

**Modelo de Machine Learning para la predicción de inundaciones en el municipio de La  
Virginia Risaralda**

Carlos Hernán Bermúdez Betancur

Asesor

Rafael Roberto Ruiz Escorcía

Universidad Nacional Abierta y a Distancia UNAD  
Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI  
Especialización en Ciencia de Datos y Analítica

2025

### **Dedicatoria**

Este trabajo está dedicado con todo mi amor y gratitud a mi esposa quien ha sido mi mayor fuente de inspiración y apoyo incondicional a lo largo de este camino, por su paciencia y comprensión, a mis hijos, los cuales me recuerdan la importancia de esforzarme para poder construirles un futuro así como demostrarles que todo en la vida se puede lograr con esfuerzo y disciplina.

### **Agradecimientos**

Quiero expresar mi más sinceros agradecimientos a todos los docentes que me acompañaron en la construcción este camino, brindándome su conocimiento, guía y apoyo para superar los retos académicos y personales, en especial agradezco al profesor Nilson Ferreira cuyo soporte incondicional y guía constante fueron determinantes para el éxito de mi carrera profesional.

## Resumen

En respuesta a la creciente amenaza de inundaciones generadas por el cambio climático y la urbanización inadecuada este proyecto propone el desarrollo de un modelo de Machine Learning para mejorar la precisión y la confiabilidad de las predicciones de inundaciones en La Virginia Risaralda, este proyecto se centró en la creación de un sistema de alerta temprana y un modelo predictivo basado en la recopilación y análisis de datos hidrometeorológicos, este busca brindar una herramienta importante para la gestión de desastres en La Virginia Risaralda realizado a través de un marco replicable para otras regiones con problemas similares contribuyendo así al cuerpo de conocimiento global sobre la aplicación de Machine Learning en la predicción y gestión de inundaciones.

***Palabras claves:*** Sistema de alertas tempranas, machine learning, inundaciones.

### **Abstract**

In response to the growing threat of floods generated by climate change and inadequate urbanization this project proposes the development of a Machine Learning model to improve the accuracy and reliability of flood predictions in La Virginia Risaralda, this project focused on the creation of an early warning system and a predictive model based on the collection and analysis of hydrometeorological data, This project seeks to provide an important tool for disaster management in La Virginia Risaralda through a replicable framework for other regions with similar problems, thus contributing to the global body of knowledge on the application of Machine Learning in flood prediction and management.

***Keywords:*** Early warning system, machine learning, floods.

## Tabla de Contenido

Introducción .....	11
Descripción del Problema .....	12
Planteamiento del Problema.....	12
Sistematización del Problema .....	14
Justificación .....	15
Objetivos .....	16
Objetivo General .....	16
Objetivos Específicos.....	16
Marco de Referencia .....	17
Estado del Arte .....	17
Marco Contextual.....	17
Marco Teórico.....	18
Definición de Inundaciones.....	18
Sistemas de Alerta Temprana (SAT).....	18
Factores Antropogénicos y Vulnerabilidad.....	18
Marco Conceptual .....	19
Marco Normativo .....	19
Metodología .....	21
Método .....	21
Tipo de Estudio .....	21
Recolección de Datos .....	21
Instrumentos y Herramientas .....	22

Hardware .....	23
Resultados .....	24
Desarrollo de Modelos .....	24
Limpieza y Exploración de los Datos .....	24
Resultados del Modelado .....	32
Evaluación del Desempeño .....	34
Conclusiones .....	35
Recomendaciones .....	37
Referencias Bibliográficas .....	38
Apéndices.....	40

**Lista de Tablas**

<b>Tabla 1</b> <i>Resultado de la Limpieza de los Datos</i> .....	27
<b>Tabla 2</b> <i>Resultados del Modelado en XGBoost de los Ríos Cauca y Risaralda</i> .....	32
<b>Tabla 3</b> <i>Resultados del Test de ADF de los Ríos Cauca y Risaralda</i> .....	34

## Lista de Figuras

<b>Figura 1</b> <i>Metodología SCRUM Aplicada al Proyecto</i> .....	22
<b>Figura 2</b> <i>Tablero con el Desarrollo del Proyecto</i> .....	23
<b>Figura 3</b> <i>Dataframes de los Ríos Cauca y Risaralda</i> .....	24
<b>Figura 4</b> <i>Series de Tiempo de los Ríos Cauca y Risaralda</i> .....	28
<b>Figura 5</b> <i>Matriz de Correlación</i> .....	29
<b>Figura 6</b> <i>Histogramas de los Ríos Cauca y Risaralda</i> .....	30
<b>Figura 7</b> <i>Boxplots de los Ríos Cauca y Risaralda</i> .....	31
<b>Figura 8</b> <i>Determinación de los Mejores Hiperparametros</i> .....	32
<b>Figura 9</b> <i>Comparación Entre Valores Reales y Predichos</i> .....	33

**Lista de Apéndices**

<b>Apéndice A</b> <i>Autorización de Uso de Datos de los Afluentes Cauca y Risaralda</i> .....	40
--	----

