MODELO DE PREDICCIÓN DE PRECIPITACIÓN ACUMULADA PARA UN DEPARTAMENTO DE COLOMBIA POR MEDIO DE LA IMPLEMENTACIÓN DE REDES NEURONALES RECURRENTES (LSTM) E INTEGRACIÓN DE DATOS SATELITALES

Jorge Iván Gómez Sepúlveda CC: 1.098.734.464

Jonathan Andres Lafaurie Suarez CC: 1.140.849.411

María Camila Mendoza García CC: 1.097.403.284

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director

David Arango Londoño

CC: 1.130.586.950

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, DICIEMBRE 05 DE 2024

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: “Modelo de predicción de precipitación acumulada para un departamento de Colombia por medio de la implementación de redes neuronales recurrentes (LSTM) e integración de datos satelitales”

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Meteorología y Ciencia de Datos
4. ESTUDIANTE (S): Jorge Iván Gómez Sepúlveda, Jonathan Andrés Lafaurie Suarez y María Camila Mendoza García.
5. CORREO ELECTRÓNICO: joanlasu@javerianacali.edu.co, mariacmendozag5@javerianacali.edu.co y joigos@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Calle 11 # 10-33 Floridablanca/Santander - 3183335544, Carrera 74 # 80 – 124 Casa 101 – 3234789340. Calle 36 # 19-50 Calarcá/Quindío – 3207140076.
7. DIRECTOR: David Arango Londoño
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Cátedra
9. CORREO ELECTRÓNICO DEL DIRECTOR:
10. CO-DIRECTOR (ES) (Si aplica): david.arango@javerianacali.edu.co
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): NO APLICA
12. OTROS GRUPOS O EMPRESAS: NO APLICA
13. PALABRAS CLAVE (al menos 5): Redes Neuronales, LSTM, ciencia de datos, Imágenes satelitales y Precipitación acumulada
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Objetivo 13: Acción por el Clima.
15. FECHA DE INICIO (Desarrollo del proyecto): 17/10/2023
16. RESUMEN (máximo 400 palabras): Este proyecto se enfoca en la predicción de la precipitación acumulada en el departamento del Valle del Cauca en Colombia, siendo una región la cual está altamente influenciada por factores climáticos variables dada su geografía y la ocurrencia de fenómenos temporales como “La Niña” o “El Niño”, los cuales generan cambios en los niveles de precipitación y afectan significativamente diversos sectores como la agricultura, la ganadería, el transporte y la economía en general. Dado esto, se propone el desarrollo de un modelo predictivo que hace uso de redes neuronales recurrentes (LSTM), a partir de información de precipitación observada (medidas terrestres) y satelital. Este enfoque permite superar los limitantes de otros métodos convencionales de series de tiempo y, de esta forma, mejorar la precisión y el rendimiento de los modelos actuales. Los objetivos específicos en este proyecto incluyen factores como la selección del departamento más idóneo para la investigación, el análisis temporal y espacial de la base de datos empleada para el estudio, la instauración y evaluación del modelo LSTM y la comparación con otros modelos tradicionales de series de tiempo. Todo esto está encaminado para el desarrollo de un modelo de predicción que logre estimaciones de la precipitación semanal acumulada.

El proyecto tiene como valor agregado la integración de información satelital por medio del procesamiento de imágenes satelitales y su potencial radica en su aplicación en futuras investigaciones que puedan convertirla en un recurso valioso para diferentes agentes y autoridades relacionadas con el clima y la meteorología. Además, se aspira a que pueda escalarse hacia otras regiones del país, contribuyendo al manejo adecuado de recursos y la planificación meteorológica.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

La precipitación es uno de los fenómenos climáticos más determinantes para el desarrollo económico y ambiental de muchas regiones, particularmente en Colombia, un país caracterizado por su diversidad geográfica y por la dependencia de varios sectores económicos respecto a los niveles de precipitación. Con un promedio anual de 3,240 mm de precipitación, Colombia se destaca por su abundancia de lluvias. Sin embargo, la marcada variabilidad climática y eventos como el fenómeno de "La Niña" o “El Niño” generan incertidumbre en la distribución y cantidad de las precipitaciones, lo que aumenta la vulnerabilidad en sectores clave como la agricultura, la energía hidroeléctrica, la gestión de recursos hídricos, entre otros.

Teniendo en cuenta lo anterior, la predicción de la precipitación es, por tanto, una herramienta indispensable para mitigar los impactos adversos de estos cambios climáticos en la economía y en el bienestar de la población. Sin embargo, este proceso presenta grandes desafíos debido a la complejidad y la naturaleza altamente no lineal de los fenómenos meteorológicos. Tradicionalmente, los modelos dinámico-estadísticos han sido utilizados para predecir los niveles de precipitación. No obstante, estos modelos enfrentan limitaciones significativas en términos de precisión y capacidad predictiva a largo plazo, lo que dificulta la toma de decisiones informadas en sectores que dependen del clima.

En este contexto, el presente proyecto se enfoca en el desarrollo de un modelo predictivo de precipitación acumulada semanal para un departamento específico de Colombia, utilizando redes neuronales Long Short-Term Memory (LSTM) a partir de información procesada de imágenes satelitales. Para lograrlo, el proyecto se estructura en varios objetivos clave: la selección del departamento más adecuado para el estudio, el análisis espacial y temporal de las series de precipitación acumulada, el diseño y entrenamiento del modelo LSTM, y la comparación de su rendimiento frente a métodos tradicionales de series temporales.

Por último, los resultados de este proyecto incluyen la evaluación precisa del modelo en términos de su capacidad para predecir la precipitación semanal en la región seleccionada. Se obtiene el modelo LSTM desarrollado supera en precisión a los modelos convencionales y puede llegar convertirse en un punto de partida para futuras investigaciones que exploren y fortalezcan el uso de redes neuronales en la predicción climática. Además, el modelo tiene el potencial de ser una herramienta útil para la gestión de recursos y la planificación preventiva frente a eventos extremos de precipitación en diferentes regiones de Colombia, contribuyendo así a la mitigación de los riesgos asociados al cambio climático.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

La precipitación, como evento climático, tiene un papel importante en el desarrollo económico de las regiones alrededor del mundo, principalmente en Colombia, que, aparte de ser catalogado como un país con geografía diversa y dependencia significativa a la agricultura (en algunos departamentos en específico), se establece que, con un promedio de 3.240 mm por año, es considerado uno de los países con mayor nivel de precipitación. [1]

Asimismo, es importante resaltar que la ocurrencia e intensidad de la precipitación en el país está sujeta a una gran incertidumbre o variabilidad, ya que es afectada por eventos que suceden en la Zona de Convergencia Intertropical (ZCIT), que cuando se desplaza hacia el norte de su posición media, trae consigo condiciones de lluvia intensas y, por el contrario, si se desplaza hacia el sur, las precipitaciones disminuyen notablemente y se presentan sequias. [2]

Adicionalmente, cabe mencionar que eventos climáticos como “La Niña” (Aumento del nivel de precipitación), en especial el ocurrido en el país en el período 2010-2011, afectó en un alto grado de vulnerabilidad los sistemas económicos de algunas regiones del país. Dado lo anterior, la información emitida por el Departamento Nacional de Planeación en el 2014 [3] y en apoyo con el IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales), proyectaron los efectos de este evento climático en los sectores económicos de mayor productividad. Para los años 2011- 2100, se tiene que el sector ganadero tendría un promedio de pérdidas anuales en un 1.6%, mientras que el sector agrícola una reducción promedio de los rendimientos para maíz tecnificado, arroz irrigado y papa del 7.4%; en cuánto, al sector pesquero se tendría una disminución promedio de la carga desembarcada del 5.3%.

Añadido a esto, se debe tener en cuenta los deslizamientos que se presentan en las temporadas de lluvias en la infraestructura vial, lo que acarrearía una inhabilidad en las vías del 5.9% del tiempo y retrasaría la entrega de los alimentos. Por lo tanto, la identificación de la poca disposición de herramientas que permitan la estimación de valores futuros de precipitación acumulada en regiones de Colombia se establece como una problemática de atención con alta prioridad. Adicionalmente, los modelos teóricos para precipitación que actualmente se utilizan en Colombia son enfocados a una escala dinámico-estadística que realiza el IDEAM [4].

Cabe mencionar, que desde la perspectiva de la ciencia de datos, el planteamiento del problema se centra en la comprensión y resolución de los desafíos asociados a la predicción de la variable precipitación acumulada y sus correspondientes análisis de la información e implementación de técnicas, que permitan el desarrollo, entrenamiento y evaluación de un modelo predictivo por medio del uso de redes neuronales recurrentes que aumente el rendimiento predictivo en comparación con los métodos tradicionales de series temporales.

Por último, es importante destacar que la falta de estudios actualizados sobre la problemática en cuestión puede incrementar el nivel de incertidumbre relacionada con la estimación de los niveles de precipitación.

De ser así, se podría limitar la efectividad de aquellos que toman decisiones relacionados con diferentes facetas susceptibles a cambios en los niveles de lluvia, incrementando los impactos tanto económicos como sociales por falta de herramientas predictivas.

1.2. FORMULACIÓN DEL PROBLEMA

En consecuencia, la pregunta general de investigación que aborda este proyecto es:

¿Cómo se desarrolla el modelo de precipitación acumulada usando redes neuronales recurrentes (LSTM) a partir de información procesada de imágenes satelitales y mediciones terrestres para un departamento de Colombia?

Con base a la pregunta de investigación, se formulan una serie de subpreguntas que cumplen la función de sistematizar la problemática abordada como se describe a continuación:

- ¿A través de qué métodos se va a realizar la selección del departamento de Colombia a analizar?
- ¿Qué estrategias se utilizan para la exploración espacial y temporal de los datos satelitales para la serie escogida?
- ¿Cómo se lleva a cabo la ejecución del entrenamiento del modelo de predicción?
- ¿Cómo es el comportamiento del modelo de predicción con redes neuronales recurrentes LSTM frente a los métodos tradicionales?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Desarrollar un modelo por medio de la implementación de redes neuronales recurrentes (LSTM) e integración de datos satelitales para la predicción de la precipitación acumulada semanal en un departamento de Colombia.

2.2. OBJETIVOS ESPECÍFICOS

- Definir el departamento de Colombia a utilizar mediante análisis, considerando la cantidad y completitud de la información suministrada por las estaciones meteorológicas y el nivel de afectación económico en la región.

- Realizar un análisis espacial y temporal de la serie de datos de precipitación acumulada obtenida a partir de los datos satelitales en el departamento seleccionado.

- Ejecutar las etapas del modelo de predicción, incluyendo la preparación de datos, entrenamiento de la red LSTM y evaluación del modelo resultante.

- Comparar el rendimiento del modelo LSTM con métodos estándar de series de tiempo a través de las métricas de desempeño como RMSE y el MAE.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

3.1.1. Climatología y precipitación en Colombia

En primer lugar, es necesario describir ¿qué es el clima?, una de las definiciones más aceptadas ha sido la planteada por Lowry [5], que define al clima como un conjunto fluctuante de condiciones atmosféricas, dado por las evoluciones y estados del tiempo en un período determinado del espacio. De esta definición, se desprenden diferentes ideas que aportan características actuales del clima, como el clima ligado a condiciones atmosféricas de un lugar y tiempo determinado acompañado de elementos y fenómenos meteorológicos propios de la ubicación espacial.

En segundo lugar, la Sociedad Americana de Meteorología (AMS) [6], define a la precipitación como cualquier producto que proviene de una condensación del vapor de agua atmosférico que cae desde las nubes en un determinado punto de la superficie terrestre en un lapso. Cabe destacar que la precipitación se puede medir de forma continua o discreta que depende de la ocurrencia o no en un periodo definido.

En tercer lugar, se definen los factores físicos que representan las condiciones de cada evento climático, como la localización geográfica de Colombia, que se encuentra ubicada en el extremo noroccidental de América del sur entre los -4° latitud sur y 13° latitud norte, junto a la Zona de confluencia intertropical (ZCIT), que influyen en las condiciones atmosféricas definiendo el tiempo y clima, al igual que sus condiciones físicas propias del territorio como sus relieves, vegetación, altitudes, distancias litorales, no continentalidad que modifican los elementos propios del estado del tiempo en cada región.

Es así como la ZCIT, es el sistema meteorológico que da lugar al mecanismo general del clima en el país [7]. Cuando un territorio se encuentra influenciado por las zonas de convergencia intertropicales, se presentan condiciones climáticas de abundantes precipitaciones. En Colombia, esta zona fluctúa entre los 0° de latitud, entre enero-febrero y 10° de latitud norte alcanzando su posición extrema en los meses de julio-agosto; cuando la zona se ubica en el centro del país, ocasiona durante el año, dos presencias de altas precipitaciones que definen al territorio como bimodal. El primer período bimodal, sucede en los meses entre abril-mayo cuando se desplaza hacia el norte dando lugar a la primera temporada de lluvias; el segundo período, se presenta entre los meses de septiembre y octubre, en donde las precipitaciones pueden ser fuertes de corta duración o moderados de una duración más prolongada que el primer período.

En cambio, en el norte de la región del pacifico, la ZCIT influencia un proceso de circulación ciclónica encargada de que se presenten sistemas productores de fuertes y abundantes precipitaciones durante todo el año, dando lugar a una de las zonas más lluviosas del mundo con 3240 mm al año.

Cabe agregar que, las componentes oceánicas del ENOS (Oscilación del Sur) que hace referencia a la presencia de aguas superficiales más cálidas (El Niño) o más frías (La Niña) que las características de las condiciones normales en el pacífico tropical central y oriental que ocurre en los litorales del norte de Perú,

Ecuador y sur de Colombia; estos ciclos El Niño y La Niña producen alteraciones en los regímenes de lluvias en Colombia por su variabilidad climática interanual que tienen un período de ocurrencia irregular que varía entre los 20 y 35 meses [8], alternándose los vientos de componente Este con los del Oeste.

3.1.2. Ciencia de datos aplicada al estudio de la precipitación

La creación de un modelo predictivo sobre la precipitación acumulada se abordará bajo la metodología de la ciencia de datos, entendida como una ciencia multidisciplinaria que combina la programación, la estadística y el dominio de experto para resolver problemas complejos a través de grandes volúmenes datos mediante el uso de algoritmos de regresión o clasificación.

A lo largo de los últimos años, la ciencia de datos ha emergido como una disciplina de gran impacto para muchos campos del proceder humano. Por medio de su uso se ha potenciado la toma de decisiones y mitigado el impacto negativo de la incertidumbre en diversos sectores. A su vez, el rápido crecimiento en el rendimiento de los sistemas tecnológicos actuales, así como la proliferación más acelerada de la información, ha hecho más viable el uso de algoritmos complejos que tiempo atrás eran difícil de implementar [9]. Uno de los métodos estándar más empleados actualmente para la ejecución de proyectos en ciencia de datos es la metodología CRISP-DM, la cual orienta a través de diferentes fases los diferentes pasos que se siguen para abordar el ciclo de un problema. Sus fases se pueden evidenciar de la siguiente manera [10]:

1. **Comprensión del negocio**: en esta etapa se realiza una comprensión sobre el contexto de la problemática abordada, la cual orienta la delimitación del proyecto y ayuda a establecer los diferentes objetivos a cumplir.
2. **Exploración de los datos:** mediante esta se realiza un análisis exploratorio a los datos, lo que brinda un primer acercamiento acerca de sus condiciones y estructura, así como de los patrones y tendencias preliminares que los datos pueden brindar.
3. **Preparación y limpieza de los datos:** una vez explorado los datos, es necesario someterlos a una fase de limpieza y preparación mediante diferentes transformaciones que permitan adecuarlos para su uso definitivo en diferentes modelos.
4. **Modelado:** se emplean diferentes algoritmos o métodos que permitan que los datos brinden información valiosa y posibiliten tomar decisiones más acertadas. Para esto, se emplean diversos modelos de Machine Learning o técnicas estadísticas, ajustando correctamente cada uno de sus parámetros y seleccionando las mejores variables o características, dado el problema en cuestión.
5. **Evaluación:** se mide los diferentes modelos generados para determinar su rendimiento y capacidad para predecir o clasificar, lo que permite escoger el más adecuado en términos de la calidad de los resultados.
6. **Despliegue**: en esta última fase se implementan los resultados en un entorno operativo, estableciendo un procedimiento de monitoreo y seguimiento del modelo para verificar que sea estable a medida que se vaya usando.

Cabe destacar que la metodología CRISP-DM no es un proceso del todo secuencial, sino que tiene connotaciones iterativas, en donde se pueden volver a etapas previas dependiendo de las necesidades y conclusiones que se vayan obteniendo a medida que se avanza en un proyecto.

3.1.3. Modelos tradicionales de predicción (ARIMA)

- **Series Temporales**

Las series temporales, se consideran como conjuntos de datos que representan observaciones recopiladas en intervalos regulares a lo largo del tiempo. Estos datos pueden representar una amplia gama de fenómenos, desde datos económicos y financieros hasta mediciones climáticas. La información en series temporales se organiza en función del tiempo, lo que permite analizar patrones, identificar tendencias, estacionalidades y modelar la variabilidad a lo largo de intervalos temporales específicos.

Las características generales más relevantes son [11]:

- **Secuencialidad:** Los datos se recopilan en orden cronológico, lo que establece una secuencia significativa para el análisis.
- **Tendencia**: Puede existir una tendencia general ascendente, descendente o fluctuaciones a lo largo del tiempo en los datos, que puede ser lineal o no lineal.
- **Estacionalidad:** Algunas series temporales presentan patrones repetitivos en intervalos fijos, como estacionalidades diarias, semanales, mensuales o anuales.

La finalidad de las series temporales es desarrollar un modelo de predicción de la información evaluada. Para este caso, se mencionan los procesos más relevantes a cumplir para la selección y creación del modelo adecuado de predicción con base en series temporales:

- **Preprocesamiento:** Corresponde a la limpieza de los datos, manejo de valores faltantes. datos atípicos, detección de anomalías, entre otros. Estos procesos se hacen para garantizar la calidad de los datos antes del análisis.
- **Identificación de Patrones**: Uso de métodos gráficos y/o estadísticos para identificar tendencias, estacionalidades y componentes relevantes en los datos.
- **Estacionariedad:** Identificar si las propiedades estadísticas de la serie temporal (media, varianza, covarianza) son constantes en el tiempo.
- **Modelado y Predicción**: Dependiendo de la serie temporal analizada, se puede establecer el uso de métodos estadísticos como ARIMA, SARIMA, y modelos basados en aprendizaje automático como redes neuronales recurrentes (RNN) y LSTM para modelar y predecir datos futuros.

