRL+C to quit
```



Enlace de despliegue para revisión de funcionamiento: 34.236.145.81:8000

Consumo para predicción: 34.236.145.81:8000/predict

Method post

```
json {
    'h_std':0,
    'h_sum':0,
    't_std':0,
    't_sum':0,
    'pr_promedio':0,
    'pr_std':0,
    'pr_sum':0,
    'vv_sum':0,'p_sum':0
}
```

•Naturaleza de las Fuentes de Datos:

•Batch: Carga periódica desde sistemas fuente.

•Streaming: Procesamiento en tiempo real desde eventos o sensores.

·Ingesta de Datos:

·Lagos de datos: Uso de S3 para almacenamiento centralizado.

·Archivos: Integración desde fuentes externas.

·Ambientes de Procesamiento:

·Desde herramientas básicas como Python y Pandas hasta plataformas avanzadas como Apache Spark.

·Aplicaciones:

·APIs para consumo de datos procesados.

·Generación y uso de archivos de salida (CSV, JSON).

5. CONCLUSIONES

- La limpieza, filtrado y transformación de los datos fueron fundamentales para garantizar la precisión de los modelos predictivos. Los procesos de normalización, imputación de datos faltantes y reducción de dimensionalidad demostraron ser esenciales para extraer valor significativo de los grandes volúmenes de datos suministrados por el SIATA.
- El modelo predictivo desarrollado, basado en la correlación de eventos en distintas zonas de Medellín, es una herramienta eficaz para generar alertas tempranas. El uso de algoritmos como CatBoost y Random Forest probó ser efectivo, especialmente para minimizar los falsos negativos en la predicción de lluvias.
- La implementación de una arquitectura Lambda que combina procesamiento batch y streaming en un data lakehouse permitió manejar datos en tiempo real y acumulados con alta calidad. Esto optimiza el análisis y las visualizaciones, alineándose con las necesidades operativas y estratégicas del SIATA.
- Este proyecto no solo tiene un impacto técnico y ambiental, sino que también fortalece la participación ciudadana. La disponibilidad de información clara y accesible fomenta la conciencia colectiva sobre el cambio climático y los riesgos asociados, ayudando a construir una comunidad más resiliente.

6. REFERENCIAS BIBLIOGRÁFICAS

- ∉ **SIATA. (s.f.).** Sistema de Alerta Temprana del Valle de Aburrá. Recuperado el 10 de octubre de 2024 de https://siata.gov.co/sitio_web/index.php/
- ∉ National Oceanic and Atmospheric Administration (NOAA). (2012). *Guía de referencia para sistemas de alerta temprana de crecidas repentinas 2012* (University Corporation for Atmospheric Research, Ed.). Estados Unidos.
- ∉ Unidad Nacional para la Gestión del Riesgo de Desastres (UNGRD). (n.d.). *Guía de referencia para sistemas de alerta temprana de crecidas repentinas*. Reportorio de Gestión del Riesgo. <https://repositorio.gestiondelriesgo.gov.co/handle/20.500.11762/38585>
- ∉ Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, 19(2), 171–209. <https://doi.org/10.1007/s11036-013-0489-0>
- ∉ Silberschatz, A., Korth, H. F., & Sudarshan, S. (2020). *Database System Concepts* (7.^a ed.). McGraw-Hill Education.
- ∉ White, T. (2015). *Hadoop: The Definitive Guide* (4.^a ed.). O'Reilly Media.
- ∉ Strang, G. (2016). *Introduction to Linear Algebra* (5.^a ed.). Wellesley-Cambridge Press.
- ∉ Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2.^a ed.). OTexts. <https://otexts.com/fpp2/>
- ∉ Khan Academy. (s.f.). *Linear Algebra y Probability & Statistics*. Recuperado de <https://www.khanacademy.org/>