## Introducción

Las inundaciones representan uno de los eventos de desastres naturales más frecuentes y arrasadores a nivel mundial, en el caso de Colombia la variabilidad climática y la ubicación geográfica del país sumados el crecimiento urbano desorganizado y la deforestación han incrementado la vulnerabilidad de diversas regiones frente a estos eventos, en particular el municipio de La Virginia Risaralda se encuentra expuesto a inundaciones debido a su proximidad a los ríos Cauca y Risaralda los cuales presentan crecidas significativas en temporadas de lluvias fuertes, a pesar de los esfuerzos locales para el monitoreo de los niveles de los ríos, la falta de herramientas tecnológicas avanzadas limitan la capacidad de respuesta oportuna ante posibles emergencias, en este contexto el desarrollo de un modelo de aprendizaje automático se presenta como una solución viable para mejorar la precisión en la predicción de inundaciones y fortalecer la gestión del riesgo de la región.

Este trabajo propone la implementación de un modelo basado en técnicas de machine learning que permita analizar datos en tiempo real optimizando la capacidad de los sistemas de alerta temprana en La Virginia Risaralda, para ello se utilizarán algoritmos avanzados capaces de identificar patrones y tendencias en los niveles de los ríos permitiendo la predicción de eventos extremos y permitiendo la toma de decisiones a tiempo por parte de las autoridades y la comunidad, buscando así un marco replicable para otras regiones con problemáticas similares impulsando el uso de la inteligencia artificial en la mitigación de desastres naturales.

## Descripción del Problema

### Planteamiento del Problema

De acuerdo con el artículo redactado por Arreguín (2016) habla de la complejidad que existe por las inundaciones en el mundo en donde se ha incrementado su frecuencia e intensidad, siendo más difíciles de controlar todo esto a raíz del cambio climático y sus deficiencias en las políticas públicas esto sumado al crecimiento urbanístico en las ciudades y su falta de planeación tanto en la ubicación como en la infraestructura de alcantarillado requiriendo una participación urgente de la sociedad y sus instituciones gubernamentales en la búsqueda de soluciones de manera coordinada haciendo sinergia entre las partes.

Particularmente Colombia se encuentra ubicada en un lugar que de acuerdo con Urrutia (2012) nuestro sistema montañoso afecta el clima de nuestro país haciéndolo muy variado desde el nivel del mar hasta los casi 4900 metros sobre el nivel del mar, hace que el estado del tiempo sea diverso como en el interior del país en donde se dan dos espacios de estado de tiempo en el año, lluvia desde aproximadamente mediados hasta fin de año y la temporada seca entre los primeros meses, sumado a esto el deterioro en el que vive de manera constante nuestra naturaleza por parte nuestra hace que sucedan de manera inevitable e incierta inundaciones particularmente que de acuerdo con la Unidad de Gestión de Riesgo de Desastres (2022) se da en los lugares poblados muy cerca a los afluentes el cual deja a su paso pérdidas materiales, económicas, sociales y humanas, además, de acuerdo con Ávila, Carvajal y Sedano (2013) comentan cómo diferentes fenómenos naturales han afectado su población Colombiana incluso llegando a duplicar sus pérdidas.

Actualmente el municipio de La Virginia Risaralda está colindando en sus orillas por los ríos Cauca que de acuerdo con Urrutia (2012) es el segundo afluente con mayor longitud en

nuestro país y el río Risaralda (éste se une al Cauca) esto permite que La Virginia presente inundaciones, que de acuerdo con Londoño y Ocampo (2017) es un hecho que se viene repitiendo desde hace mucho tiempo y esto ha dejado como consecuencia pérdidas físicas y económicas limitando así su crecimiento económico.

De acuerdo con lo expresado anteriormente se genera la necesidad de responder a las siguientes presuntas que cuyo núcleo central se busca resolver a través de este proyecto:

¿Cómo puede la implementación de un modelo de aprendizaje automático avanzar en la precisión y confiabilidad de las alertas tempranas de inundaciones en el municipio de La Virginia Risaralda? ¿Qué métodos y técnicas de aprendizaje automático podrían utilizarse que permitan aumentar la calidad de las predicciones de las inundaciones? ¿Qué modelo de aprendizaje automático podría predecir de manera precisa las inundaciones en el municipio de La Virginia?

### **Sistematización del Problema**

- ¿Cómo un modelo predictivo puede mejorar las alertas tempranas en La Virginia

Risaralda?

- ¿Qué variables hidrometeorológicas son importantes para garantizar esta precisión?

- ¿Cómo validar el desempeño del modelo en situaciones de riesgo extremo?