Dado lo anterior, los métodos de predicción para series temporales se pueden clasificar en dos categorías principales: modelos tradicionales y modelos basados en aprendizaje automático.

En esta sección, se presentará la información relacionada con los modelos tradicionales: ARIMA y el modelo basado en aprendizaje automático, relacionado con Redes Neuronales Recurrentes.

- **Modelos tradicionales de predicción**

La creación de modelos de predicción con el paso del tiempo se ha convertido en una herramienta invaluable en varios campos de todos los sectores, desde la medicina, ingeniería hasta la meteorología. El hecho de poder desarrollar un mecanismo con la capacidad de predecir eventos futuros o estimar valores es

fundamental para la toma de decisiones y una debida planificación estratégica.

Los modelos de predicción, a partir del análisis de datos históricos y del uso de herramientas computacionales, tienen como objetivo la predicción del comportamiento de valores futuros, por medio del establecimiento de un proceso de aprendizaje supervisado [11]. En la literatura, los métodos de predicción para series temporales se pueden clasificar en dos categorías principales: modelos tradicionales y modelos basados en aprendizaje automático. se destacan dos métodos tradicionales de predicción, como lo son: ARIMA y la Regresión Lineal, que han aportado soluciones valiosas antes del auge de las técnicas más modernas de aprendizaje automático.

El modelo ARIMA, siglas en inglés de Autoregressive Integrated Moving Average, es una técnica estadística utilizada en el análisis y la predicción de series temporales, tendencias y comportamientos cíclicos. Este, se establece como un modelo paramétrico que busca obtener la representación de una serie temporal en términos de la interrelación temporal de sus elementos y fue propuesto por Yule y Slutky en la década los 20 [12].

Este método, dentro de todo su proceso de análisis, trabaja con formulaciones estadísticas que permiten establecer dicha interrelación temporal, como lo es el coeficiente de autocorrelación, que mide el grado de asociación lineal que existe entre observaciones separadas n periodos.

A continuación, se detallan los componentes de este modelo:

- **Autoregresión (AR):** Esta componente hace referencia a la relación entre una observación actual y múltiples observaciones anteriores en la serie temporal. En un modelo AR, la observación actual se modela como una combinación lineal de las p observaciones previas, utilizando coeficientes que se determinan durante la estimación del modelo.
- **Integración (I):** Proceso de diferenciación de la serie temporal para hacerla estacionaria. Si la serie temporal tiene tendencia o estacionalidad, se aplican diferencias de primer orden para lograr la estacionariedad.
- **Media móvil (MA):** Modela la relación entre la observación actual y los errores residuales de predicciones pasadas en un modelo de media móvil MA(q), donde q es el orden de la media móvil.

Una vez detallados los componentes, se destacan posibles aplicaciones y ventajas del modelo ARIMA:

- **Predicción de series temporales:** ARIMA ha demostrado ser eficaz en la predicción de datos secuenciales en diversos campos, como finanzas, climatología, ventas y economía.
- **Flexibilidad:** Adaptable a una amplia gama de patrones y comportamientos en series temporales.
- **Identificación de tendencias:** Permite identificar tendencias a largo plazo, lo que resulta valioso para comprender la dinámica temporal de los datos.

Sin embargo, la efectividad del modelo ARIMA se puede ver limitada en casos donde existen relaciones no lineales o cambios abruptos en la serie temporal que se está evaluando, ya que este modelo se establece con base en supuestos lineales y continuidad de los patrones históricos. Pero a pesar de esta limitación, ARIMA sigue siendo considerado como una herramienta valiosa para el análisis y predicción de series temporales.

Teniendo en cuenta la definición y las características presentadas del modelo tradicional detallado, se considera que se ha convertido en piedra angular de procesos de optimización en el campo de la predicción, ya que su menor nivel de complejidad, en comparación con el uso de métodos de predicción basados en aprendizaje automático, ha permitido una buena comprensión de los datos resultantes, pero con las limitaciones que se presentaron a lo largo de la sección.

3.1.4. Redes neuronales recurrentes (RRN)

Las Redes Neuronales Recurrentes (RNN) son un tipo de arquitectura de redes neuronales diseñadas para trabajar con datos secuenciales, donde la información tiene una dependencia temporal entre las observaciones. A diferencia de las redes neuronales tradicionales, las RNN tienen conexiones retroalimentadas que les permiten mantener y utilizar información previa en su procesamiento. Esta capacidad las hace especialmente adecuadas para modelar datos secuenciales y series temporales [12].

Sus características principales son:

- **Memoria a corto plazo:** La característica clave de las RNN es su capacidad para recordar y utilizar información anterior en la secuencia de datos, lo que les permite capturar dependencias temporales a corto plazo
- **Estructura recursiva:** Las RNN utilizan conexiones cicladas en su arquitectura, lo que les permite mantener una especie de memoria interna y procesar secuencias de longitud variable.
- **Flexibilidad en la entrada y salida**: Pueden manejar secuencias de entrada y salida de diferentes longitudes, lo que las hace útiles para tareas como el procesamiento del lenguaje natural (NLP), reconocimiento de voz, traducción automática, entre otros.
- **Aplicaciones en series temporales:** Se utilizan ampliamente en la predicción y modelado de series temporales debido a su capacidad para capturar patrones secuenciales y temporales complejos.

Las RNN tradicionales tienen dificultades para retener información relevante en secuencias muy largas, lo que puede generar dificultades en la modelización de dependencias temporales a largo plazo. Es por esto, que se han diseñado mejoras en la arquitectura RNN, tales como:

- **LSTM (Long Short-Term Memory):** Las LSTM son un tipo de unidad recurrente que supera el problema del olvido a largo plazo en las RNN tradicionales. Introducen compuertas (gates) para regular y controlar el flujo de información, permitiendo retener información relevante durante más tiempo.
- **GRU (Gated Recurrent Unit):** Similar a las LSTM, las GRU son unidades recurrentes que utilizan menos compuertas, lo que resulta en una estructura más simple. A pesar de su simplicidad, las GRU han demostrado ser igualmente efectivas en la retención de información a largo plazo.

A continuación, se presenta la información correspondiente a la RNN seleccionada para el desarrollo del proyecto:

- **Redes Long Short –Term Memory (LSTM)**

Las Redes Neuronales LSTM son un tipo de arquitectura especializada de Redes Neuronales Recurrentes (RNN), diseñadas para superar el problema del olvido a largo plazo que enfrentan las RNN tradicionales. Las LSTM fueron propuestas por Hochreiter y Schmidhuber en 1997, y se han convertido en una de las arquitecturas más utilizadas en el procesamiento de secuencias.

Sus características principales son:

- **Unidades de memoria:** Las LSTM contienen unidades de memoria llamadas "celdas de memoria" que pueden mantener y recordar información durante largos períodos de tiempo.
- **Compuertas (Gates):** Utilizan compuertas (input gate, forget gate, output gate) para regular el flujo de información dentro de la celda de memoria.
- **Olvido a largo plazo reducido:** La estructura de compuertas de las LSTM permite mantener información relevante durante más pasos temporales, mitigando el problema del olvido a largo plazo que afecta a las RNN tradicionales.
- **Conexiones de celda:** Las conexiones de las celdas de memoria a lo largo del tiempo permiten que la información fluya de manera más consistente y controlada en comparación con las RNN tradicionales.

Sus componentes principales son:

- **Celda de memoria:** La unidad central que almacena y procesa información. Mantienen un estado de celda que puede ser modificado a través de las compuertas.
- **Compuerta de olvido (Forget gate):** Decide qué información debe ser descartada de la celda de memoria.
- **Compuerta de entrada (Input gate):** Regula qué nueva información debe ser almacenada en la celda de memoria.
- **Compuerta de salida (Output gate):** Determina qué información debe ser entregada como salida basada en el estado actual de la celda de memoria.

Actualmente, este tipo de red es ampliamente utilizada en la predicción de series temporales, permitiendo capturar patrones temporales complejos y realizar predicciones precisas en datos secuenciales. En resumen, las Redes Neuronales LSTM han demostrado ser una arquitectura poderosa y efectiva en el procesamiento de secuencias, especialmente en tareas donde se requiere retener y procesar información a largo plazo.

3.1.5. Métricas de evaluación

Las métricas de evaluación o desempeño pueden ser absolutas cuando el error el calculado despreciando el signo, relativas al valor del error puede compararse con otras medidas de evaluación y tamaño de los errores si arrojan un valor porcentual en sus cálculos. También pueden clasificarse como de método simple si el valor obtenido de la métrica declara el pronóstico.

En esta investigación para la evaluación del rendimiento del modelo de predicción se desarrolla por medio de las métricas de desempeño absolutas como el RMSE (*Root Mean Square Error*) y el MAE (*Mean Absolute Error*).

3.1.5.1 RMSE (Root Mean Square Error)

Se define RMSE (Root Mean Square Error) como la raíz cuadrada de la media de los errores al cuadrado:

El error cuadrático medio (RMSE) también llamado desviación cuadrática media es una medida de uso frecuente de la diferencia entre los valores pronosticados por un modelo y los valores realmente observados. Estas diferencias individuales son también llamadas residuos y el RMSE sirve para agregar en una sola medida la capacidad de predicción.

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

Donde:

$\hat{y}_1$, $\hat{y}_{2...}$ $\hat{y}_n$ Son los valores predichos

$y_1, y_2 \ldots y_n$ Son los valores observados

n: Es el número de observaciones

3.1.5.2 MAE (Mean Absolute Error)

Se define MAE (Mean Absolute Error) como la magnitud promedio de los errores de un ejercicio de pronóstico sin tener en cuenta su signo, es decir, el promedio de los valores absolutos de los errores calculados:

$$MAE = \frac{1}{n}\sum |y - \hat{y}|$$

Donde:

$\frac{1}{n}$: Se divide en el total de datos

y: Valor del punto actual observado

ŷ: Valor del punto predicho

3.2. MARCO DE ANTECEDENTES

Teniendo en cuenta el marco teórico descrito, se establece que el estudio y análisis de series temporales se puede catalogar como hito fundamental en gran cantidad de campos científicos y aplicaciones prácticas. Es

por esto que, el uso de herramientas analíticas para comprender la dinámica temporal de los datos ha sido sujeto de numerosos procesos de investigación. Desde métodos tradicionales como la Regresión Lineal y ARIMA hasta enfoques más avanzados basados en redes neuronales recurrentes (RNN) y técnicas de aprendizaje profundo.

La sección de antecedentes busca proporcionar una visión integral sobre la importancia y evolución del análisis de series temporales, así como establecer las bases para comprender las metodologías actuales utilizadas en la predicción de datos dinámicos y en el estudio de la variabilidad climática frente a las fluctuaciones observadas en períodos cortos. Durante un año determinado, pueden registrarse valores climáticos que se desvíen de lo considerado normal. La Comisión de Climatología de la Organización Meteorológica Mundial (OMM) define la Normal Climatológica [8], o valor normal, como el promedio de una serie continua de mediciones de una variable climatológica durante un período mínimo de 30 años, a escala local, nacional o mundial. Para el análisis y predicción de la variable de precipitación, es esencial que la base de datos abarque un período de al menos 30 años, lo que permite comprender la dinámica de la variable y las fluctuaciones que puedan presentarse.

A nivel mundial, se han realizado diversas investigaciones relacionadas con la predicción de variables climáticas. En este sentido, con información obtenida a partir del año 1749 hasta el 2018 por el Centro Mundial de Datos SILSO, del Real Observatorio de Bélgica en Bruselas, se llevó a cabo una investigación, la cual tuvo como objetivo principal, predecir el número de manchas solares en el Ciclo Solar 25 mediante la utilización de dos modelos distintos: un modelo de redes neuronales recurrentes Long short-term memory (LSTM) y un modelo Autoregressive Integrated Moving Average (ARIMA).

Los resultados obtenidos revelaron un mejor desempeño por parte del modelo LSTM, evidenciando así, un rendimiento significativamente superior en comparación con el modelo ARIMA. Cuantitativamente, el modelo LSTM demostró una precisión notable con un Root Mean Squared Error (RMSE) de 3.6, mientras que el modelo ARIMA alcanzó un RMSE de 32.6. Esta diferencia, resaltó la superioridad del modelo LSTM en términos de precisión de predicción, mostrando una mejora del 89% en la reducción del RMSE [14]

Por otra parte, en investigaciones relacionadas con la generación de energía renovable como la fotovoltaica, se modelo una predicción de futuros comportamientos de las nubes a través de imágenes, por ello, se hace uso de la técnica híbrida de deep learning junto a una red generativa adversaria (GAN) y el modelo de predicción de memoria a corto plazo (LSTM) [15].

De esta forma, se utilizó la red GAN, que es la generación de las imágenes de las nubes teniendo los vectores latentes aleatorios y el LSTM para que aprendiera los patrones de las imágenes de series temporales de entrada, prediciendo los vectores latentes futuros. Lo anterior, con el fin de evaluar la efectividad de los métodos y las técnicas propuestas, por medio de la comparación de varios modelos de pronóstico para la generación fotovoltaica, conforme a la precisión de cada modelo, en donde se utilizaron imágenes satelitales e información meteorológica.

En relación con lo anterior, se obtuvieron 30.507 imágenes infrarrojas cada 15 minutos tomadas por el satélite Communication, Ocean, and Meteorological Satellite 1 del Centro Nacional de Satélites Meteorológicos de Corea. Al ejecutar la metodología propuesta, se concluyó que el modelo propuesto LSTM-GAN presenta una

mayor precisión de predicción en comparación con CNN-ANN, CNN-LSTM, GRU-GAN y BILSTM-GAN; afirmando que el rendimiento de predicción de los modelos lineales convencionales tiende a ser menos efectivo cuando los datos varían repentinamente, mientras que los modelos no lineales están diseñados para soportar la fluctuación.

Investigaciones realizadas para predicción en modelos de sequias, se desarrolló una predicción a través de un modelo híbrido que involucra la red neuronal a corto plazo y un modelo climático (LSTM-CM) [16], para el proceso de predicción de la sequía se tomó información a escala diaria de variables climáticas como la precipitación, temperatura máxima, temperatura mínima y velocidad del viento.

En el proceso de validación del modelo propuesto, se comparó con el modelo aislado de memoria a corto y largo plazo (LSTM-SA) y modelo de predicción climática GloSea5 (GS5) para evaluar la viabilidad y rendimiento de estos modelos, usando técnicas estadísticas convencionales como el coeficiente de correlación de Pearson, error absoluto medio (MAE), error cuadrático medio (RMSE) y puntaje de habilidad (SS). En primer lugar, el modelo GS5 demostró un buen comportamiento en las predicciones y no desplazó los resultados, no obstante, la predicción de este modelo genera un sesgo causado por las entradas, su estructura y los parámetros.

Posteriormente, se evalúo el modelo aislado LSTM-SA que disminuye notablemente el sesgo, pero las predicciones no tuvieron la capacidad para proyectar la ocurrencia de sequías en períodos de tiempo largos. En contraste, el LSTM-CM proporcionó predicciones mejoradas de sequías en comparación con el LSTM-SA y a su vez, con la capacidad de simulación de procesos físicos de GS5; mejorando así, las limitaciones de los modelos estándar y optimizando sus predicciones.

En conclusión, LSTM-CM en comparación con los de GS5 para las predicciones a 1, 2 y 3 meses de anticipación mejoraron del 29.17 al 54.29, del 22.47 al 34.15 y del 1.75 al 35.09%, respectivamente. LSTM-CM puede detectar con precisión eventos de sequía y mostró menos incertidumbre en la predicción que LSTM-SA y GS5.

En Colombia también se han realizado diferentes investigaciones utilizando modelos de Machine Learning para predecir las sequias y las cantidades acumuladas de lluvia sobre ciertas regiones del país. En primer lugar, Herrera & Aristizábal [17], llevaron a cabo una investigación donde plantean un modelo de clasificación de dos etapas basado en un Random Forest (RF) y un Árbol de decisión en bolsa (DTC), que demostró ser adecuado para la predicción espacial y temporal de eventos de sequía en el departamento del Magdalena. Para entrenarlo usaron como variables explicativas el índice de vegetación, la temperatura de la superficie terrestre, el índice del agua, el índice de sequía multibanda, la evapotranspiración, la humedad del suelo, la humedad superficial del suelo, el índice ENSO multivariado, el índice de oscilación del sur y el índice del Niño oceánico, algunas de las cuales tuvieron que someterse a un proceso de normalización. Cada modelo tuvo una precisión 0,33 y 0,59 para el árbol de decisión y el Random Forest respectivamente.

En segundo lugar, Guerrero [18], probó diferentes modelos y concluyó que las redes neuronales Long Short Term Memory (LSTM) tenían un mejor rendimiento con respecto a otros algoritmos para predecir los niveles de precipitación. El enfoque de esta investigación estaba orientado a identificar las mejores zonas agrícolas que tuvieran un equilibrio adecuado en términos de lluvias. Adicionalmente, se probó tres LSTM tuneados

con diferentes hiperparámetros relacionados con la época, el tamaño del lote y la proporción de datos de entrenamiento y prueba, obteniendo una precisión de 44,2% en el mejor modelo, compuesto con 100, 32 y 80% en los hiperparámetros de época, lote y porcentaje de entrenamiento respectivamente.