### **Justificación**

De acuerdo con datos entregados por parte del Cuerpo de Bomberos Voluntarios de La Virginia Risaralda la medición de los niveles del río Cauca y del río Risaralda se realizaba de manera manual y visual utilizando regletas ubicadas en los puentes Bernardo Arango y Jaramillo Ochoa para el río Cauca y en el Puente Negro para el río Risaralda, la frecuencia de estas mediciones variaba según las condiciones climáticas oscilando entre cada 12 horas y cada 30 minutos, esta situación implica que ante un incremento repentino del caudal de los ríos no se cuentan con medidas de prevención inmediatas que permitan atenuar las consecuencias de un posible desastre, hace poco se implementó un sistema que se encuentra en fase piloto que permite la toma de muestras con una mayor frecuencia, siendo éste el primer municipio en implementarlo a través de la Gobernación.

Un Sistema de Alerta Temprana “permite conocer anticipadamente y con cierto nivel de certeza en qué tiempo y espacio, una amenaza o evento adverso de tipo natural o generado por la actividad humana puede desencadenar situaciones potencialmente peligrosas” (Buitrago, 2015, p.7) todo esto se puede llevar con éxito si se cuenta con un modelo de predicción de Machine Learning que permita analizar los datos en tiempo real para poder generar las alertas necesarias y así prevenir consecuencias.

Incluso, de acuerdo con las Naciones Unidas (2022) hoy en día con los Objetivos de Desarrollo Sostenibles se busca generar cambios positivos y aportar a sus objetivos, en este caso al 9b, 10.3, 11.1, 11.5 y 11b los cuales buscan reducir el número de muertes por desastres y también económicas de personas vulnerables, la implementación de políticas en la adaptación del cambio climático y la resiliencia a desastres garantizando así la vivienda, la reducción de la desigualdad a través de políticas mediante la tecnología, la investigación y la innovación.

## Objetivos

### Objetivo General

Desarrollar un modelo de aprendizaje automático para cada afluente que permita mejorar la precisión en la predicción de inundaciones de los Ríos Cauca y Risaralda en el municipio de La Virginia Risaralda.

### Objetivos Específicos

Realizar un análisis exploratorio y limpieza de los datos que permita la identificación de patrones, tendencias y posibles valores atípicos en el desarrollo del modelo.

Desarrollar modelos de aprendizaje automático para cada afluente de los ríos Cauca y Risaralda evaluando diferentes parámetros y seleccionando aquellos que sean óptimos para la precisión en la predicción de las inundaciones.

Evaluar el desempeño de los modelos desarrollados para los ríos Cauca y Risaralda aplicando las métricas de validación como el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE), la Validación Cruzada (Cross-Validation Score) y el Coeficiente de Correlación ( $R^2$ ) que permitan determinar su efectividad y precisión en la predicción de inundaciones.

## **Marco de Referencia**

### **Estado del Arte**

El avance de los sistemas de predicción y mitigación de desastres naturales ha llevado al desarrollo de herramientas cada vez más sofisticadas para enfrentar las inundaciones, según Zhou & Kang (2023) los modelos de aprendizaje automático como Regresión de Vectores de Soporte (SVR), Regresión de Procesos Gaussianos (GPR), Regresión con Bosques Aleatorios (RFR), Percepción Multicapa (MLP), Memoria a Largo Corto Plazo (LSTM) y Unidad Recurrente Cerrada (GRU) han demostrado diferentes niveles de eficacia en la predicción de las inundaciones, en donde el modelo GRU particularmente se muestra como el más eficiente logrando reducciones significativas en métricas de error como MAPE y RMSE, en Colombia Sedano-Cruz et al. (2013) identifican que las inundaciones no solo son un resultado de fenómenos naturales sino también de factores antropogénicos como la deforestación, la mala gestión de recursos hídricos y la planificación urbana deficiente, por otro lado Domínguez-Calle y Lozano-Báez (2014) muestran la importancia de los Sistemas de Alerta Temprana (SAT) como herramientas de suma importancia para mitigar el impacto de las inundaciones a través de advertencias tempranas basadas en datos climáticos y ambientales.

### **Marco Contextual**

Colombia es un país altamente vulnerable a inundaciones debido a su compleja geografía y variabilidad climática, en donde regiones como La Virginia, Risaralda se encuentran particularmente expuestas debido a su proximidad a ríos caudalosos y patrones climáticos cambiantes, en donde "el fenómeno de inundación de manera frecuente, de acuerdo a información y recuentos históricos dicho fenómeno evidencia niveles de afectación de la población que tiene que enfrentar las pérdidas físicas y económicas como resultado de las

mismas, es por esto que dichas limitantes atrasan el desarrollo urbano del municipio de La Virginia Risaralda" (Salinas Díaz & Patiño Rodríguez, 2024, p. 10).

## **Marco Teórico**

### ***Definición de Inundaciones***

Según Smith y Ward (1998) las inundaciones se definen como la acumulación temporal de agua en terrenos normalmente secos debido a fenómenos como lluvias intensas, deshielos rápidos o fallos en estructuras artificiales, estas constituyen uno de los desastres naturales más comunes y devastadores capaces de generar daños significativos en infraestructura, economías locales y vidas humanas.