4. MODELO PREDICTIVO DE LA PRECIPITACIÓN SEMANAL ACUMULADA

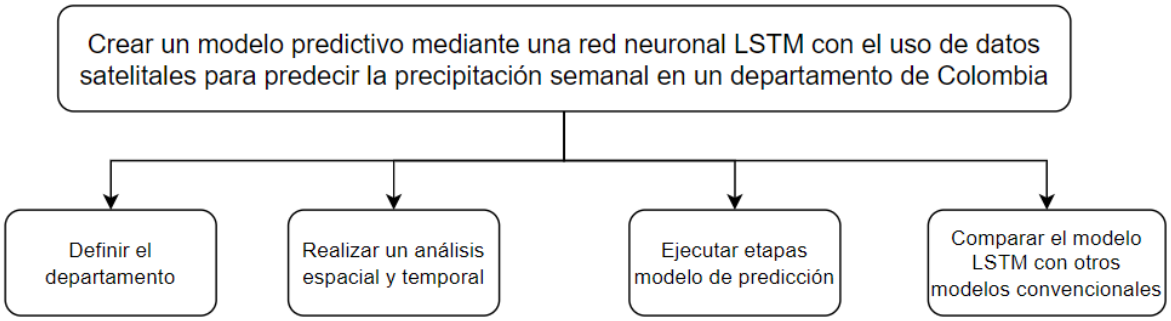

Figura 1. Flujo metodológico objetivo general

La Figura 1, presenta el esquema de flujo de trabajo para la creación del modelo predictivo mediante una red neuronal LSTM, utilizando datos satelitales para predecir la precipitación semanal en un departamento de Colombia. Este diagrama se estructura en cuatro pasos principales, vinculado a cada objetivo específico:

Definir el departamento: El primer paso consiste en seleccionar el departamento de Colombia que será objeto de estudio, en el que se aplicarán los datos satelitales y las técnicas de predicción de precipitación. (Objetivo específico 1)

Realizar un análisis espacial y temporal: En esta fase, se lleva a cabo un estudio detallado de los datos, considerando tanto la dimensión espacial (ubicación geográfica) como la temporal (comportamiento de la precipitación a lo largo del tiempo). Esto permite establecer patrones históricos y preparar los datos para el modelo LSTM. (Objetivo específico 2)

Ejecutar etapas del modelo de predicción: Una vez procesados los datos, se construye y ejecuta el modelo LSTM, que es una red neuronal recurrente adecuada para series temporales, con el objetivo de predecir la precipitación semanal acumulada. (Objetivo específico 3)

Comparar el modelo LSTM con otros modelos convencionales: Finalmente, se compara el rendimiento del modelo LSTM con modelos predictivos convencionales para evaluar su efectividad y precisión en la predicción de la precipitación. (Objetivo específico 4)

Este flujo de trabajo está orientado al desarrollo de un modelo factible, evaluando su precisión a través de la comparación con otros enfoques.

5. DEFINICIÓN DE ÁREA GEOGRÁFICA DE ESTUDIO

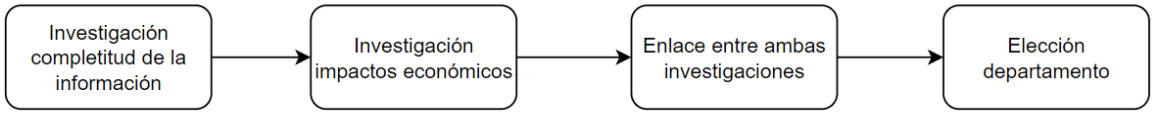

Figura 2. Flujo metodológico objetivo específico 1

El diagrama de flujo de la Figura 2, representa los pasos generales para seleccionar un departamento de Colombia para el proyecto de investigación basado en el análisis de datos económicos y la disponibilidad de información. Este proceso se organiza en las siguientes etapas:

Investigación sobre la completitud de la información: En este primer paso, se lleva a cabo una evaluación exhaustiva de la disponibilidad y calidad de la información necesaria para el análisis, asegurándose de que los datos requeridos estén completos y sean adecuados para el estudio.

Investigación de impactos económicos: A continuación, se realiza una investigación centrada en los impactos económicos que pueden estar relacionados con la precipitación, tales como la agricultura, infraestructura o turismo. Este análisis económico es clave para determinar la relevancia del estudio en el departamento a escoger.

Enlace entre ambas investigaciones: En esta fase, se conecta la información obtenida de las dos investigaciones previas (disponibilidad de información y análisis económico), para identificar posibles relaciones entre la calidad de los datos y los impactos económicos. Este paso es crucial para establecer un enfoque integral y justificado para la elección del departamento.

Elección del departamento: Finalmente, se selecciona el departamento más adecuado para el estudio, considerando tanto la completitud de la información disponible como la relevancia de los impactos económicos identificados en fases previas.

Este flujo de trabajo garantiza una toma de decisiones fundamentada y orientada a obtener resultados útiles y aplicables en la investigación.

El desarrollo y los resultados obtenidos en relación con el primer objetivo específico se basaron en diversas actividades que permitieron recopilar, procesar y evaluar la calidad de la información meteorológica requerida para el análisis de la precipitación en el contexto de este estudio. A continuación, se describe en detalle el proceso implementado:

En primer lugar, la información meteorológica utilizada en este estudio fue suministrada por el Director del proyecto, quien proporcionó datos históricos de precipitación recopilados por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Esta información fue entregada en formato de archivos planos (.txt), con un archivo para cada estación meteorológica incluida en el estudio. Cada archivo contenía los registros históricos de precipitación correspondientes a su estación. Estos archivos de texto se consideran

como la base de datos inicial sobre la cual se realizan las fases posteriores de procesamiento y análisis.

Para el procesamiento de la información recopilada, se utilizaron las herramientas Jupyter Notebook y R Studio, las cuales, gracias a su amplia variedad de códigos y bibliotecas predeterminadas, se consideran idóneas para el desarrollo de los códigos necesarios para consolidar y organizar los datos meteorológicos.

Inicialmente, se diseña un código que permite crear una base de datos estructurada que incluye la totalidad de la información de todas las estaciones meteorológicas. El objetivo, es consolidar los registros de precipitación en una única estructura de datos que facilite su posterior análisis.

Este proceso, inicia con la definición de una ruta de acceso local que contiene los archivos .txt de las estaciones meteorológicas. A partir de allí, se realizan iteraciones sobre los archivos en el directorio para extraer y procesar la información de cada estación. Se crea una lista vacía, denominada data_prec, para almacenar los datos de precipitación de cada archivo. Es importante resaltar, que los nombres de los archivos permiten identificar tanto el código de la estación meteorológica correspondiente como el tipo de medición realizada. Posteriormente, los datos contenidos en cada archivo de texto son leídos y transformados en DataFrames, para facilitar su manipulación y análisis. Paso seguido, las fechas de cada registro fueron estandarizadas al formato YYYY/MM/DD para asegurar la consistencia de los datos, y se agrega una columna con el código de la estación correspondiente a cada archivo.

Una vez consolidada la información en una estructura de datos común, se procede a evaluar la calidad de los datos obtenidos. Esta evaluación es esencial para asegurar que los datos procesados sean confiables y adecuados para su utilización en modelos predictivos. Lo anterior, se logra a través de un código específico, que evalúa de manera detallada la calidad de la información, con el objetivo de identificar y corregir posibles problemas en los datos, como valores faltantes, inconsistencias en las mediciones o errores de formato.

Para llevar a cabo esta evaluación, se implementaron varias acciones clave. En primer lugar, se establecieron rutas específicas para acceder a los archivos de datos procesados, y se asignaron variables representativas a los diferentes archivos de precipitación. Posteriormente, el archivo consolidado (prec_data.txt) es leído en un DataFrame que contenía los registros de todas las estaciones meteorológicas. En esta fase, se realiza una conversión de las columnas de fecha al tipo datetime, con el objetivo de facilitar las operaciones y análisis sobre las series temporales. A su vez, los datos fueron ordenados cronológicamente, asegurando que cada registro estuviera dispuesto desde el más antiguo hasta el más reciente, lo que es esencial para la posterior modelización de series temporales.

Durante el proceso de revisión de calidad, se implementan estrategias para identificar posibles problemas en los datos, como la presencia de valores nulos o erróneos, y se realizan acciones correctivas en caso de considerarse necesario. También, se verifica que la información esté correctamente vinculada a la matriz de códigos de estaciones meteorológicas por departamento, lo que permite asociar los registros de precipitación con las regiones geográficas correspondientes. Esta vinculación es esencial, ya que asegura que los análisis posteriores tengan una codificación territorial precisa.

El proceso descrito permite obtener una base de datos robusta, organizada y de alta calidad, que constituye el punto de partida para los análisis y modelamientos posteriores que se realizan en el marco de este proyecto. Esta base de datos contiene toda la información necesaria para desarrollar modelos predictivos de

precipitación, así como para llevar a cabo comparaciones y análisis que permitan evaluar el comportamiento de la precipitación. Asimismo, es importante destacar que el uso de herramientas de programación avanzadas permite automatizar y optimizar el procesamiento de grandes volúmenes de datos, garantizando un manejo eficiente y fiable de la información a lo largo de todo el proyecto. A continuación, se presentan los primeros registros de la base de datos df_prec:

Tabla 1. Primeros registros del DataFrames df_prec

ID	Date	StationCode	Value
20339567	1968-01-01	48015010	NaN
20339568	1968-01-02	48015010	NaN
20339569	1968-01-03	48015010	NaN
20339570	1968-01-04	48015010	NaN
20339571	1968-01-05	48015010	NaN

Seguidamente, se continúa con el desarrollo de las actividades que complementan el procesamiento de los datos meteorológicos, enfocándose específicamente en el manejo y análisis de datos faltantes. Este conjunto de actividades resulta esencial para asegurar la integridad y confiabilidad de la base de datos antes de su utilización en el modelo predictivo. Teniendo en cuenta lo anterior, se procede a presentar las acciones que se llevan a cabo de manera secuencial:

Para empezar, se procede con el cálculo de la información faltante, lo cual permite identificar la cantidad total de datos faltantes en el conjunto de datos meteorológicos. Esta etapa es importante, dado que los valores ausentes pueden afectar significativamente el rendimiento de los modelos predictivos si no se gestionan adecuadamente. Para este fin, se desarrolla un código que calcula tanto el total de datos disponibles como la cantidad de datos faltantes por tipo de medición (en este caso, precipitación). También, se genera el porcentaje de datos faltantes. Este análisis porcentual (Figura 3) resulta indispensable para evaluar el impacto de los datos faltantes en los resultados finales del estudio y determinar la necesidad de aplicar técnicas de imputación o de eliminación de registros.

```
Para Precipitación: Total de datos = 21806050, Datos faltantes = 1477675, Porcentaje faltante = 6.78%
```

Figura 3. Información General del DataFrame df_prec

Una vez calculado, se desarrollan funciones específicas para la creación y manipulación de DataFrames que acumulan los datos faltantes por mes, lo que permite un análisis más detallado y la visualización de los patrones temporales de los valores faltantes. Para todas las estaciones meteorológicas, se genera una visualización (Figura 4) que denota el acumulado mensual de datos faltantes, permitiendo identificar períodos de tiempo en los que la recolección de datos fue menos precisa o no fue registrada. Así mismo, se destaca que el uso de gráficos y diagramas permite no solo cuantificar los datos faltantes, sino también ofrecer una herramienta visual que facilite la comprensión de la magnitud del problema a lo largo del tiempo.

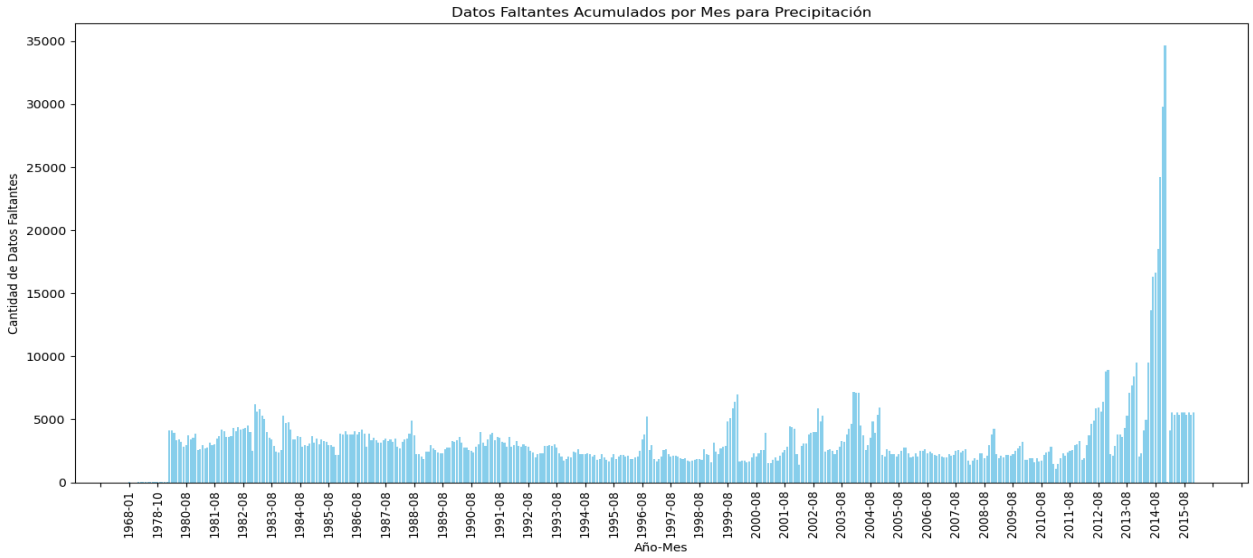

Figura 4. Datos faltantes acumulados por mes – df_prec

Con respecto a lo anterior, se realiza la carga de datos de las estaciones meteorológicas desde un archivo Excel que contenía información detallada sobre las estaciones utilizadas en el estudio (Figura 5). Este archivo incluye datos sobre la ubicación geográfica de cada estación, su código de identificación y la asociación de cada estación a un departamento en Colombia. Esta información, fue cargada y procesada en un DataFrame para facilitar su análisis y manipulación. Durante este proceso, se realiza una evaluación de la distribución de estaciones por departamento, lo cual permite identificar las regiones con mayor densidad de estaciones meteorológicas, así como aquellas que cuentan con un menor número de estaciones. Este análisis es fundamental, ya que la cantidad de estaciones en cada departamento (Figura 6) puede influir en la calidad y representatividad de los datos utilizados para modelar la precipitación. Asimismo, se identifican posibles áreas geográficas con menor cobertura de estaciones, lo que puede constituir una limitación en términos de la fiabilidad de las predicciones para esas regiones.

Tabla 2. Información Base de Datos df_estaciones

	Code	Variable	StartDate	EndDate	Category	Name	Municipality	Departament	Elevation	LongitudeDD	LatitudeDD	LongitudeDMS	LatitudeDMS	Entity	OperativeArea	AirportCity
0	11010010	prec	19900101	20141231	PM	VUELTA LA	LLORO	CHOCO	100	-76.545000	5.458861	76°32'42.0"W	5°27'31.9"N	1.0	1.0	Medellin
1	11020010	prec	19800101	20141231	PM	CARMEN DE ATRAT	EL CARMEN DE ATRA	CHOCO	1850	-76.142083	5.908528	76°8'31.5"W	5°54'30.7"N	1.0	1.0	Medellin
2	11020010	wsmeand	19800101	20141231	PM	CARMEN DE ATRAT	EL CARMEN DE ATRA	CHOCO	1850	-76.142083	5.908528	76°8'31.5"W	5°54'30.7"N	1.0	1.0	Medellin
3	11020020	prec	19800101	19981231	PM	GUADUAS	EL CARMEN DE ATRA	CHOCO	1500	-76.183333	5.766667	76°11'0.0"W	5°46'0.0"N	1.0	1.0	Medellin
4	11020020	wsmeand	19800101	19981231	PM	GUADUAS	EL CARMEN DE ATRA	CHOCO	1500	-76.183333	5.766667					

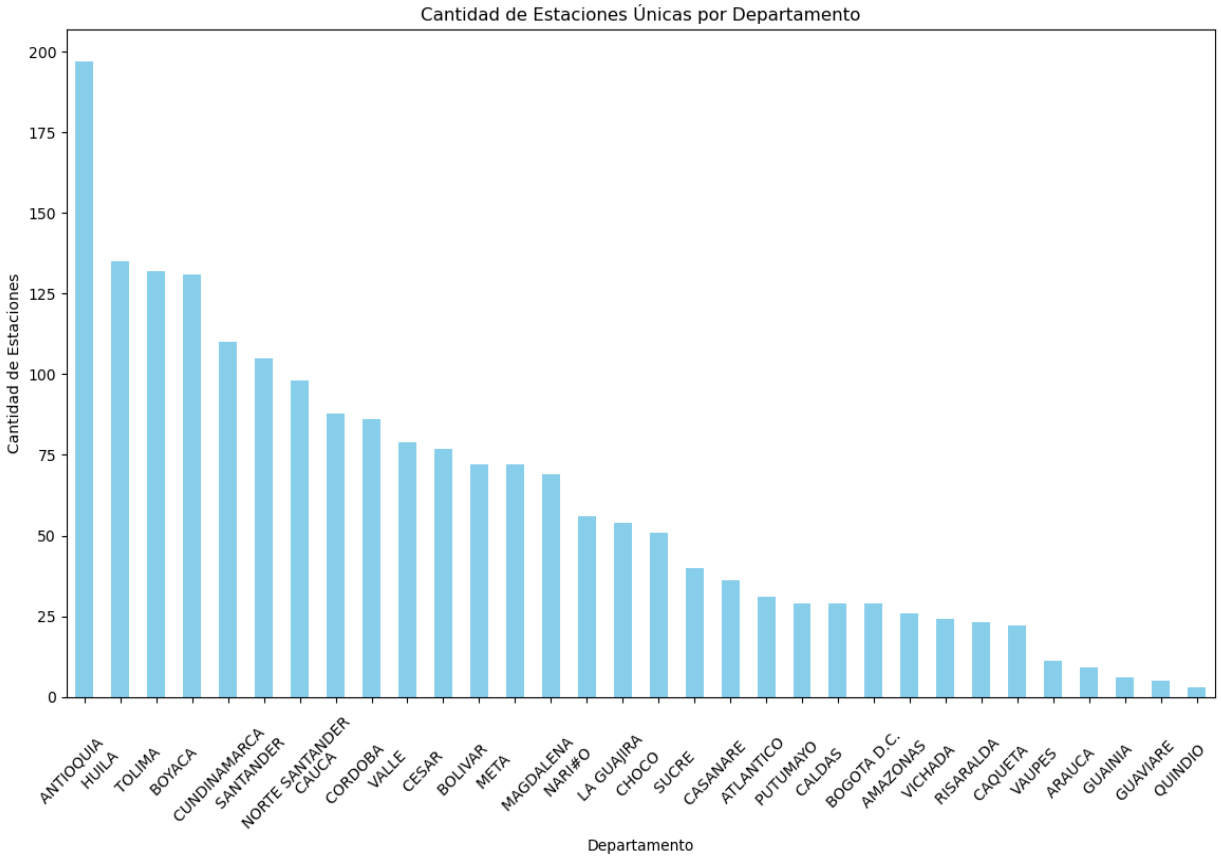

Figura 5. Cantidad de estaciones meteorológicas por departamento - df_estaciones

Tabla 3. Top 10 de departamentos con más estaciones – df_estaciones

Departamento	StationCount
Antioquia	197
Huila	135
Tolima	132
Boyacá	131
Cundinamarca	110
Santander	105
Norte Santander	98
Cauca	88
Córdoba	86
Valle del Cauca	79

En paralelo a este análisis general de los datos faltantes, se llevó a cabo un análisis detallado por departamento, que permite desglosar la información por regiones geográficas. Este análisis se enfoca en identificar los departamentos y estaciones meteorológicas que presentan mayores carencias de información, agregando una evaluación más exhaustiva de la cobertura de los datos. Para cada departamento, se realiza una evaluación comparativa entre las distintas estaciones meteorológicas, permitiendo así identificar

aquellas que registran un porcentaje más alto de datos faltantes. Este enfoque departamental proporciona una visión territorial más clara del problema, destacando las áreas geográficas en las que la disponibilidad de datos es más limitada. Esta información es crucial, ya que la calidad de los datos por departamento influye directamente en la capacidad del modelo para generar predicciones precisas y confiables en diferentes regiones del país.