### ***Sistemas de Alerta Temprana (SAT)***

Son herramientas que permiten anticipar eventos adversos a través de la monitorización continua de factores ambientales, según Domínguez-Calle y Lozano-Báez (2014) estos sistemas son importantes para reducir las pérdidas económicas, ambientales y humanas, además incorporan análisis de variables como precipitación, niveles de ríos y condiciones del suelo para emitir advertencias oportunas a las comunidades vulnerables.

### ***Factores Antropogénicos y Vulnerabilidad***

Las actividades humanas como la deforestación y la inadecuada planificación urbana alteran los ciclos hidrológicos y aumentan el riesgo de inundaciones (Sedano-Cruz et al., 2013), además la mala gestión de recursos hídricos, la expansión descontrolada de asentamientos humanos y la falta de infraestructura adecuada aceleran los impactos de estos eventos en especial en regiones donde la geografía y la hidrografía son factores determinantes.

## **Marco Conceptual**

**Inundaciones:** Fenómeno natural caracterizado por la acumulación temporal de agua en zonas normalmente secas, que puede ser causado por lluvias excesivas, deshielos rápidos o fallos en infraestructuras (Smith y Ward, 1998).

**Sistemas de Alerta Temprana (SAT):** Herramientas de monitorización y análisis continuo de variables ambientales para emitir advertencias tempranas y mitigar los impactos de desastres naturales (Domínguez-Calle y Lozano-Báez, 2014).

**Aprendizaje Automático o Machine Learning:** Es un método computacional que emplea algoritmos para modelar y predecir relaciones complejas entre variables hidrometeorológicas mejorando la precisión en la predicción de inundaciones (Zhou & Kang, 2023).

**Variables Hidrometeorológicas:** Factores como la precipitación, los niveles de ríos, las condiciones del suelo y los datos topográficos, que son de importancia para la precisión de los modelos predictivos (Di Baldassarre et al., 2009).

**Umbrales:** Estos están definidos de manera semaforizada en centímetros (cm), para el río Cauca esta entre 350 cm y 409 cm como alerta amarilla, entre 410 cm y 519 cm como alerta naranja y entre 520 cm hasta 1000 cm como alerta roja; para el río Risaralda está definido entre 200 cm y 250 cm como alerta amarilla, entre 251 cm y 300 cm como alerta naranja y entre 301 cm y 600 cm como alerta roja (Elkin Ramírez, comunicación personal, 26 de diciembre de 2024).

## **Marco Normativo**

En Colombia el marco legal para la gestión del riesgo de desastres está establecido por la Ley 1523 de 2012 (Congreso de Colombia, 2012) la cual regula la política nacional de gestión del riesgo y promueve el desarrollo de capacidades institucionales para prevenir y mitigar los

impactos de desastres naturales, incluidas las inundaciones, esta normativa muestra la importancia de incorporar tecnologías disruptivas como los SAT y fomenta una gestión integral de los recursos hídricos, de igual forma promueve la adopción de opciones sostenibles en la planificación urbana y el manejo ambiental para reducir la vulnerabilidad de las comunidades.

## **Metodología**

### **Método**

Este proyecto utiliza un enfoque cuantitativo, aplicado y experimental centrado en el desarrollo de un modelo de aprendizaje automático para predecir inundaciones en los ríos Cauca y Risaralda en el municipio de La Virginia Risaralda, éste permite establecer relaciones entre las variables hidrometeorológicas y los patrones de inundación mediante técnicas avanzadas de análisis de datos y machine learning, además se implementará la metodología ágil SCRUM, lo que permite un desarrollo iterativo e incremental del modelo, generando una adaptabilidad y un enfoque en los resultados prácticos.

### **Tipo de Estudio**

El estudio es de tipo aplicado-experimental ya que busca resolver un problema práctico mediante la implementación y evaluación de un modelo predictivo, se diseñarán experimentos para evaluar diferentes técnicas de machine learning tales como redes neuronales, regresión y modelos de aprendizaje profundo, los resultados serán validados con datos históricos y mediciones actuales de los niveles de los ríos Cauca y Risaralda.

### **Recolección de Datos**

La recolección de datos se llevará a cabo en tres fases principales:

- Fuentes de Datos Existentes: Datos hidrometeorológicos históricos proporcionados por el Cuerpo de Bomberos Voluntarios de La Virginia (ver Anexo A).
- Monitoreo en Tiempo Real: Uso de estaciones meteorológicas locales para obtener información detallada y precisa sobre las condiciones climáticas actuales, además la incorporación de sensores automatizados y dispositivos IoT para capturar datos en tiempo real como niveles de los ríos, precipitación, temperatura y humedad.

- **Análisis y Limpieza de Datos:** Se realizará un análisis exploratorio y limpieza de los datos recolectados, eliminando valores atípicos, registros incompletos e información irrelevante para obtener una calidad y confiabilidad de los datos procesados, también se identificarán patrones y tendencias que servirán como insumo para entrenar y evaluar los modelos de machine learning.

## Instrumentos y Herramientas

**Software y Lenguajes de Programación:** Python, XGBRegressor y Scikit-learn para el desarrollo y validación de los modelos de machine learning; librerías de análisis de datos como Pandas, Matplotlib y Numpy para el procesamiento de información.