Tabla 4. Cantidad de Datos faltantes por estación - df_final_aggregated

ID	StationCode	Departamento	Municipio	Datos_faltantes
0	11025010	Antioquia	Ciudad Bolívar	258
1	11060010	Antioquia	Vigia del fuerte	322
2	11070010	Antioquia	Vigia del fuerte	166
...	...	...	...	...
1156	54075040	Valle del Cauca	Buenaventura	2263

Tabla 5. Cantidad de Datos faltantes por departamento - df_departament_summary

Departamento	Total de estaciones	Total datos faltantes precipitación
Antioquia	196	124582
Huila	134	83100
Tolima	132	94261
Boyacá	131	83409
Cundinamarca	110	87728
Santander	105	54384
Norte Santander	98	63364
Cauca	88	53200
Córdoba	85	81363
Valle del Cauca	79	57812

Tabla 6. Cantidad de Datos (porcentaje) faltantes por departamento - df_departament_summary

Departamento	Total de Estaciones	Total datos faltantes precipitación	% datos faltantes
Antioquia	196	124582	0.57
Huila	134	83100	0.38
Tolima	132	94261	0.43
Boyacá	131	83409	0.38
Cundinamarca	110	87728	0.40
Santander	105	54384	0.25
Norte Santander	98	63364	0.29
Cauca	88	53200	0.24
Córdoba	85	81363	0.37
Valle del Cauca	79	57812	0.27

Una vez realizado el análisis de la base de datos, se procede con la evaluación de la contextualización del impacto de la variable meteorológica de precipitación en Colombia. En cuanto a esto, la evaluación económica de los sectores económicos del país, según el informe de la Asociación Nacional de Comercio Exterior (ANALDEX) en el 2023 [19], denota que el Producto Interno Bruto (PIB) de Colombia aumentó positivamente un 0,6 % respecto al año anterior. Este incremento, aunque positivo, representa el crecimiento más moderado en relación con años anteriores, a causa del periodo de la pandemia en 2020. Se destaca que los años 2021 y 2022 fueron períodos de recuperación económica a nivel global.

Según el reporte emitido por el Departamento Administrativo Nacional de Estadística (DANE), en lo que respecta al enfoque de producción del PIB, se observó un crecimiento en actividades económicas productivas en comparación con el año 2022. Específicamente, se destacó un crecimiento del 6,0% en el sector agrícola. Sin embargo, se registraron disminuciones en sectores como la construcción (-1,6%), actividades artísticas (-3,0%) e industrias manufactureras (-4,8%).

De acuerdo con el crecimiento en actividades económicas, en 1991, Banguero [20] afirmó que la región del Valle del Cauca es uno de los sectores económicos más importantes del país, debido a diversos factores. Entre estos, destaca su ubicación estratégica, suelo fértil y relieve plano, así como su acceso al mercado mundial por su proximidad al océano Pacífico y su puerto seco en Buenaventura. A su vez, es el municipio con más lluvia al año en el mundo, con un promedio de 258 días al año [21] y es el tercer departamento con mayor aporte al PIB en Colombia con 9.6%, por detrás solo de Bogotá y Antioquia. Todo esto lo convierte en una región estratégica y de gran impacto económico (ver Figura 10).

Por consiguiente, la región ha apostado por la inversión en capital humano y el desarrollo tecnológico, lo que ha generado avances notables en los sectores industrial y manufacturero. Esto ha contribuido a que el Valle del Cauca se posicione como uno de los departamentos con mayor fortaleza económica en el país. Entre los logros más destacados se encuentran las transformaciones de los cultivos tradicionales, como el arroz, maíz, frijol y caña de azúcar, hacia procesos industrializados con mejores rendimientos. Sin embargo, estos avances podrían enfrentar desafíos debido a la variabilidad climática, lo que representa una vulnerabilidad en el desarrollo económico en la región.

El Impacto de la precipitación en las actividades económicas como lo afirma López y Velásquez [22], se enfoca en el sector agrícola, energético y del sector de alimentos. Por ello, se propusieron modelos que se inscriben dentro del enfoque espacial que buscan estimar efectos en la agricultura con base en las diferencias observadas en los valores de la tierra, la producción agrícola y otros impactos climáticos relacionados entre regiones, utilizando métodos estadísticos para analizar cambios en los patrones espaciales de la producción [22]. Dada esta situación, es crucial reconocer la existencia de la vulnerabilidad de estos sectores económicos frente al aumento de las precipitaciones en comparación con la media histórica.

Por lo tanto, teniendo en cuenta la información obtenida de la descripción de la base de datos (análisis de disposición de información) presentada en la sección anterior, y el análisis del impacto económico en la región, el presente proyecto se realiza para el departamento del Valle del Cauca.

Esta selección, se considera de gran relevancia para el departamento del Valle del Cauca, ya que al implementar un modelo de predicción acumulada de precipitaciones que utiliza redes neuronales LSTM, puede obtener un primer avisto de un pronóstico de esta variable, y así, tener soporte adicional al

establecimiento de criterios y estrategias en cuanto a la evaluación de la exposición/sensibilidad y capacidad adaptativa, para las entidades gubernamentales en la adaptación de los cultivos, las actividades manufactureras, las industrias y los servicios comerciales. A su vez, cuantificar la posible vulnerabilidad a los cambios en el clima, que puede mejorar la resiliencia y adaptabilidad de los sectores económicos del Valle del Cauca, mitigando así el impacto de eventos climáticos extremos en la economía.

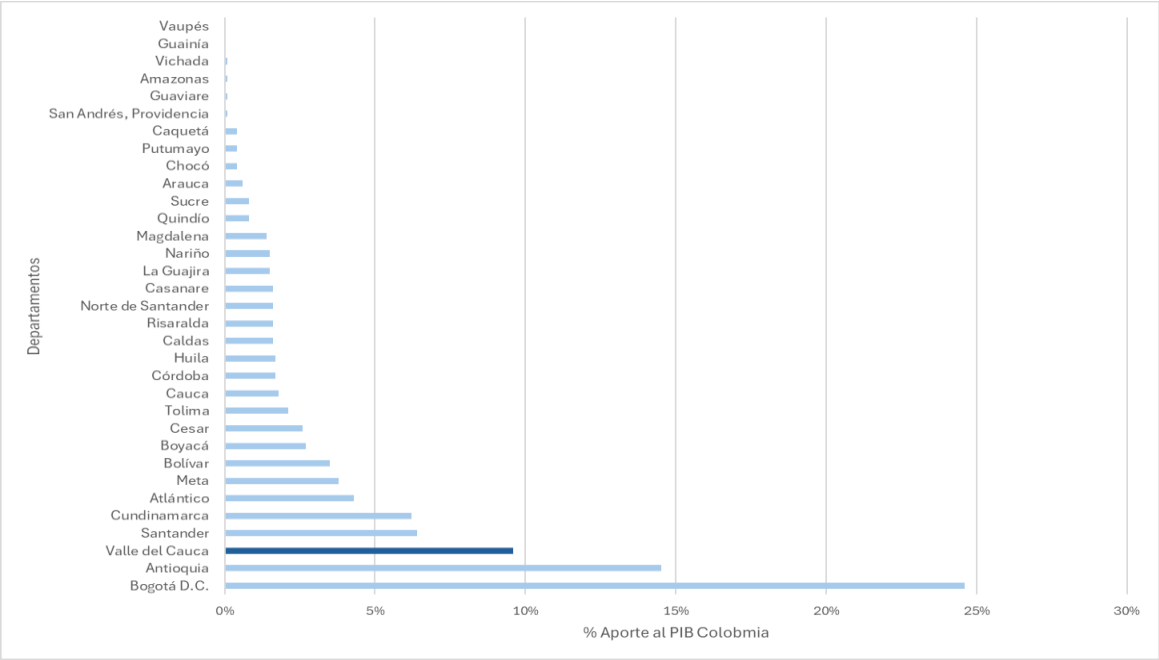

Figura 6. Aporte al PIB de Colombia por departamento
Elaboración: Propia. Fuente: DANE, 2022

6. ANÁLISIS TEMPORAL Y ESPACIAL DE LA PRECIPITACIÓN ACUMULADA

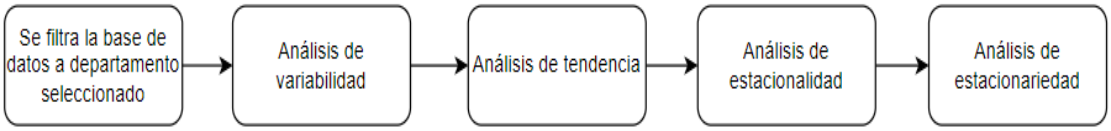

Figura 7. Flujo metodológico objetivo específico 2

La Figura 8, muestra el diagrama de flujo que describe los pasos generales en el análisis de datos para el departamento del Valle del Cauca, enfocado en la predicción de la precipitación semanal mediante un modelo predictivo. Los pasos son los siguientes:

Filtrar la base de datos con el departamento elegido: En esta primera etapa, se filtran los datos de la base de datos general para que solo se incluyan aquellos que correspondan al departamento previamente elegido. Esto permite concentrar el análisis en la región específica de interés.

Análisis de variabilidad: Una vez filtrados los datos, se procede a realizar un análisis de la variabilidad de las precipitaciones en el departamento seleccionado. Este paso busca identificar cómo fluctúan los datos a lo largo del tiempo, lo que es esencial para comprender las posibles variaciones climáticas.

Análisis de tendencia: En este punto, se busca detectar tendencias a largo plazo en los datos de precipitación, como posibles aumentos o disminuciones de los niveles de lluvia en el tiempo. Este análisis es importante para modelar patrones continuos en los datos.

Análisis de estacionalidad: Se examina la presencia de estacionalidad en los datos, es decir, si existen patrones recurrentes de precipitación en ciertas épocas del año. Este análisis es crucial en modelos predictivos, ya que la estacionalidad puede tener un gran impacto en las predicciones climáticas.

Análisis de estacionariedad: Finalmente, se realiza un análisis de estacionariedad para determinar si los datos presentan características estadísticas constantes a lo largo del tiempo. Este paso es fundamental para preparar los datos para ser usados en un modelo LSTM, ya que muchos modelos predictivos requieren que los datos sean estacionarios para obtener resultados precisos.

En esta sección se presentan las actividades desarrolladas para procesar y analizar los datos meteorológicos y satelitales, con el enfoque particular en el departamento del Valle del Cauca. Este proceso, realizado en múltiples fases, permite no solo la depuración y consolidación de la información meteorológica existente, sino también la integración de datos satelitales, lo que en conjunto proporciona una base de datos robusta y precisa para la modelización de la precipitación. A continuación, se detalla cada una de las actividades ejecutadas:

La primera actividad corresponde al procesamiento de los datos de precipitación acumulada para el

departamento seleccionado, en este caso, Valle del Cauca. Para ello, se inicia con la carga de los datos meteorológicos de precipitación desde archivos en formato TXT almacenados en una ruta predefinida. Esta información se estructura en un DataFrame inicial (Tabla 6), que se ajusta para incluir columnas relacionadas con la fecha, el código de estación, datos meteorológicos específicos y detalles geográficos, lo que permite establecer una estructura de datos coherente y uniforme. Adicionalmente, se define un rango temporal de fechas desde 1968 hasta 2015, abarcando todo el período de registro disponible para las estaciones meteorológicas, sin distinción del departamento al que pertenecen.

Tabla 7. Estructura DataFrame Final – Valle del Cauca

Columna	Total no nulos	Tipo de dato
Date	0	datetime64
Code	0	Object
Departament	0	Object
Municipality	0	Object
LatitudeDD	0	Object
LongitudeDD	0	Object
LongitudDMS	0	Object
Elevation	0	Object
Prec	0	Object

Posteriormente, se realiza un proceso de filtrado y preparación de estaciones específicas correspondientes al departamento del Valle del Cauca. En este paso, se seleccionan las estaciones meteorológicas activas en dicha región, asegurando la unicidad de cada estación mediante la eliminación de duplicados. En total, se identifican 79 estaciones para este departamento (Tabla 7).

Tabla 8. Validación estaciones DataFrame Final – Valle del Cauca

Columna	Total no nulos	Tipo de dato
Code	79	Int64
Variable	79	Object
StartDate	79	Int64
EndDate	79	Int64
Category	79	Object
Name	79	Object
Departament	79	Object
Municipality	79	Object
LatitudeDD	79	Float64
LongitudeDD	79	Float64
LongitudDMS	79	Object
LatitudDMS	79	Object
Elevation	79	Int64
Entity	79	Float64

OperativeArea	79	Float64
AirportCity	79	Object
InstallationDate	79	Datetime64
Status	79	Object

La siguiente fase del procesamiento consiste en combinar las fechas con los códigos de estación, mediante el uso de un producto cartesiano, para crear una base de datos (Tabla 8) que vincula cada fecha con cada estación del Valle del Cauca. Esta estructura es clave para la correcta integración de los datos meteorológicos.

Tabla 9. Registros DataFrame Final – Valle del Cauca

Date	Code	Prec	Departament	Municipality	LatDD	LonDD	LatiDMS	LonDMS	Elevation
1980-01-01	26055050	1.0	Valle	Jamundi	3.233333	-76.583333	3°14’0”N	75°35’0”W	1010
1980-01-01	26055070	0.0	Valle	Cali	3.378000	-76.533778	3°22’40,8”N	75°32’1.6”W	985
1980-01-01	26060020	0.0	Valle	Candelaria	3.320444	-76.345972	3°19’13.6”N	75°20’45.5”W	1000
1980-01-01	26065040	NaN	Valle	Candelaria	3.316667	-76.350000	3°19’0”N	75°21’0”W	1000
1980-01-01	26070110	0.0	Valle	Palmira	3.527194	-76.210694	3°31’37.9”N	75°12’38.5”W	1120

Posteriormente, se procede con la integración de los datos meteorológicos, que se lleva a cabo utilizando una serie de uniones secuenciales, que logran la incorporación de la información de precipitación en la base de datos final (df_final). Cabe mencionar, que se manejan cuidadosamente los sufijos para evitar la duplicación de columnas y garantizar que cada estación esté correctamente relacionada con sus correspondientes datos de precipitación.

Una vez consolidada la base de datos meteorológica, se procede a determinar las estaciones meteorológicas de estudio. En esta etapa, se filtran las estaciones con mayor cantidad de datos faltantes. Para lograrlo, se implementa un análisis de la cantidad de datos faltantes por estación, lo que permite identificar aquellas con mayores carencias de información. Adicionalmente, se utilizó un gráfico de barras (Figura 9) para representar visualmente los primeros 20 registros de estaciones con mayor cantidad de datos faltantes, organizados de manera descendente. Esta visualización facilita la identificación de una tendencia estable en los datos faltantes para las estaciones con mayores carencias. Con base en este análisis, se decide excluir estas 20 estaciones del conjunto de datos final, asegurando que las estaciones seleccionadas para el análisis posterior tuvieran un nivel adecuado de integridad de datos.

Tabla 10. Registros df missing_data_sorted

Code	Precipitación	Datos faltantes
26070760	12697	12697
26100300	12676	12676
26060020	12666	12666
54030010	12637	12637
26110160	12595	12595

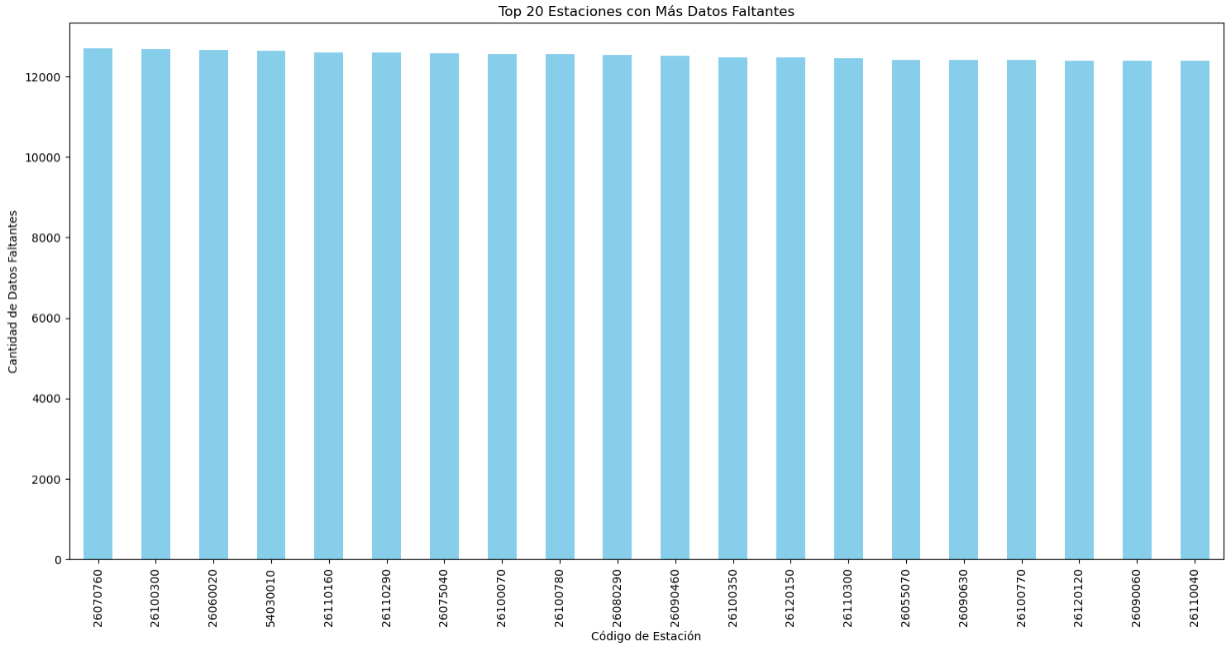

Figura 8. Diagrama de barras - df missing_data_sorted

A continuación, se presentan los códigos de las estaciones seleccionadas, junto con su longitud y latitud, que posteriormente, permiten la integración con la información de imágenes satelitales:

Tabla 11. Estaciones meteorológicas finales – Valle del Cauca

Code	LatitudeDD	LongitudeDD
26055050	3.23333333	-76.5833333
26065040	3.31666667	-76.35
26070110	3.52719444	-76.2106944
26070170	3.45980556	-76.2502778
26075010	3.51338889	-76.3150278
26075050	3.57397222	-76.2778333
26075060	3.43333333	-76.4333333
26075080	3.36072222	-76.2994167

Code	LatitudeDD	LongitudeDD
26110060	4.46666667	-76.1166667
26110070	4.61666667	-76.0666667
26110100	4.48333333	-76.1
26110110	4.58333333	-76.05
26110120	4.97083333	-76.0381111
26110150	4.79383333	-75.9865
26110210	4.776	-76.1437778
26110230	4.41866667	-76.1005833

26075100	3.63333333	-76.4333333
26080070	3.69852778	-76.4297778
26080280	3.43816667	-76.6015556
26080310	3.47613889	-76.5229444
26080380	3.7	-76.5833333
26085120	3.45730556	-76.5032222
26095080	3.72994444	-76.0748333
26095110	3.93333333	-76.3
26095230	3.83519444	-76.2995278
26100400	4.30688889	-76.0188056
26100410	4.17188889	-76.0564444
26100690	3.87894444	-76.1004167
26100700	4.08727778	-76.1008056
26100790	4.05722222	-75.80575
26100800	3.88866667	-76.0658889
26100820	4.11666667	-76.2
26100830	4.68841667	-75.9615556
26105110	4.31822222	-76.0824167
26105150	4.23897222	-76.0331667
26105160	4.09038889	-76.2231111
26105230	4.02869444	-76.1680556
26110250	4.41666667	-76.1
26115030	4.41666667	-76.1
26115040	4.53127778	-76.0621667
26120130	4.40277778	-75.9154167
26120180	4.41022222	-75.8753611
26125130	4.18502778	-75.8323611
53080010	3.26066667	-77.25925
53090030	3.48247222	-77.2114444
53090040	3.69780556	-77.0671389
53100040	3.51772222	-76.7087778
53110020	3.78472222	-76.7618333
53110030	3.63333333	-76.7151667
53110040	3.67808333	-76.5399444
53110100	3.64786111	-76.5663889
53110130	3.69013889	-76.5925556
53115010	3.82019444	-76.9923333
53115020	3.88333333	-77.0666667
54030030	4.49858333	-76.3555278
54070030	4.18377778	-77.2163889
54070150	4.06955556	-77.0842222
54075020	3.95361111	-76.99025
54075040	4.22252778	-77.2763056

A continuación, se lleva a cabo una visualización geográfica de las estaciones meteorológicas restantes. Se crea un mapa centrado en las coordenadas geográficas (latitud y longitud) de las estaciones dentro del departamento del Valle del Cauca. Esta representación espacial es esencial para validar la ubicación de las estaciones y asegurar que las estaciones utilizadas en el análisis estén distribuidas de manera representativa a lo largo del departamento. Esta visualización permite además identificar posibles áreas sin cobertura de estaciones meteorológicas, lo que puede afectar la representatividad de las predicciones climáticas en esos sectores:

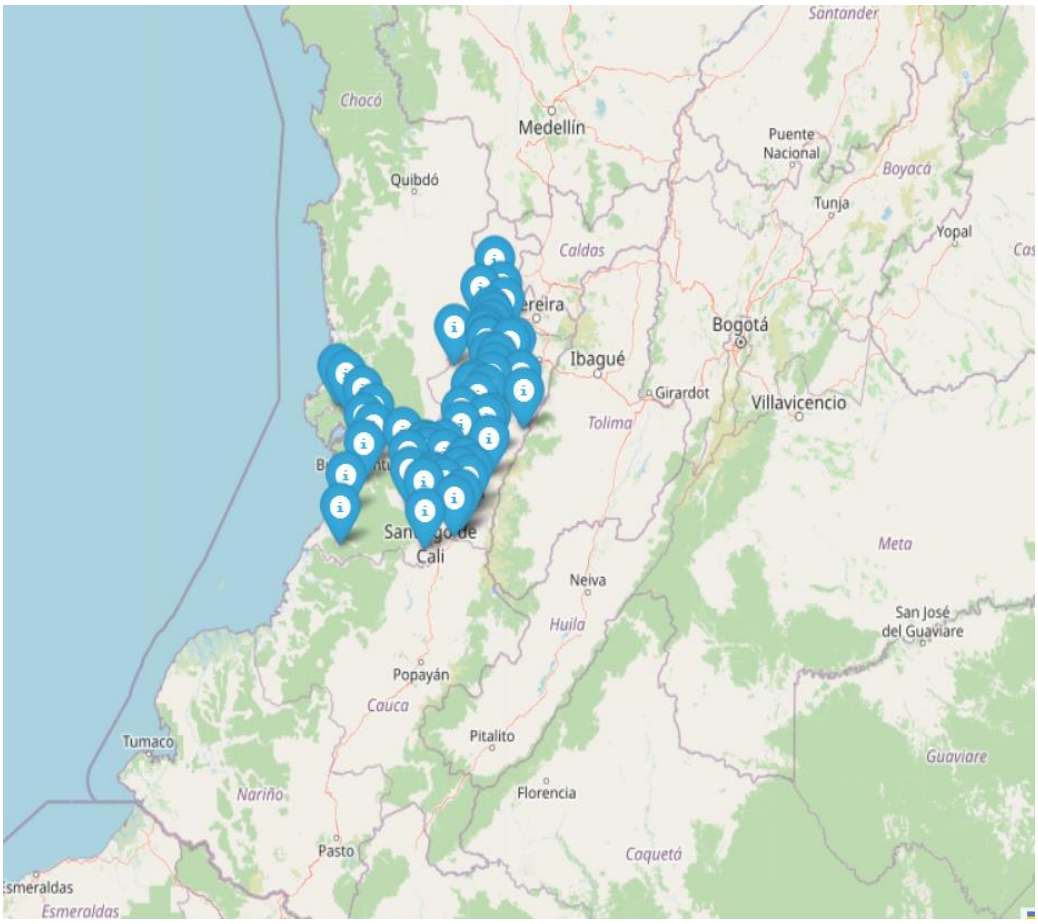

Figura 9. Visualización geográfica de estaciones meteorológicas

La Figura 10, denota la concentración significativa de estaciones en el suroeste de Colombia, particularmente alrededor del Valle del Cauca y áreas adyacentes.

Una vez verificada la ubicación geográfica de las estacione, una de las etapas más importantes del proceso corresponde a la integración de los datos satelitales por medio de herramientas geoespaciales propias de los sistemas de información geográfica (GIS). Esta actividad se centra en vincular los datos de las estaciones meteorológicas con la información geoespacial proporcionada por el grupo CHIRPS (Climate Hazards Group InfraRed Precipitation with Stations).

A través del software R Studio, se ejecuta un código para descargar y procesar los archivos NetCDF de CHIRPS, que contienen las estimaciones de precipitación diaria basadas en imágenes satelitales. Se utiliza un formato espacial de 0.05 grados de resolución, adecuado para el análisis a nivel regional. El código implementado, carga los archivos NetCDF y extrae las variables relevantes, como precipitación, latitud, longitud y tiempo. Posteriormente, estos datos son vinculados con las estaciones meteorológicas por su ubicación geográfica, lo que permite enriquecer la base de datos final con estimaciones satelitales de precipitación.

Como resultado de esta integración, se genera un conjunto de datos que combina observaciones de

estaciones meteorológicas y estimaciones satelitales, ofreciendo una representación más completa y detallada de la precipitación a lo largo del departamento del Valle del Cauca. Esta información es clave para mejorar la precisión de los modelos predictivos, ya que permite incorporar tanto observaciones terrestres como datos satelitales para la estimación de precipitaciones.

Finalmente, se realiza un análisis temporal de la información de datos satelitales. Este análisis se implementa utilizando Jupyter Notebook, donde se estructura la base de datos para asegurar la coherencia temporal de las observaciones. Para esto, se unifica el formato de fecha en todas las fuentes de datos al estándar YYYY-MM-DD, lo que facilita la manipulación y análisis de series temporales.

Este análisis temporal permite detectar patrones en la precipitación a lo largo del tiempo y facilita la identificación de tendencias y estacionalidades, lo que es esencial para el desarrollo de modelos predictivos precisos. A continuación, se presenta la estructura de la base de datos que entra al proceso de análisis temporal:

Tabla 12. Estructura table df – Mediciones Satelitales

Fecha	26055070_pre	26060020_pre	26070110_pre	26070760_pre	26075010_pre	26075040_pre	26080070_pre	26080290_pre	26090460_pre	26090630_pre
1983-01-01	2.603280	3.926486	17.761583	5.874582	2.253405	4.280465	6.928794	0.000000	26.951990	7.741584
1983-01-02	2.339656	3.528867	0.000000	5.279688	2.025212	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-03	1.812410	2.733629	0.000000	4.089900	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-04	2.174892	3.280356	0.000000	4.907879	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-05	1.788872	0.000000	0.000000	4.036784	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-06	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-07	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-08	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-09	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1983-01-11	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	8.391286	0.000000	0.000000
1983-01-12	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	9.427739	0.000000	6.411820
1983-01-13	8.486757	16.605419	0.000000	0.000000	0.000000	0.000000	0.000000	10.364534	0.000000	7.048935

6.1 Análisis de Tendencia:

Teniendo en cuenta que la información brindada es diaria, y por estación meteorológica, se procede a crear una columna con el valor promedio día, para que, a partir de este resultado, se obtenga un primer resultado que permita determinar si existe alguna tendencia estadísticamente significativa. Para ello, se utiliza la

prueba de Mann – Kendall, y se obtiene el siguiente resultado:

```
Trend Test Results: Mann_Kendall_Test(trend='no trend', h=False, p=0.9993784454459074, z=0.0007790031885462679, Tau=4.817066553
469995e-06, s=329.0, var_s=177283867265.66666, slope=0.0, intercept=4.317384045672413)
```

Figura 10. Resultado de Prueba Mann – Kendall

6.1.1 Trend: 'no trend'

La prueba de Mann-Kendall determina que no hay una tendencia estadísticamente significativa en los datos de precipitación. Esto significa que, basado en los datos proporcionados y el período de tiempo analizado, no hay evidencia suficiente para afirmar que las precipitaciones están aumentando o disminuyendo de manera significativa.

6.1.2 H (hypothesis test result): False

El resultado de la hipótesis de la prueba indica que no se rechaza la hipótesis nula, lo que significa que no hay evidencia suficiente para afirmar la presencia de una tendencia.

6.1.3 P-value (p): 0.9993784454459074

El valor p es alto, lo que confirma aún más que no hay suficiente evidencia estadística para rechazar la hipótesis nula de no tendencia. Un valor p alto como este sugiere que las diferencias en los datos pueden ser atribuidas al azar.

6.1.4 Z-score (z): 0.0007790031885462679

El valor z es muy cercano a cero, lo que indica que el resultado de la prueba está muy cerca de la media esperada bajo la hipótesis nula. En términos de estadística, muestra que no hay desviación significativa que indique una tendencia.

6.1.5 Var_s (variance of S): 177283867265.66666

La varianza de S es bastante alta, indicando que hay una gran variabilidad en los datos a lo largo del tiempo, pero esta variabilidad no se traduce necesariamente en una tendencia.

6.1.6 Slope: 0.0

La pendiente de la tendencia estimada es cero, reflejando la conclusión de que no hay tendencia en los datos.

Basado en todos estos resultados, se puede concluir que no hay evidencia de cambios significativos en las precipitaciones a lo largo del tiempo en el conjunto de datos analizado. Esto podría ser útil para demostrar que, en el contexto y la región estudiada, las condiciones de precipitación han permanecido estables a lo largo del período observado.

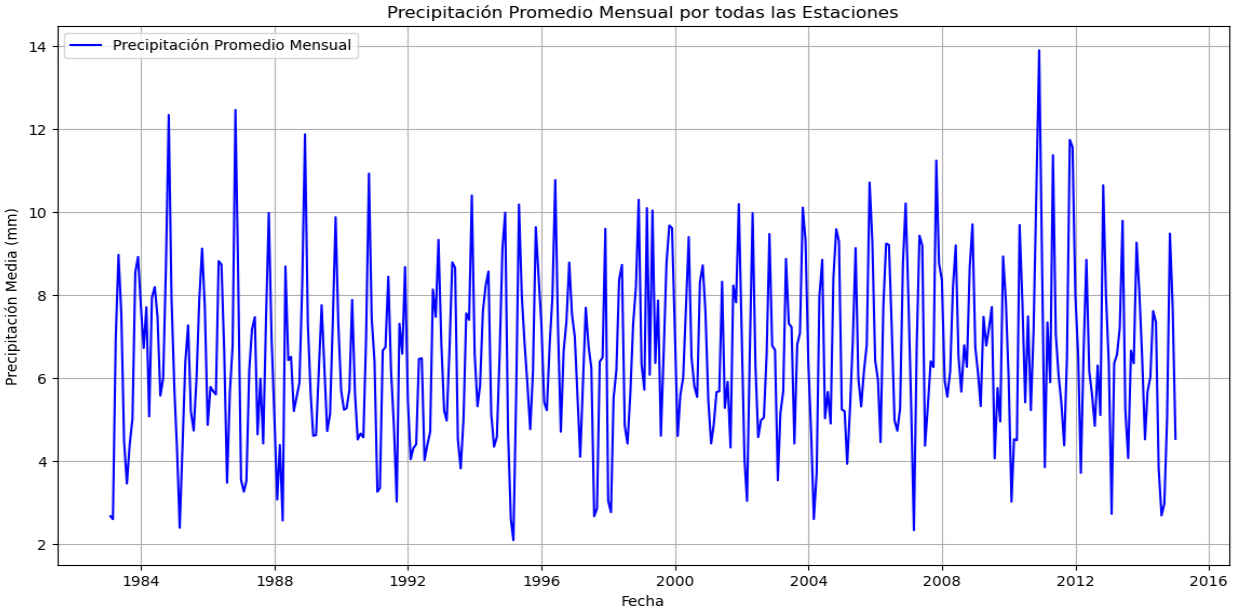

Figura 11. Precipitación Promedio Mensual por todas las estaciones

6.2 Análisis de la información acumulada mensual y anual

Se procede con el análisis de la información de la precipitación acumulada mensual y anual, para el promedio diario de todas las estaciones, para evaluar su comportamiento:

- El gráfico muestra una considerable variabilidad en la precipitación mensual a lo largo del tiempo, con valores que fluctúan entre cerca de 2 mm y más de 12 mm. Esta variabilidad podría ser atribuible a la estacionalidad inherente, con meses más lluviosos intercalados con periodos más secos.

- A simple vista, no parece haber una tendencia lineal clara que indique un aumento o disminución sistemática en la precipitación promedio a lo largo de los años. Esto es consistente con la prueba de Mann-Kendall previamente mencionada que también indicó la ausencia de una tendencia.

- Los picos pronunciados pueden estar relacionados con eventos específicos de precipitación más intensa, posiblemente vinculados a fenómenos meteorológicos como El Niño o La Niña, que son conocidos por afectar patrones de precipitación globalmente.

Existe un grado significativo de "ruido" o variabilidad mensual en los datos. Esta variabilidad puede complicar la detección de patrones subyacentes a menos que se apliquen métodos de suavizado o se realice un análisis de frecuencia para identificar ciclos estacionales más claros.

Aunque hay variabilidad mensual y estacional, no hay evidencia de cambios significativos o tendencias a largo plazo en la precipitación total. Esto puede indicar que, a pesar de la variabilidad mensual y anual, las condiciones climáticas en términos de precipitación han permanecido relativamente estables durante el período observado.

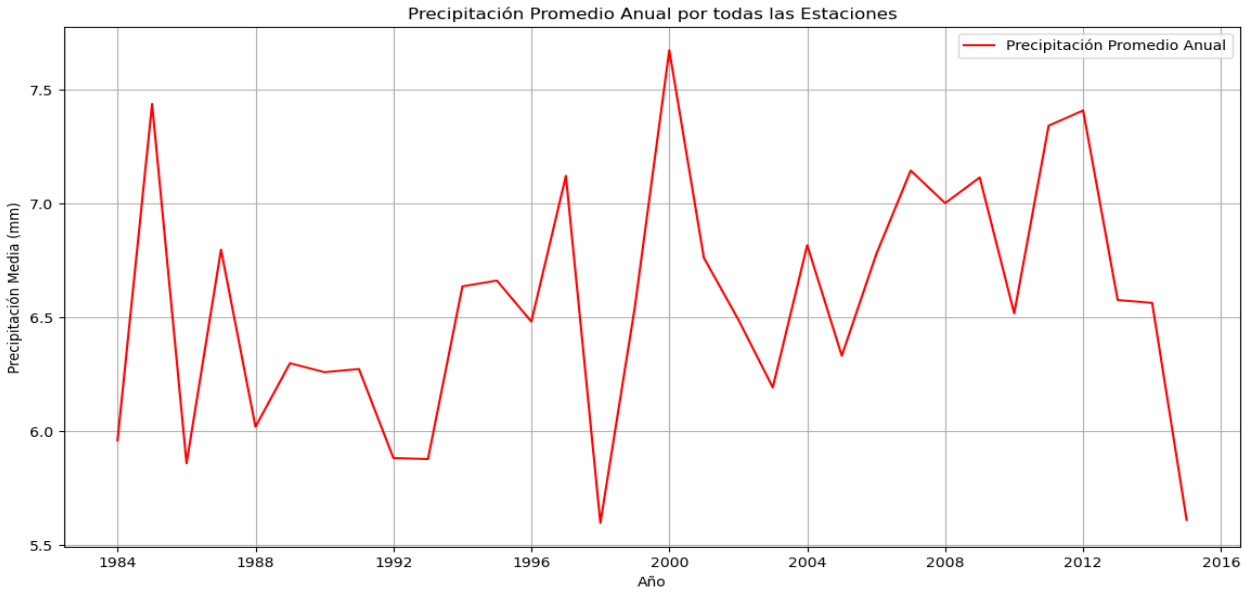

Figura 12. Precipitación Promedio Anual por todas las estaciones

- El gráfico presenta una serie de fluctuaciones significativas en la precipitación anual a lo largo de los años. Estas fluctuaciones pueden reflejar variaciones climáticas naturales, como ciclos de El Niño y La Niña, que son conocidos por influir en los patrones de precipitación a nivel mundial.

- No se observa una tendencia clara y consistente a lo largo del tiempo que indique un aumento o una disminución en los niveles de precipitación. Sin embargo, hay períodos notables de alta y baja precipitación, como los picos altos en los años cerca de 1987, 1998 y 2011, y una notable caída hacia el final del período alrededor de 2015-2016.

- La variabilidad en la precipitación anual es bastante pronunciada, con picos que a veces superan los 7 mm y caídas por debajo de los 6 mm. Esto sugiere una heterogeneidad en las condiciones anuales, que podrían estar relacionadas con eventos meteorológicos específicos o cambios en los patrones climáticos a largo plazo.

- Los picos en la precipitación pueden correlacionarse con años de fuertes fenómenos de La Niña, que generalmente aumentan la precipitación en muchas partes del mundo.

- A pesar de la variabilidad anual, no hay una tendencia clara de largo plazo hacia un aumento o disminución en la precipitación. Esto podría indicar que, a nivel anual, las condiciones de precipitación han permanecido relativamente estables a lo largo del período observado, aunque con variabilidad significativa de un año a otro.

6.3 Implementar análisis de descomposición de serie temporal por medio de STL: Se implementa la técnica STL (Descomposición de Series Temporales basada en Loess) para una mejor visualización y comprensión de los componentes estacionales.

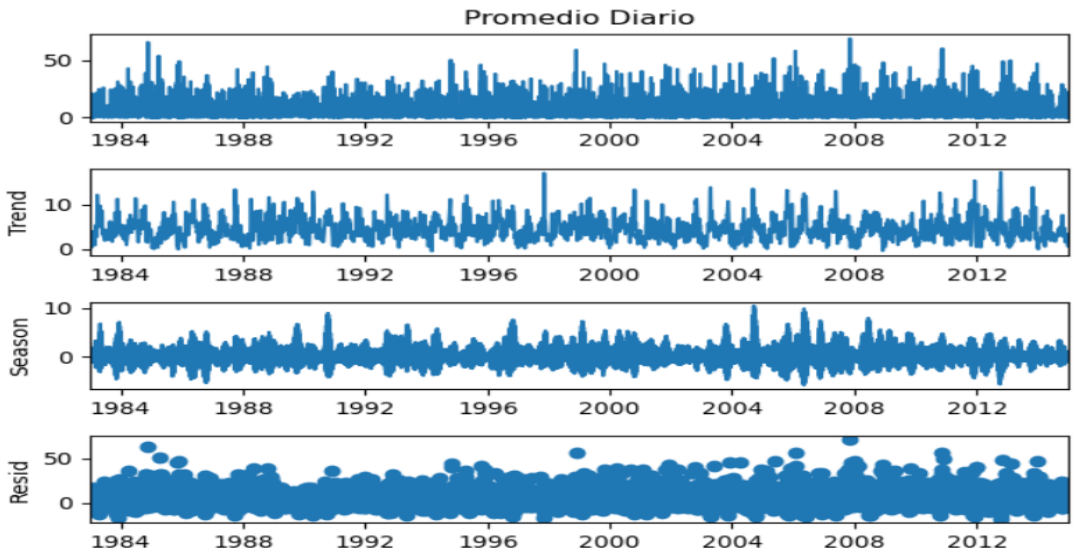

Figura 13. Descomposición de serie temporal - STL

La Figura 15, muestra la descomposición de la serie temporal que incluye el promedio diario de precipitación junto con sus componentes de tendencia, estacionalidad y residuos. A continuación, se presenta el análisis de cada uno:

6.3.1 Promedio Diario:

La serie de promedio diario muestra picos considerables, lo que indica días de alta precipitación intercalados con períodos de baja precipitación. Este comportamiento, es común en series de precipitación.

La gran variabilidad en los datos diarios puede necesitar suavizado o modelado para predecir tendencias de manera efectiva.

6.3.2 Tendencia:

La tendencia muestra variaciones, pero sigue un patrón ascendente o descendente claro a lo largo del tiempo. Los picos y valles no son consistentes y no muestran un aumento o disminución significativos a largo plazo.

La ausencia de una tendencia clara en la precipitación sugiere que no ha habido un cambio significativo en

los niveles de precipitación a lo largo de los años cubiertos por la serie. Esto podría indicar que no hay influencias externas consistentes afectando los niveles de precipitación durante el período observado, o que estas influencias se han balanceado a lo largo del periodo de estudio.

6.3.3 Estacionalidad:

El componente estacional muestra oscilaciones que sugieren una estacionalidad en los datos, probablemente reflejando los patrones climáticos típicos de las estaciones húmedas y secas.

6.3.4 Residuos:

Los residuos muestran variaciones aleatorias alrededor de cero, sin patrones consistentes. Esto indica que la mayor parte de la información en los datos ha sido capturada por los componentes de tendencia y estacionalidad.

Los residuos bajos, indican que las anomalías o eventos extremos no explicados por la tendencia o la estacionalidad son mínimos, lo que es positivo para el modelado predictivo.

7. EJECUCIÓN DEL MODELO DE PREDICCIÓN

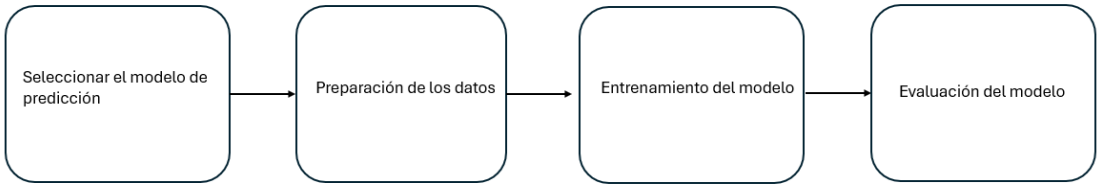

Figura 14. Flujo metodológico objetivo específico 3

La Figura 16, presenta el flujo de trabajo para la construcción y evaluación del modelo predictivo de la precipitación acumulada semanal del departamento del Valle del Cauca. Los pasos descritos son los siguientes:

Seleccionar el modelo de predicción: En esta fase inicial, se elige el modelo predictivo que se utilizará para el análisis de los datos. La elección del modelo depende de la naturaleza de los datos y del tipo de predicción a realizar.

Preparación de los datos: Una vez seleccionado el modelo, se procede a la preparación de los datos. Este paso incluye la limpieza, normalización, transformación y cualquier otro preprocesamiento necesario para asegurar que los datos estén en el formato adecuado para ser utilizados por el modelo. También implica la segmentación de los datos en conjuntos de entrenamiento y prueba.

Entrenamiento del modelo: En esta etapa, el modelo es entrenado utilizando el conjunto de datos preparado. El entrenamiento implica alimentar al modelo con los datos históricos para que aprenda los patrones y las relaciones presentes en los mismos. Durante este proceso, se ajustan los parámetros del modelo para optimizar su capacidad de predicción.

Evaluación del modelo: Finalmente, se evalúa el rendimiento del modelo utilizando el conjunto de datos de prueba. Esto implica medir su precisión, error y otros indicadores relevantes para determinar qué tan bien el modelo es capaz de predecir la variable objetivo, en este caso la precipitación.

Este flujo de trabajo asegura una secuencia ordenada y lógica en la construcción del modelo predictivo, enfocándose en la correcta elección del modelo, la preparación detallada de los datos y la evaluación adecuada de los resultados obtenidos.

A continuación, se presenta el desarrollo y resultados obtenidos del tercer objetivo específico por medio de las siguientes actividades:

La selección del modelo de predicción constituye un paso crucial en el desarrollo del proyecto, dado que permite establecer una arquitectura adecuada para abordar el problema de la predicción de precipitaciones semanales en el departamento del Valle del Cauca. En este sentido, se evalúan diferentes enfoques, entre ellos, las Redes Neuronales Convolucionales (CNN) y las Redes Neuronales Recurrentes (RNN). Las CNN, ampliamente reconocidas por su capacidad para extraer características espaciales en datos complejos, han demostrado ser muy eficaces en tareas de clasificación y reconocimiento. Sin embargo, dado que este proyecto se centra en el análisis de datos de series temporales, se opta por utilizar una arquitectura basada en Redes Neuronales Recurrentes, específicamente Long Short-Term Memory (LSTM), debido a su capacidad para capturar dependencias a largo plazo en secuencias temporales y minimizar el problema del sobreajuste.

7.1 Aplicación del Modelo a las Estaciones seleccionadas

Dado que el proyecto busca la creación de un modelo de predicción para el departamento del Valle del Cauca, se establece que la metodología de selección de las estaciones a las cuales se realizará la implementación y evaluación del modelo es aquella estación por municipio, que cuente con el mayor nivel de completitud de información. Por lo tanto, teniendo en cuenta la información presentada en la Tabla 11, se procede a determinar las estaciones finales que serán implementadas para la ejecución del modelo (Tabla 12).

Tabla 13. Datos faltantes por Estaciones meteorológicas Disponibles – Valle del Cauca

Número de estación	Station_code	missings	Municipio
1	26070760	3	FLORIDA
2	26080280	20	CALI
3	26090060	34	GINEBRA
4	26060020	57	CANDELARIA
5	26100070	129	BUGALAGRANDE
6	26080290	156	YOTOCO
7	26100300	166	OBANDO
8	26110300	182	TORO
9	26110290	189	LA UNIÓN
10	26090630	195	GUACARI
11	26075040	202	PALMIRA
12	54030010	206	VERSALLES
13	26110160	248	TORO
14	26100780	278	ZARZAL

Número de estación	Station_code	missings	Municipio
30	26070110	468	PALMIRA
31	53115010	501	BUENAVENTURA
32	26110120	504	EL AGUILA
33	26115040	513	LA UNIÓN
34	53110030	534	DAGUA
35	26110150	563	ANSERMANUEVO
36	26125130	569	SEVILLA
37	53090040	655	BUENAVENTURA
38	26100700	661	ANDALUCIA
39	53110100	698	LA CUMBRE
40	26070170	766	PRADERA
41	53090030	796	BUENAVENTURA
42	54070030	829	BUENAVENTURA
43	53110040	840	LA CUMBRE

15	26055070	279	CALI
16	26120180	283	SEVILLA
17	26100830	309	CARTAGO
18	26090460	314	EL CERRITO
19	26120150	340	ALCALÁ
20	26110040	340	BOLÍVAR
21	26100350	345	TULUA
22	26080070	354	VIJES
23	26100770	379	BUGALAGRANDE
24	26120120	385	CAICEDONIA
25	26100690	397	GUADALAJARA DE BU
26	26100410	403	BUGALAGRANDE
27	53100040	404	DAGUA
28	26120130	436	ZARZAL
29	26110230	437	ROLDANILLO
44	53080010	882	BUENAVENTURA
45	26100400	918	ZARZAL
46	54070150	1001	BUENAVENTURA
47	26100800	1002	GUADALAJARA DE BU
48	26105230	1142	TULUA
49	54030030	1256	EL DOVIO
50	26105160	1302	TULUA
51	26095230	1319	GUADALAJARA DE BU
52	26105150	1360	BUGALAGRANDE
53	26110210	1533	ANSERMANUEVO
54	26095080	1560	EL CERRITO
55	26075010	1561	PALMIRA
56	54075040	1881	BUENAVENTURA
57	26100790	2749	SEVILLA
58	54075020	3185	BUENAVENTURA

Tabla 14. Información Selección Final Estaciones Meteorológicas

Número de estación	Station_code	missing_count	Municipio
1	26120150	340	ALCALÁ
2	26100700	661	ANDALUCÍA
3	26110150	563	ANSERMANUEVO
4	26110040	340	BOLÍVAR
5	53115010	501	BUENAVENTURA
6	26100070	129	BUGALAGRANDE
7	26120120	385	CAICEDONIA
8	26080280	20	CALI
9	26060020	57	CANDELARIA
10	26100830	309	CARTAGO
11	53100040	404	DAGUA
12	26110120	504	EL ÁGUILA
13	26090460	314	EL CERRITO
14	26070760	3	FLORIDA
15	26090060	34	GINEBRA

16	26090630	195	GUACARI
17	26100690	397	GUADALAJARA DE BU
18	53110100	698	LA CUMBRE
19	26110290	189	LA UNIÓN
20	26100300	166	OBANDO
21	26075040	202	PALMIRA
22	26070170	766	PRADERA
23	26110230	437	ROLDANILLO
24	26120180	283	SEVILLA
25	26110300	182	TORO
26	26100350	345	TULUÁ
27	54030010	206	VERSALLES
28	26080070	354	VIJES
29	26080290	156	YOTOCO
30	26100780	278	ZARZAL

7.2 Preparación de los Datos

Una vez seleccionado el modelo y las estaciones por municipio, se procede con la preparación de los datos, ya que es una etapa fundamental en la implementación de cualquier modelo de machine learning. En este caso, se trabaja con una base de datos que incluye información diaria de precipitación proveniente de estaciones meteorológicas en el departamento del Valle del Cauca, tanto medidas terrestres (observadas) como mediciones satelitales. A partir de esta base de datos, se lleva a cabo un proceso de estructuración y transformación para garantizar que los datos estén en un formato adecuado para el entrenamiento del modelo LSTM.

Inicialmente, se realiza la conversión de la columna de fecha al formato datetime, lo que facilita la manipulación temporal de los datos. Esta conversión permite generar nuevas variables, como el número del mes y el día de la semana, lo que resulta útil para capturar posibles patrones estacionales. Después, se procede con la acumulación de los valores diarios en periodicidad semanal, lo que asegura, tener el periodo de tiempo establecido en el alcance del proyecto. Posteriormente, se procede al escalado de las variables meteorológicas utilizando la técnica de normalización MinMaxScaler, que transforma los datos a un rango entre 0 y 1. Esto es esencial para mejorar el rendimiento y estabilidad del modelo LSTM, ya que este tipo de redes neuronales es particularmente sensible a la magnitud de los datos de entrada.

Adicionalmente, se procede con la creación de la variable de rezago, ya que es un paso clave en la preparación de los datos, al permitir capturar la autocorrelación temporal inherente a los datos de precipitación. Para ello, se implementa un rezago de un día (lag1), lo que permite que el modelo LSTM capture la dependencia temporal entre los registros diarios de precipitación.

En cuanto al manejo de valores faltantes, se opta por reemplazar los valores NaN resultantes del rezago con ceros, ya que la ausencia de precipitación puede ser interpretada como días sin lluvia.

Por último, el conjunto de datos final se divide en train y test, utilizando el 80% de los datos para entrenamiento y el 20% para prueba. Esta división se realiza manteniendo la secuencia temporal, asegurando así que el modelo preserve la estructura de la serie temporal en su proceso de aprendizaje.

7.3 Entrenamiento del Modelo LSTM

Para este paso, se diseña el modelo con las siguientes capas:

- Una capa LSTM con un número ajustable de neuronas, que permite al modelo retener la información necesaria para predecir la precipitación futura basándose en los valores pasados.
- Una capa de Dropout con una tasa de 20%, destinada a prevenir el sobreajuste al eliminar aleatoriamente conexiones durante el entrenamiento.
- Una capa Dense de salida, con una sola neurona, que genera la predicción de la cantidad de precipitación diaria en milímetros.

Para optimizar el modelo, se utiliza el optimizador Adam, conocido por su capacidad de ajustar eficientemente los pesos de la red, y la función de pérdida mean_squared_error (MSE), ya que se trata de un problema de regresión en el que se predicen valores continuos de precipitación.

Durante el entrenamiento, se implementa un proceso de tuning de hiperparámetros mediante GridSearchCV. Este proceso permite encontrar la combinación óptima de parámetros como el número de neuronas (50, 75, 100), la función de activación (ReLU, Leaky ReLU), el tamaño del batch (64), y el número de épocas (50). Adicionalmente, se implementa EarlyStopping para evitar el sobreajuste. Este mecanismo interrumpe el entrenamiento si el error en el conjunto de validación no mejora tras 5 épocas consecutivas.

7.4 Evaluación del Modelo

El rendimiento del modelo se evalúa utilizando varias métricas clave, tanto en los conjuntos de entrenamiento como de prueba:

- **Mean Absolute Error (MAE)**: Esta métrica mide el error absoluto promedio entre las predicciones y los valores reales, lo que proporciona una idea clara de la desviación promedio en milímetros entre las predicciones de precipitación y los valores reales.
- **Mean Squared Error (MSE):** El MSE penaliza más los errores grandes, lo que lo convierte en una medida útil para identificar predicciones que se desvían considerablemente de los valores observados.
- **Root Mean Squared Error (RMSE):** Al estar en la misma escala que los valores originales de precipitación, el RMSE ofrece una interpretación más directa de la magnitud promedio del error en las predicciones.