**Herramientas de Gestión:** Jira y/o Trello para el seguimiento de las tareas y Sprints en el marco SCRUM.

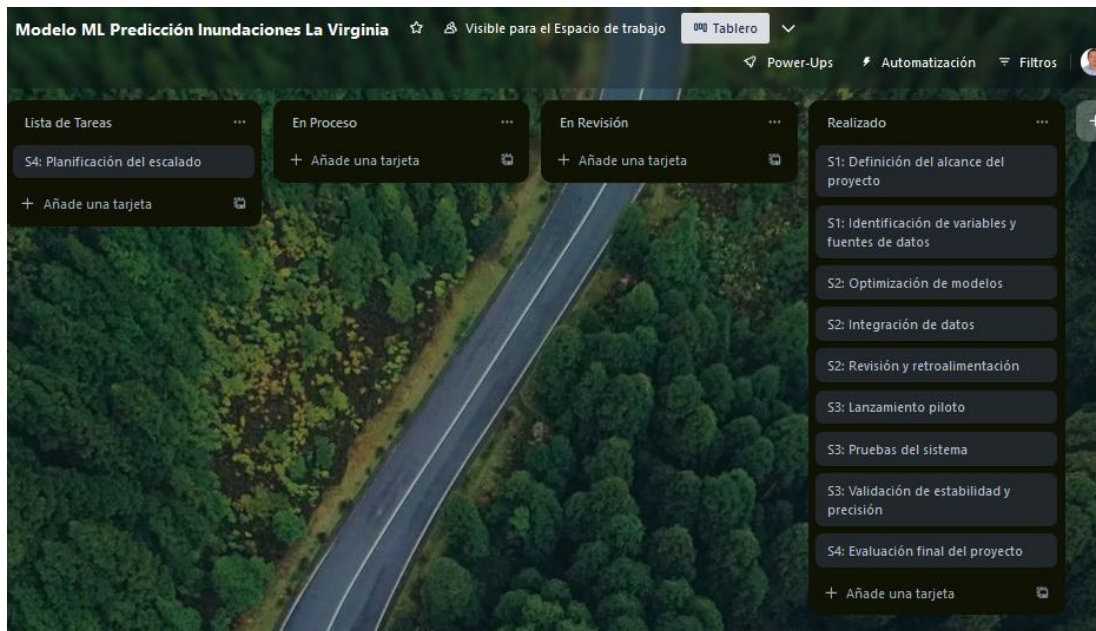
## Figura 1

*Metodología SCRUM Aplicada al Proyecto*



## Figura 2

### Tablero con el Desarrollo del Proyecto



*Nota.* Tomado del sitio web de Trello.

## Hardware

- Sensores hidrológicos y estaciones meteorológicas para capturar datos ambientales.
- Infraestructura de servidores para el almacenamiento y procesamiento de grandes volúmenes de datos.

## Resultados

### Desarrollo de Modelos

Para mejorar la precisión en la predicción de inundaciones en los ríos Cauca y Risaralda en el municipio de La Virginia Risaralda se implementó un modelo de aprendizaje automático basado en *Gradient Boosting (XGBoost)*, este modelo fue seleccionado debido a su capacidad para manejar datos con alta variabilidad y su eficiencia en la predicción de series temporales, además se evaluaron múltiples arquitecturas y parámetros con el objetivo de encontrar el mejor ajuste para la predicción de los niveles de los ríos permitiendo tener una baja tasa de error y una alta capacidad de generalización.

### Limpieza y Exploración de los Datos

#### Figura 3

##### *Dataframes de los Ríos Cauca y Risaralda*

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112709 entries, 0 to 112708
Data columns (total 16 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Fecha                                     112709 non-null object
1   Hora                                     112709 non-null object
2   Temperatura del aire (°C)               112709 non-null float64
3   Punto rocío (°C)                       112709 non-null float64
4   Presión barométrica (mmHg)             112709 non-null float64
5   Humedad relativa (%)                   112709 non-null float64
6   Precipitación acumulada (mm)           112709 non-null float64
7   Precipitación real (mm)                 112709 non-null float64
8   Intensidad de Precipitación (mm/h)     112709 non-null float64
9   Nivel (cms)                             112709 non-null object
10  Voltaje de la Batería (V)               112709 non-null float64
11  Puerta_Abierta                          112709 non-null bool
12  Batería_Baja                             112709 non-null bool
13  Solar_Energia                            112709 non-null object
14  Observaciones                           720 non-null    object
15  Unnamed: 15                             0 non-null     float64
dtypes: bool(2), float64(9), object(5)
memory usage: 12.3+ MB