A continuación, se presenta los resultados de cada una de las métricas relacionadas anteriormente, para la implementación del modelo en todas las estaciones, tanto para TEST como para TRAIN:

Tabla 15. Métricas de Evaluación Modelo LSTM - TEST

				TEST		
Número de estación	Station_code	Municipio	R2 Modelo	MAE	MSE	RMSE
1	26120150	ALCALA	0.3434	24.67	1,181.47	4.97
2	26100700	ANDALUCIA	0.3014	22.69	1,100.05	4.76
3	26110150	ANSERMANUEVO	0.3205	18.21	584.28	4.27
4	26110040	BOLIVAR	0.3869	13.56	371.77	3.68
5	53115010	BUENAVENTURA	0.1628	64.58	11,052.59	8.04
6	26100070	BUGALAGRANDE	0.3646	15.60	469.78	3.95
7	26120120	CAICEDONIA	0.4534	19.87	815.32	4.46
8	26080280	CALI	0.4063	18.61	624.30	4.31
9	26060020	CANDELARIA	0.4159	15.50	511.84	3.94
10	26100830	CARTAGO	0.3528	15.92	448.48	3.99
11	53100040	DAGUA	0.2691	14.91	375.93	3.86
12	26110120	EL AGUILA	0.3119	31.27	1,816.44	5.59
13	26090460	EL CERRITO	0.4199	20.17	1,021.27	4.49
14	26070760	FLORIDA	0.3995	22.05	1,113.50	4.27
15	26090060	GINEBRA	0.4916	16.63	544.00	4.08
16	26090630	GUACARI	0.3976	12.42	290.70	3.52
17	26100690	GUADALAJARA DE BU	0.4608	15.22	422.40	3.90
18	53110100	LA CUMBRE	0.2910	16.49	460.63	4.06
19	26110290	LA UNION	0.3007	12.79	353.58	3.58
20	26100300	OBANDO	0.2775	16.55	505.41	4.07
21	26075040	PALMIRA	0.4032	11.08	279.23	3.33
22	26070170	PRADERA	0.2087	23.60	2,223.99	4.25
23	26110230	ROLDANILLO	0.3362	12.13	305.76	3.64
24	26120180	SEVILLA	0.3419	20.75	808.48	4.55
25	26110300	TORO	0.2285	13.94	410.26	3.73
26	26100350	TULUA	0.3607	23.70	1,242.82	4.87
27	54030010	VERSALLES	0.3118	14.36	404.88	3.79
28	26080070	VIJES	0.3632	11.20	232.46	3.35
29	26080290	YOTOCO	0.3058	14.50	427.85	3.81
30	26100780	ZARZAL	0.2852	15.60	493.89	3.95

Tabla 16. Métricas de Evaluación Modelo LSTM - TRAIN

				TRAIN		
Número de estación	Station_code	Municipio	R2 Modelo	MAE	MSE	RMSE
1	26120150	ALCALA	0.3434	23.85	1,000.10	4.88
2	26100700	ANDALUCIA	0.3014	20.27	766.26	4.50
3	26110150	ANSERMANUEVO	0.3205	18.70	702.20	4.32
4	26110040	BOLIVAR	0.3869	13.80	370.48	3.72
5	53115010	BUENAVENTURA	0.1628	50.47	4,331.32	7.10
6	26100070	BUGALAGRANDE	0.3646	14.25	369.68	3.78
7	26120120	CAICEDONIA	0.4534	21.29	1,007.11	4.61
8	26080280	CALI	0.4063	19.87	741.57	4.46
9	26060020	CANDELARIA	0.4159	15.79	513.19	3.97
10	26100830	CARTAGO	0.3528	17.73	598.10	4.21
11	53100040	DAGUA	0.2691	14.26	420.59	3.78
12	26110120	EL AGUILA	0.3119	26.65	2,431.53	5.16
13	26090460	EL CERRITO	0.4199	20.25	808.06	4.50
14	26070760	FLORIDA	0.3995	18.30	4.70	4.28
15	26090060	GINEBRA	0.4916	16.72	590.25	4.09
16	26090630	GUACARI	0.3976	12.51	350.14	3.54
17	26100690	GUADALAJARA DE BU	0.4608	16.34	546.11	4.04
18	53110100	LA CUMBRE	0.2910	17.26	626.72	4.15
19	26110290	LA UNION	0.3007	14.30	477.16	3.78
20	26100300	OBANDO	0.2775	16.31	499.86	4.04
21	26075040	PALMIRA	0.4032	11.22	270.54	3.35
22	26070170	PRADERA	0.2087	18.14	4.86	4.26
23	26110230	ROLDANILLO	0.3362	13.03	3.48	3.61
24	26120180	SEVILLA	0.3419	19.99	727.97	4.47
25	26110300	TORO	0.2285	13.05	349.43	3.61
26	26100350	TULUA	0.3607	22.49	1,040.89	4.74
27	54030010	VERSALLES	0.3118	17.46	685.39	4.18
28	26080070	VIJES	0.3632	12.28	289.04	3.50
29	26080290	YOTOCO	0.3058	15.16	682.89	3.89
30	26100780	ZARZAL	0.2852	15.00	416.35	3.87

Los resultados obtenidos en el modelo de predicción de precipitaciones, mediante redes neuronales LSTM, evaluados a través de diversas métricas (MAE, MSE, RMSE y R^2), muestran un comportamiento variado entre las diferentes estaciones meteorológicas del departamento del Valle del Cauca. A continuación, se presentan un análisis del rendimiento general del modelo, y así, identificar las posibles causas que generen los valores extremos (máximos y mínimos) y obtener una mayor comprensión de los factores que afectan el rendimiento del modelo predictivo.

7.4.1 Variabilidad en las métricas de error

El rendimiento del modelo varía significativamente entre las estaciones, lo que se puede observar en la amplitud de los valores de MAE, MSE y RMSE. Esta variabilidad sugiere que el modelo es capaz de predecir con precisión en algunas estaciones, mientras que en otras tiene dificultades para capturar los patrones de precipitación.

MAE: Oscila entre 11.08 mm y 64.58 mm en el conjunto de prueba. Las estaciones con un MAE más bajo indican una menor desviación promedio entre las predicciones y los valores reales, lo que refleja un mejor rendimiento del modelo en esas estaciones.

RMSE: Muestra valores que varían de 3.33 mm a 8.04 mm, lo que indica que el modelo tiene errores más grandes en algunas estaciones, posiblemente debido a eventos de precipitación más extremos.

R^2, que mide la proporción de la variabilidad explicada por el modelo, también fluctúa de manera considerable, con valores que van desde 0.1628 hasta 0.4916. Las estaciones con un R^2 bajo indican que el modelo no está capturando adecuadamente la variabilidad en los datos de precipitación.

7.5 Posibles factores que afectan el rendimiento del modelo

La variabilidad en los resultados puede estar influenciada por varios factores, tanto inherentes a los datos como al entorno geográfico de las estaciones meteorológicas:

Calidad y cantidad de los datos disponibles: El número de datos faltantes puede generar un impacto directo en el rendimiento del modelo. Estaciones como Buenaventura (con 501 datos faltantes) y La Cumbre (con 698 datos faltantes) muestran peores métricas de error, lo que sugiere que la falta de información dificulta el entrenamiento adecuado del modelo. A su vez, que el modelo no capture bien los patrones subyacentes en los datos, lo que se traduce en predicciones menos precisas.

Patrones de precipitación locales: Las estaciones ubicadas en regiones con patrones de precipitación más regulares, como Palmira y Cali, tienden a tener errores menores. Estas zonas presentan menos variabilidad en sus patrones de lluvia, lo que facilita al modelo aprender relaciones más claras entre las entradas y las salidas. Por el contrario, en estaciones ubicadas en áreas con alta variabilidad climática, como Buenaventura o Pradera, los errores son más grandes, lo que indica que el modelo tiene dificultades para predecir en regiones con patrones de precipitación más complejos y menos predecibles.

Factores geográficos: Las estaciones situadas en áreas montañosas o cerca de la costa, como Buenaventura y El Águila, muestran peores resultados en las métricas de error, lo que podría deberse a las complejas interacciones climáticas que afectan estas regiones. Los modelos LSTM pueden tener dificultades para capturar los efectos de factores topográficos como montañas o cuerpos de agua cercanos, que influencian de manera significativa los patrones de lluvia.

Dependencia temporal: Dado que el modelo LSTM es diseñado para capturar la dependencia temporal en los datos, es probable que el rendimiento del modelo sea más bajo en estaciones donde las precipitaciones no presentan una clara correlación temporal o estacional. En áreas con precipitaciones irregulares o altamente estacionales, el modelo puede tener dificultades para aprender patrones consistentes, lo que explicaría los altos valores de error en estaciones como Pradera y El Águila.

7.6 Relación entre el conjunto de entrenamiento y prueba

Una observación clave en el análisis de los resultados es la consistencia entre las métricas de error en los conjuntos de entrenamiento y prueba. En la mayoría de las estaciones, las métricas en el conjunto de prueba son similares a las del conjunto de entrenamiento, lo que sugiere que el modelo no está sobreajustando los datos. Sin embargo, en algunas estaciones como Buenaventura y Pradera, los errores son significativamente mayores en el conjunto de prueba, lo que indica que el modelo podría estar sobreajustando los datos de entrenamiento o que los patrones de precipitación en esas estaciones son especialmente difíciles de predecir.

7.7 Casos extremos y posibles causas

Buenaventura tiene los valores de error más altos en casi todas las métricas. Esto puede explicarse por las condiciones climáticas extremas en la región costera, donde las precipitaciones son mucho más volátiles y difíciles de predecir debido a la influencia del océano y fenómenos climáticos como el fenómeno de La Niña.

Pradera también muestra un alto nivel de error, tanto en MAE como en MSE. Dado que esta estación tiene la mayor cantidad de datos faltantes (766), esto sugiere que la falta de datos ha impedido que el modelo capture correctamente los patrones de precipitación.

7.8 Estaciones con buen rendimiento

Estaciones como Palmira, Cali y Florida muestran resultados relativamente buenos, con valores bajos en las métricas de error. Esto puede explicarse por la menor variabilidad en los patrones de precipitación y una cantidad más completa de datos, lo que facilita al modelo la identificación de relaciones claras en los datos.

A continuación, se presenta el gráfico correspondiente a la comparación entre las predicciones y los valores reales de las primeras 35 semanas del periodo de estudio para la estación Florida Código 26070760.

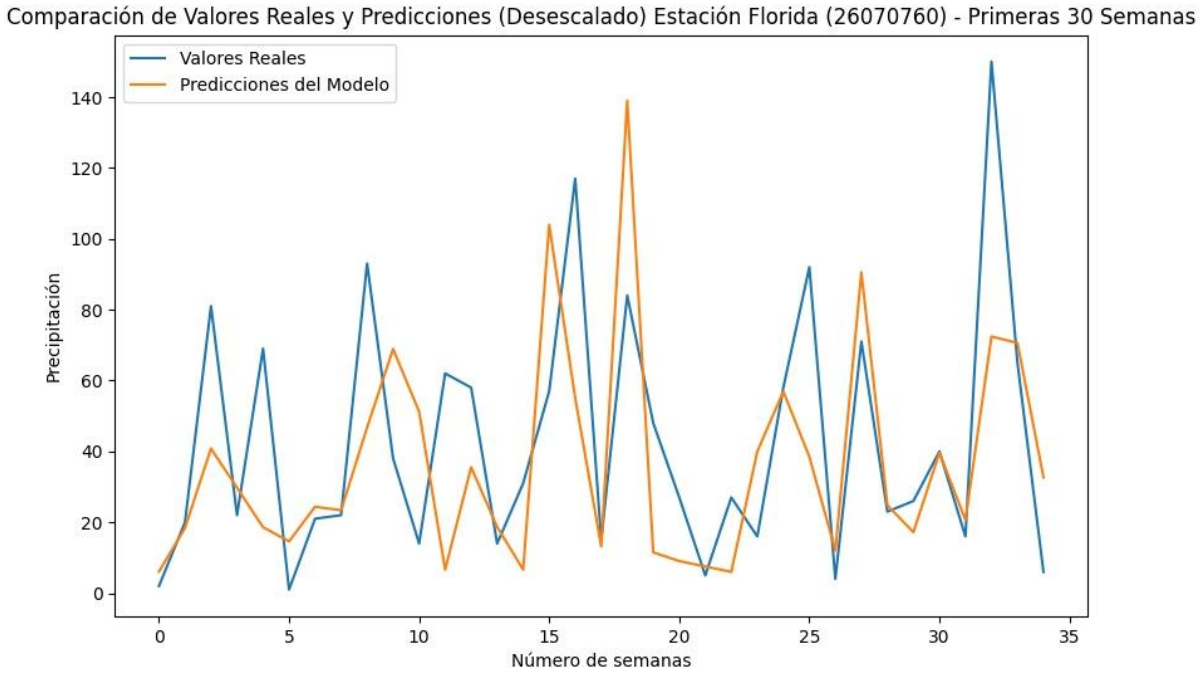

Figura 15. Comparación de valores reales y predicciones Estación Florida (26070760)

8. COMPARACIÓN DEL RENDIMIENTO DEL MODELO LSTM CON MODELOS DE SERIES TEMPORALES

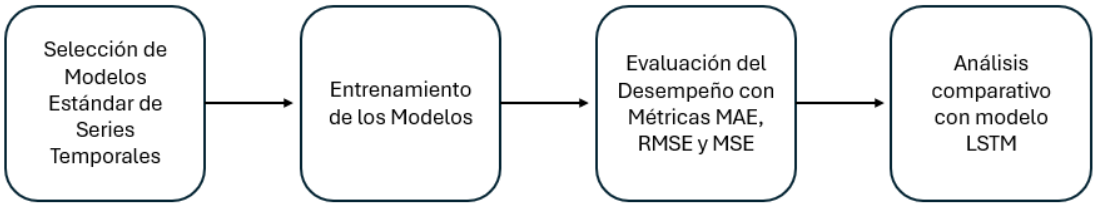

Figura 16. Flujo metodológico objetivo específico 4

La Figura 17, presenta el flujo de trabajo para la selección y entrenamiento de modelos estándar de Series temporales para su comparación con el modelo LSTM desarrollado. Los pasos descritos son los siguientes:

Selección de Modelos Estándar de Series Temporales: Identificar los modelos tradicionales como ARIMA, SARIMA para compararlos con el modelo LSTM.

Entrenamiento de los Modelos: Entrenar cada uno de los modelos estándar utilizando los mismos datos y particiones temporales que el LSTM.

Evaluación del Desempeño con Métricas MAE y RMSE: Calcular las métricas de error (MAE, RMSE y MSE) para medir el rendimiento predictivo de los modelos estándar.

Análisis Comparativo con el Modelo LSTM: Comparar los resultados obtenidos de los modelos estándar con los del LSTM, destacando cuál ofrece un mejor rendimiento en términos de precisión y error.

Este flujo de trabajo asegura un proceso de evaluación del modelo LSTM en comparación con los métodos tradicionales de series temporales en términos de precisión (MAE) y capacidad de manejar errores grandes (RMSE). A su vez, a partir del análisis comparativo, establecer conclusiones acerca de si el uso de LSTM permite la obtención de mejores resultados en comparación con los modelos tradicionales.

En esta sección, se aborda la predicción de series temporales con el objetivo de modelar el comportamiento de la precipitación a través de métodos avanzados de modelado estadístico. Inicialmente, se realiza la selección de un modelo de series temporales adecuado para capturar las dinámicas complejas que presentan los datos de precipitación. Para este propósito, se considera el uso de modelos ARIMA (Autoregressive Integrated Moving Average) y SARIMA (Seasonal Autoregressive Integrated Moving Average), los cuales son ampliamente utilizados en series temporales debido a su capacidad para modelar dependencias tanto a corto como a largo plazo, así como para incorporar estacionalidades en los datos.

La primera etapa del proceso consiste en la obtención y preparación de los datos, específicamente aquellos correspondientes a la precipitación diaria observada en diferentes estaciones meteorológicas. Estos datos son agrupados por semana para obtener una representación más adecuada del comportamiento a lo largo del tiempo, lo que también permite manejar de mejor manera cualquier patrón estacional que pueda estar presente en la serie temporal. Una vez organizados los datos, se procede a separar el conjunto en dos partes: un 80% de los registros es utilizado para el entrenamiento del modelo y el 20% restante se reserva para la validación del modelo en un contexto no visto, es decir, el conjunto de prueba.

Con los datos preparados, se comienza el proceso de ajuste del modelo ARIMA, en el cual es necesario identificar los hiperparámetros más apropiados para capturar las dependencias en los datos de precipitación. Este modelo se caracteriza por tres parámetros clave: p, que representa el número de retardos en la parte autorregresiva; d, que indica el grado de diferenciación aplicada a los datos para hacer la serie estacionaria; y q, que define el número de términos en el componente de media móvil. Inicialmente, se prueba con el modelo ARIMA (1, 1, 1), que es un punto de partida común en el análisis de series temporales, ya que contempla un término autorregresivo, una diferenciación para lograr la estacionariedad y un término de media móvil para capturar la dependencia de los errores pasados.

Teniendo en cuenta los valores obtenidos del modelo inicial, se procede con la búsqueda de mejores combinaciones de hiperparámetros utilizando una estrategia de búsqueda en cuadrícula (Grid Search), la cual permite probar múltiples configuraciones de forma sistemática. Se evalúan diferentes combinaciones para ARIMA y SARIMA.

Tabla 17. Métricas de Evaluación Modelo SARIMA

				TEST			TRAIN		
Número de estación	Station_code	missing_count	Municipio	MAE	MSE	RMSE	MAE	MSE	RMSE
1	26120150	340	ALCALA	23.64	881.22	29.69	21.12	715.09	26.74
2	26100700	661	ANDALUCIA	21.71	735.44	27.12	19.43	595.78	24.41
3	26110150	563	ANSERMANUEVO	21.39	717.44	26.79	18.91	594.33	24.38
4	26110040	340	BOLIVAR	18.43	543.35	23.31	16.54	447.69	21.16
5	53115010	501	BUENAVENTURA	54.18	4,971.79	70.51	48.97	4,098.67	64.02
6	26100070	129	BUGALAGRANDE	19.29	605.48	24.61	17.08	479.85	21.91
7	26120120	385	CAICEDONIA	24.87	1,009.58	31.77	21.73	750.99	27.40
8	26080280	20	CALI	30.03	1,360.97	36.89	27.06	1,160.82	34.07
9	26060020	57	CANDELARIA	21.32	786.76	28.05	18.65	551.49	23.48
10	26100830	309	CARTAGO	19.28	602.45	24.54	17.79	508.26	22.54
11	53100040	404	DAGUA	24.14	935.94	30.59	22.05	789.23	28.09
12	26110120	504	EL AGUILA	27.22	1,258.73	35.48	21.82	778.57	27.90
13	26090460	314	EL CERRITO	21.03	704.63	26.54	19.13	576.42	24.01
14	26070760	3	FLORIDA	25.36	1,192.66	34.53	21.73	747.16	27.33
15	26090060	34	GINEBRA	19.83	572.68	23.93	18.98	559.86	23.66
16	26090630	195	GUACARI	18.30	511.37	22.61	17.14	466.76	21.60
17	26100690	397	GUADALAJARA DE BU	23.31	854.23	29.23	21.46	717.38	26.78
18	53110100	698	LA CUMBRE	14.40	302.14	17.38	13.27	285.08	16.88
19	26110290	189	LA UNION	17.64	488.30	22.10	16.50	446.78	21.14
20	26100300	166	OBANDO	18.08	496.31	22.28	16.49	436.98	20.90
21	26075040	202	PALMIRA	16.30	407.95	20.20	14.89	352.86	18.78
22	26070170	766	PRADERA	18.74	559.69	23.66	16.93	453.35	21.29
23	26110230	437	ROLDANILLO	17.31	464.60	21.55	16.85	457.49	21.39
24	26120180	283	SEVILLA	21.50	794.95	28.19	19.86	627.58	25.05
25	26110300	182	TORO	17.34	464.49	21.55	16.85	453.67	21.30
26	26100350	345	TULUA	25.73	1,115.05	33.39	22.77	833.07	28.86

27	54030010	206	VERSALLES	15.73	403.89	20.10	14.30	327.90	18.11
28	26080070	354	VIJES	14.67	313.55	17.71	13.98	307.43	17.53
29	26080290	156	YOTOCO	17.47	454.62	21.32	17.09	479.37	21.89
30	26100780	278	ZARZAL	17.70	493.16	22.21	16.52	444.07	21.07

Tabla 18. Métricas de Evaluación Modelo ARIMA

				TEST			TRAIN		
Número de estación	**Station_code**	**missing_count**	**Municipio**	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
1	26120150	340	ALCALA	23.20	869.30	29.48	20.63	670.11	25.89
2	26100700	661	ANDALUCIA	21.19	714.91	26.74	18.92	559.30	23.65
3	26110150	563	ANSERMANUEVO	20.75	709.31	26.63	18.42	550.45	23.46
4	26110040	340	BOLIVAR	18.23	531.48	23.05	16.29	428.22	20.69
5	53115010	501	BUENAVENTURA	53.21	4,865.21	69.75	47.19	3,840.64	61.97
6	26100070	129	BUGALAGRANDE	18.84	586.17	24.21	16.72	454.72	21.32
7	26120120	385	CAICEDONIA	24.30	987.18	31.42	21.32	709.80	26.64
8	26080280	20	CALI	29.81	1,344.34	36.67	26.70	1,111.16	33.33
9	26060020	57	CANDELARIA	21.56	763.68	27.63	18.29	517.73	22.75
10	26100830	309	CARTAGO	18.43	582.17	24.13	17.61	482.27	21.96
11	53100040	404	DAGUA	24.05	922.54	30.37	21.67	752.86	27.44
12	26110120	504	EL AGUILA	25.84	1,248.13	35.33	21.20	728.31	26.99
13	26090460	314	EL CERRITO	20.55	695.91	26.38	18.71	546.67	23.38
14	26070760	3	FLORIDA	25.56	1,162.17	34.09	21.19	700.54	26.47
15	26090060	34	GINEBRA	19.45	561.06	23.69	18.55	531.64	23.06
16	26090630	195	GUACARI	18.01	499.62	22.35	16.77	444.86	21.09
17	26100690	397	GUADALAJARA DE BU	23.16	841.86	29.01	20.90	670.02	25.88
18	53110100	698	LA CUMBRE	14.18	296.70	17.22	13.11	273.94	16.55
19	26110290	189	LA UNION	16.74	461.90	21.49	16.28	415.96	20.40
20	26100300	166	OBANDO	17.44	480.69	21.92	16.21	408.75	20.22
21	26075040	202	PALMIRA	15.96	403.40	20.08	14.47	330.23	18.17

22	26070170	766	PRADERA	18.50	555.53	23.57	16.55	422.60	20.56
23	26110230	437	ROLDANILLO	16.82	450.90	21.23	16.58	430.62	20.75
24	26120180	283	SEVILLA	20.77	780.00	27.93	19.43	593.33	24.36
25	26110300	182	TORO	16.46	436.46	20.89	16.56	426.90	20.66
26	26100350	345	TULUA	25.31	1,091.09	33.03	22.25	792.10	28.14
27	54030010	206	VERSALLES	15.34	395.33	19.88	13.98	310.92	17.63
28	26080070	354	VIJES	14.51	310.01	17.61	13.78	296.85	17.23
29	26080290	156	YOTOCO	17.00	439.84	20.97	16.72	451.03	21.24
30	26100780	278	ZARZAL	17.09	476.53	21.83	16.20	417.32	20.43

A partir de los resultados presentados, se realiza una comparación entre los tres enfoques de modelado de series temporales: ARIMA, SARIMA y LSTM. Para ello, se consideran las principales métricas de evaluación: MAE, MSE y RMSE, que permiten analizar la precisión y el ajuste de cada uno de los modelos.

8.1 Comparación ARIMA vs. LSTM:

MAE: En cuanto al MAE, LSTM registra una mejor precisión en la mayoría de las estaciones. Por ejemplo, en la estación Cartago (código 26100830), el MAE del LSTM es de 15.92, mientras que en ARIMA es 18.43, lo que sugiere que el error promedio absoluto es menor en el enfoque basado en redes neuronales recurrentes. Esto también se observa en estaciones como Zarzal (código 26100780), donde el MAE del LSTM es de 15.60, superando al ARIMA que obtiene un MAE de 17.09.

MSE y RMSE: En términos de MSE y RMSE, que penalizan errores grandes, el LSTM también muestra un mejor rendimiento en casi todas las estaciones. Por ejemplo, en Cali (código 26080280), el MSE del LSTM es de 624.30, considerablemente menor que el de ARIMA que alcanza 1,344.34. Esto refuerza la idea de que LSTM maneja mejor los errores grandes y se ajusta con más precisión a las series temporales, lo que se ve reflejado en valores menores de RMSE en todas las estaciones.

En resumen, LSTM presenta mejores resultados a ARIMA en todas las métricas de evaluación. Esto se debe principalmente a que los modelos basados en redes neuronales como LSTM pueden capturar patrones no lineales y dependencias a largo plazo en los datos, mientras que ARIMA, al ser un modelo lineal, es más limitado para este tipo de series temporales complejas.

8.2 Comparación SARIMA vs. LSTM:

MAE: Al comparar el MAE, LSTM sigue mostrando un mejor desempeño en casi todas las estaciones. En la estación Yotoco (código 26080290), el MAE del SARIMA es de 17.47, mientras que en LSTM es de 14.50, lo que refleja una mejora significativa en la precisión del modelo basado en redes neuronales. Asimismo, en la estación Tuluá (código 26100350), LSTM logra un MAE de 23.70, mientras que SARIMA tiene un valor más

alto de 25.73.

MSE y RMSE: Similar a la comparación con ARIMA, LSTM también presenta mejores valores de MSE y RMSE frente a SARIMA. Por ejemplo, en la estación La Cumbre (código 53110100), el LSTM obtiene un RMSE de 4.06, mientras que SARIMA obtiene un RMSE mayor de 17.38, lo que implica que LSTM es más efectivo en la predicción de la serie temporal al minimizar los errores grandes.

En esta comparación, aunque SARIMA mejora ligeramente respecto a ARIMA en algunas estaciones, en general LSTM sigue superando a SARIMA en todas las métricas clave. Esto es coherente con las limitaciones de los modelos SARIMA, que asumen una estacionalidad regular y lineal, mientras que los datos de precipitación tienden a ser más irregulares y presentan relaciones no lineales que LSTM puede modelar mejor.

8.3 Comparación ARIMA vs. SARIMA vs. LSTM:

MAE y RMSE: Al observar los valores de MAE y RMSE, LSTM también muestra un rendimiento más consistente y preciso en la mayoría de las estaciones. En términos de error promedio absoluto (MAE), LSTM suele estar por debajo de ARIMA y SARIMA, lo que implica una mayor precisión en las predicciones. En la estación Candelaria (código 26060020), por ejemplo, LSTM muestra un MAE de 15.50, mientras que ARIMA tiene 21.56 y SARIMA 21.32.

MSE: LSTM también tiene un desempeño superior. Por ejemplo, en la estación Pradera (código 26070170), LSTM presenta un MSE de 2,223.99, menor que el MSE de SARIMA (559.69), lo que indica que este último tiene dificultades para captar correctamente los patrones de largo plazo y presenta errores más grandes.

9. CONCLUSIONES Y TRABAJOS FUTUROS

9.1. CONCLUSIONES

El análisis de los resultados obtenidos en este estudio denota una amplia variabilidad en el rendimiento del modelo LSTM implementado para la predicción de precipitaciones entre las distintas estaciones meteorológicas del Valle del Cauca. El rendimiento del modelo, medido a través de métricas como el MAE, MSE y RMSE, muestra que las estaciones con patrones de precipitación más regulares y menos datos faltantes, como Palmira y Cali, tienden a presentar los mejores resultados. En estas estaciones, los errores de predicción son menores y el coeficiente de determinación (R^2) indica que el modelo es capaz de capturar una buena parte de la variabilidad en los datos.

Por otro lado, estaciones ubicadas en regiones geográficamente más complejas, como Buenaventura y Pradera, presentan valores de error significativamente más altos. Esto sugiere que la variabilidad climática y las características topográficas influyen en gran medida en la capacidad del modelo para predecir con precisión las precipitaciones. En estas áreas, el modelo tiene dificultades para capturar la dinámica de las lluvias, posiblemente debido a la influencia de factores externos no considerados, como la cercanía al océano, el fenómeno de la Niña, o la orografía accidentada que altera los patrones de lluvia.

La presencia de datos faltantes también juega un papel fundamental en el rendimiento del modelo. Las estaciones con mayor número de datos faltantes, como La Cumbre, Pradera y El Águila, muestran consistentemente errores más altos en comparación con aquellas que tienen un registro más completo. Esta observación resalta la importancia de la calidad y continuidad en la recolección de datos para el desarrollo de modelos predictivos precisos y fiables.

Otro aspecto relevante que se puede concluir es la capacidad del modelo para generalizar correctamente entre el conjunto de entrenamiento y el conjunto de prueba. En la mayoría de las estaciones, no se observan diferencias significativas entre los errores de entrenamiento y prueba, lo que sugiere que el modelo no está sobreajustado y tiene una capacidad razonable para predecir en datos no conocidos. Sin embargo, en estaciones como Buenaventura y Pradera, las diferencias son mayores, lo que podría indicar que el modelo ha capturado patrones específicos de los datos de entrenamiento que no generalizan bien al conjunto de prueba.

También, es importante destacar que las características locales de las estaciones, como la ubicación geográfica y el tipo de clima, parecen influir significativamente en el rendimiento del modelo. Esto sugiere que un único modelo LSTM para todas las estaciones del departamento puede no ser la solución óptima, dado que los patrones de precipitación son altamente específicos a cada región. Por tanto, el desarrollo de modelos adaptados a las características de cada área podría ser una alternativa viable para mejorar la precisión predictiva.

A su vez, y teniendo en cuenta la comparación con modelos de series temporales lineales, los resultados demuestran que el modelo LSTM supera claramente tanto a ARIMA como a SARIMA en todas las métricas clave, así como en la reducción del error promedio absoluto (MAE) y el error cuadrático medio (MSE y RMSE). Los modelos ARIMA y SARIMA no logran modelar eficazmente las relaciones no lineales presentes en las series temporales de precipitación, lo que sugiere que, para este tipo de datos meteorológicos, las redes neuronales recurrentes LSTM son una mejor alternativa para la predicción

Finalmente, a pesar de los datos atípicos en métricas de evaluación observados en algunas estaciones, el modelo LSTM ha mostrado ser una herramienta útil para predecir precipitaciones, especialmente en estaciones con condiciones climáticas más predecibles. Sin embargo, el desempeño del modelo en estaciones con alta variabilidad climática y datos faltantes plantea la necesidad de implementar estrategias más robustas para abordar estos desafíos.

9.2. TRABAJOS FUTUROS

En relación con futuros trabajos relacionados con la predicción de precipitaciones utilizando modelos de aprendizaje automático, es esencial enfocar los esfuerzos en mejorar la calidad y continuidad de los datos meteorológicos. La presencia de datos faltantes ha demostrado tener un impacto significativo en el rendimiento del modelo, afectando negativamente su capacidad predictiva. Por lo tanto, sería beneficioso implementar estrategias que mitiguen el impacto de los datos faltantes, como técnicas de imputación de datos diferentes a la implementada en este proyecto, que podrían llenar los vacíos de información con base en patrones históricos o estaciones cercanas. Además, es recomendable revisar la infraestructura de recolección de datos en aquellas estaciones con un mayor número de faltantes, ya que la falta de continuidad en la recolección puede estar relacionada con problemas técnicos en los equipos de medición.

Otro aspecto para considerar para el desarrollo de futuras investigaciones es la incorporación de más características meteorológicas en el proceso de modelado. El modelo actual se basa únicamente en la precipitación histórica para realizar sus predicciones, lo cual puede ser una limitación. La inclusión de variables adicionales, como la temperatura, la presión atmosférica, la humedad, e incluso información topográfica, podría enriquecer el modelo y permitirle capturar patrones más complejos. Estas características adicionales podrían ser particularmente útiles en estaciones donde los patrones de precipitación son altamente irregulares, como en Buenaventura y Pradera. Al incluir más información sobre el entorno climático, el modelo podría aprender interacciones más precisas entre las variables y mejorar la precisión de las predicciones.

Otra línea de investigación futura podría centrarse en el desarrollo de modelos específicos para diferentes regiones o tipos de clima. El análisis de los resultados ha mostrado que las estaciones ubicadas en áreas con climas similares tienen un comportamiento predictivo más uniforme, mientras que estaciones con climas o topografías particulares, como áreas costeras o montañosas, presentan mayores dificultades. En lugar de utilizar un único modelo para todo el departamento, podría ser más efectivo dividir las estaciones en grupos con características climáticas similares y entrenar modelos específicos para cada grupo. Esto permitiría a los modelos adaptarse mejor a las particularidades de cada región, mejorando así el rendimiento general.

Además, se podrían explorar arquitecturas de redes neuronales más complejas, como modelos híbridos que combinen LSTM y CNN. Estos, podrían capturar tanto la dependencia temporal como las características espaciales de las precipitaciones, proporcionando una solución más robusta para estaciones con alta variabilidad geográfica.

Finalmente, es recomendable que futuros estudios realicen un análisis más detallado de los patrones estacionales de precipitación, ya que la precipitación tiende a variar significativamente entre temporadas en

regiones como el Valle del Cauca. Incorporar esta estacionalidad en los modelos podría mejorar su capacidad para predecir eventos extremos de lluvia o períodos secos, que son críticos para la gestión de recursos naturales y la planificación agrícola.

10. REFERENCIAS BIBLIOGRÁFICAS

[1] National Geographic. (2023). ¿Cuál es el país donde más llueve? [En Línea]. Disponible: https://www.nationalgeographicla.com/medio-ambiente/2023/05/cual-es-el-pais-donde-mas-llueve

[2] M. Zuluaga y G. Poveda, “Diagnóstico de sistemas convectivos de mesoescala sobre Colombia y el océano Pacífico Oriental durante 1998-2002", Av. Recur. Hidraul., n.º 11, pp. 145–160, ene. 2004

[3] Departamento Nacional de Planeación. (2014). Impactos económicos del cambio climático en Colombia. [En Línea]
Disponible:https://colaboracion.dnp.gov.co/CDT/Ambiente/Impactos%20economicos%20Cambio%20climático.pdf

[4] Instituto de Hidrología, Meteorología y Estudios Ambientales. (2023). Informe de predicción climática a corto, mediano y largo plazo. [En Línea]

[5] W.P Lowry, Compendio de apuntes de climatología para la formación de personal meteorológico de la clase IV. Ginebra: Organización Meteorológica Mundial, 1973, pp. 167.

[6] American Meteorological Society. Precipitation. (2021). Glossary of Meteorology. [En Línea]. Disponible: https://glossary.ametsoc.org/wiki/Precipitation

[7] J. Eslava, “Climatología y diversidad climática de Colombia”, Rev. Acad. Colomb. Cienc. 18(71), pp. 507-538, marzo 1993.

[8] J. Montealegre. (2009). Estudio de la variabilidad climática de la precipitación en Colombia asociada a procesos oceánicos y atmosféricos de meso y gran escala. [En Línea]. Disponible: http://www.ideam.gov.co/documents/21021/21789/Estudio+de+la+variabilidad+climática+de+la.pdf/643c4c0e-83d7-414f-b2b4-6953f64078d3

[9] J. Guerrero, Predicción niveles de precipitación para la identificación de zonas agrícolas, Colombia: Universidad Internacional de la Rioja, 2021.

[10] D. Herrera y E. Aristizábal, «Artificial Intelligence and machine learning model for spatial and temporal prediction of Drought events in the Magdalena department, Colombia» INGE CUC, vol. 18, nº 2, pp. 249-265, 2022.

[11] P. González, “Análisis de Series Temporales: Modelos ARIMA", Facultad de Ciencias Económicas y Empresariales, pp. 6 - 7, abril 209.

[12] R. Cruz, “Modelo Híbrido Predictivo y de Recomendación con Técnicas de Minería de Datos e Inteligencia Artificial", Doctor en Ciencias Computacionales, Computación y electrónica, Universidad Autónoma del Estado de Hidalgo, Hidalgo, Marzo de 2018.

[13] Saavedra O., Tomas S., & Espinoza I. "(2023). Predicción del ciclo solar 25 mediante modelos ARIMA y redes neuronales LSTM." Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales,

47(183):400-411, abril-junio de 2023

[14] Y. Son et al, “LSTM–GAN based cloud movement prediction in satellite images for PV forecast”, Rev. Journal of Ambient Intelligence and Humanized Computing. 14, pp. 12373–12386, 2023.

[15] T. Vo et al, “LSTM-CM: a hybrid approach for natural drought prediction based on deep learning and climate models”, Rev. Stochastic Environmental Research and Risk Assessment. 37, pp. 2035–2051, 2023

[16] C. Batista et al, “Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal", Tecnología en Marcha, Encuentro de Investigación y Extensión 2016, pp. 37 - 38, mayo 2016.

[17] J. Pérez, “Redes Recurrentes", Trabajo de Grado Matemáticas y Estadística, Computación y electrónica, Universidad de Sevilla, Sevilla, junio de 2020.

[18] Asociación Nacional de Comercio Exterior ANALDEX. (2023). Informe Producto Interno Bruto 2023. [En Línea]. Disponible : https://www.analdex.org/2024/03/04/informe-producto-interno-bruto-de-2023/#:~:text=En%202023%2C%20el%20Producto%20Interno,las%20economías%20a%20nivel%20global

[19] H. Banguero (1991). El Modelo de desarrollo del Valle del Cauca en retrospectiva y prospectiva. [En Línea]. Disponible: https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/4015/1/Modelo_desarrollo_valle.pdf

[20] Buddget Direct (2021). The wettest city in every country. [En Línea]. Disponible: https://www.budgetdirect.com.au/home-contents-insurance/home-safety/storm-season/wettest-cities.html

[21] I. López y R. Velázquez, «Impacto en el riesgo climático de actividades económicas, análisis del sector líneas áreas» Contabilidad y Negocios, vol. 15, nº 29, pp. 40-57, 2020.

[22] E. Molua, y C. Lambi, (2007). The Economic Impact of Climate Change on Agriculture in Cameroon. Pretoria, Southafrica: The Wold Bank.

[23] C. Zheng et al, “TISE-LSTM: A LSTM model for precipitation nowcasting with temporal interactions and spatial extract blocks”, Neurocomputing 590, pp. 1-19, Abril 2024.