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67269 entries, 0 to 67268
Data columns (total 16 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   Fecha                                     67269 non-null object
1   Hora                                     67269 non-null object
2   Temperatura del aire (°C)               67269 non-null float64
3   Punto rocío (°C)                       67269 non-null float64
4   Presión barométrica (mmHg)             67269 non-null float64
5   Humedad relativa (%)                   67269 non-null float64
6   Precipitación acumulada (mm)           67269 non-null float64
7   Precipitación real (mm)                 67269 non-null float64
8   Intensidad de Precipitación (mm/h)     67269 non-null float64
9   Nivel (cms)                             67269 non-null object
10  Voltaje de la Batería (V)               67269 non-null float64
11  Puerta_Abierta                          67269 non-null bool
12  Batería_Baja                             67269 non-null bool
13  Solar_Energia                            67269 non-null object
14  Observaciones                           67269 non-null object
15  Unnamed: 15                             0 non-null     float64
dtypes: bool(2), float64(9), object(5)
memory usage: 7.3+ MB

```

De acuerdo con el análisis exploratorio de los datos entregados por el Cuerpo de Bomberos de La Virginia se obtuvieron los siguientes campos:

Fecha: Indica el día en que se realizó el registro, este se encuentra en formato de texto y se espera que contenga la información de la fecha en formato AAAA-MM-DD.

Hora: Especifica el momento del día en que se efectuó la medición, está en formato de texto y junto con la columna “Fecha” permite reconstruir la marca temporal completa de cada registro.

Temperatura del aire (°C): Registra la temperatura del aire en grados Celsius, es un valor numérico de tipo float que ayuda a conocer las condiciones ambientales durante el registro.

Punto rocío (°C): Mide la temperatura del punto de rocío en grados Celsius, este valor es útil para analizar la humedad y la formación de rocío en el ambiente.

Presión barométrica (mmHg): Indica la presión atmosférica expresada en milímetros de mercurio, es un valor numérico que permite evaluar las condiciones climáticas y predecir posibles cambios en el tiempo.

Humedad relativa (%): Representa el porcentaje de humedad en el aire, es un dato de tipo float que Ayuda al análisis de la atmósfera y su influencia en eventos meteorológicos.

Precipitación acumulada (mm): Registra la cantidad total de lluvia acumulada en milímetros durante un período determinado, es de tipo numérico y es importante para el análisis de patrones de lluvia.

Precipitación real (mm): Muestra la cantidad de precipitación real en milímetros, este campo puede ser utilizado para comparar con la precipitación acumulada y evaluar discrepancias o validar la calidad de la medición.

Intensidad de Precipitación (mm/h): Mide la intensidad de la precipitación en milímetros por hora, es de tipo numérico y ayuda a identificar la rapidez con la que se produce la lluvia.

Nivel (cms): Muestra el nivel del agua medido en centímetros, aunque el campo se almacena como objeto se espera que contenga valores numéricos que representan la altura del agua, este campo es importante para poder desarrollar este proyecto.

Voltaje de la Batería (V): Registra el voltaje de la batería del dispositivo en voltios, es de tipo numérico que permite monitorear el estado energético de la del equipo.

Puerta Abierta: Es un indicador booleano que muestra si la puerta del dispositivo se encontraba abierta en el momento del registro.

Batería Baja: Indicador booleano que señala si la batería del dispositivo se encuentra en un estado bajo, este dato es importante para el mantenimiento y/o la fiabilidad de las mediciones.

Solar Energía: Campo de tipo objeto que contiene información relacionada con la energía solar, indicando si el dispositivo está operando con energía solar o si existe alguna condición particular en el suministro de energía.

Observaciones: Campo de texto donde se registran notas o comentarios adicionales, vale la pena aclarar que solo 720 registros contienen información en este campo.

Unnamed: 15: Columna sin nombre y sin registros (0 registros no nulos) por lo que no aporta información relevante al análisis.

Posterior a esto se realiza la limpieza de datos donde se logró procesar, normalizar y limpiar los datos eliminando valores atípicos y valores vacíos, el uso de la codificación correcta (utf-8) y la separación entre campos con punto y coma, todo esto realizado usando el módulo de Pandas en Python, también se realizó la depuración y unificación de los campos fecha y hora, todo estos procesos permitieron identificar patrones estacionales mediante el uso de la prueba Dickey-Fuller Aumentada (ADF) y también diarios aunque no se exploró completamente la correlación de los datos con factores externos como la precipitación que podría mejorar la

interpretación de los resultados y la efectividad del modelo, pero para llegar a aplicarlo se requiere datos de puntos diferentes de sus cauces y en los datos sólo contamos con un punto de monitoreo en cada río ya que se debe tener en cuenta la precipitación en el punto no afecta de manera inmediata el cauce del río.

**Tabla 1**

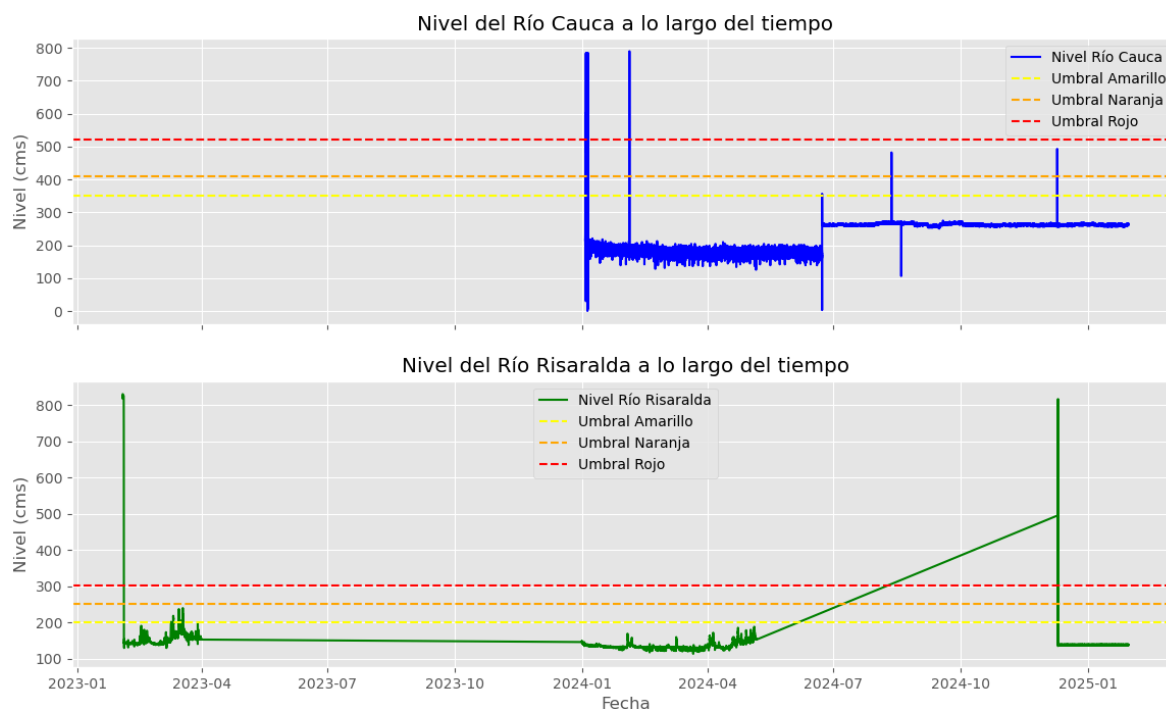
*Resultado de la Limpieza de los Datos*

Columna	Unidad	Risaralda	Cauca	Tipo de dato
Temperatura del aire	°C	67.269	112.709	Float64
Punto rocío	°C	67.269	112.709	Float64
Presión barométrica	mmHg	67.269	112.709	Float64
Humedad relativa	%	67.269	112.709	Float64
Precipitación acumulada	mm	67.269	112.709	Float64
Precipitación real	mm	67.269	112.709	Float64
Intensidad de Precipitación	mm/h	67.269	112.709	Float64
Nivel	cm	67.269	112.709	Float64
Fecha_Hora	-	67.269	112.709	Float64

Luego de la normalización de los datos podemos identificar variables interesantes como la temperatura, presión, humedad que podrían ayudar a mejorar el modelo a implementar, además existe una buena cantidad de registros por cada uno de los ríos, pero hay más registros en el río Cauca que del Risaralda, por lo que es posible que se tengan menos tiempo.

## Figura 4

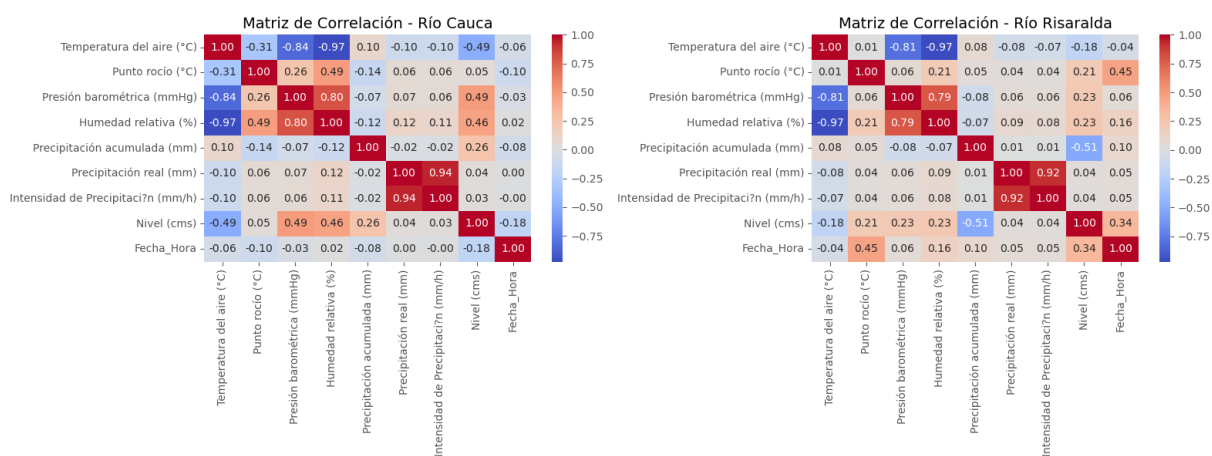
### *Series de Tiempo de los Ríos Cauca y Risaralda*



Según la gráfica hay una serie de evidencias que afectarían el uso de modelos, empezando por los momentos en los que se sobrepasa los umbrales, siendo estos valores atípicos ya que el Cuerpo de Bomberos informó que durante el tiempo que se adquirió la información no se presentaron eventos, además hay valores cercanos a cero, cosa que no podría suceder, también se evidencia que existen espacios de tiempos sin datos, también una variación en la amplitud y ubicación del nivel del río Cauca, también se puede evidenciar en el río Risaralda registros muy iguales salvo en dos espacios de tiempo el Cauca del 23/06/2024 al 30/01/2025 y el Risaralda del 01/01/2024 al 06/05/2024, esto se debe al nuevo del sistema que captura la información, los posibles fallos que se pueden presentar y por los ajustes al mismo, por lo que no es posible usar todos los datos por lo que se va a usar la información más reciente y acertada para ambos ríos.

Figura 5

## Matriz de Correlación



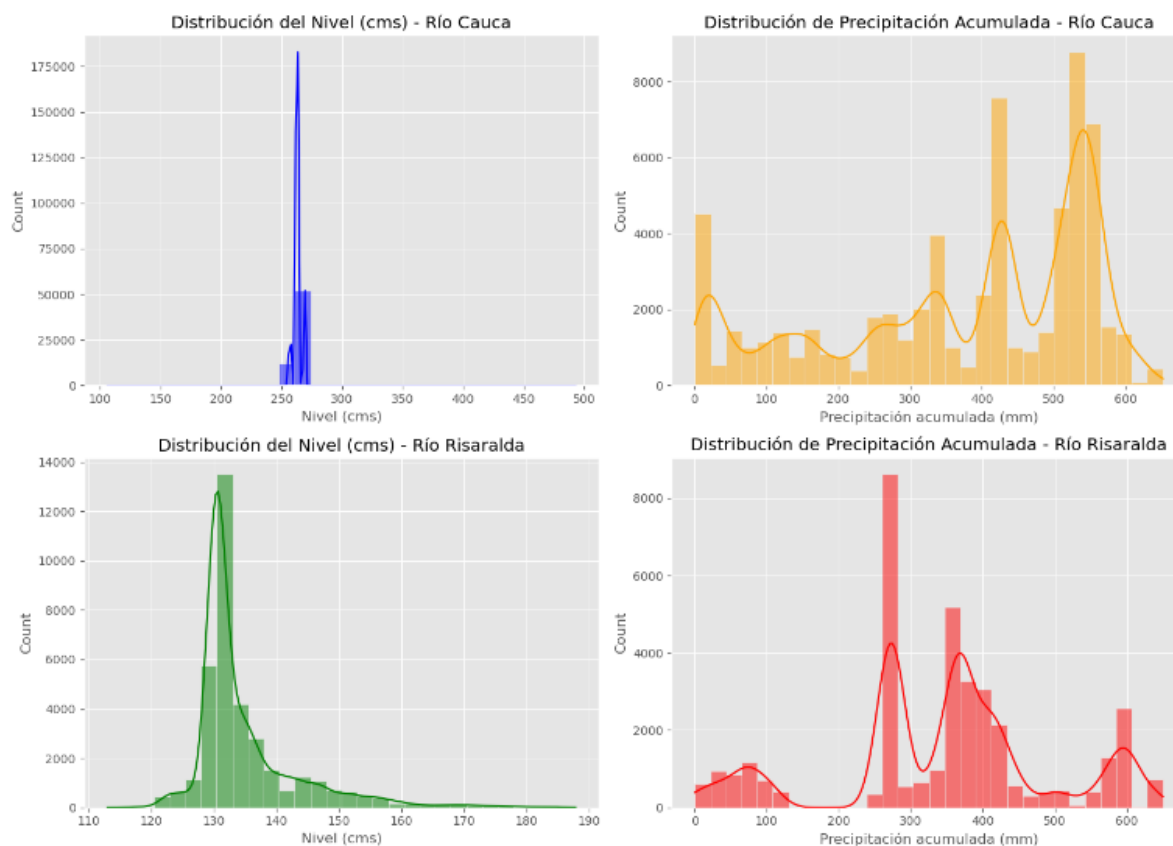
La matriz de correlación del río Cauca muestran un correlación moderada entre la humedad relativa y la presión versus con el nivel del río, mientras que el río Risaralda muestra una baja correlación (entre sus mejores valores) entre el nivel del río y la precipitación acumulada y la intensidad de la precipitación.

En el caso del río Cauca su mayor tamaño y extensión hacen que factores como la humedad relativa y la presión barométrica puedan influir indirectamente en su nivel, posiblemente a través de efectos en la generación de precipitaciones en áreas más amplias de la cuenca, especialmente en el Valle del Cauca, además la humedad puede estar relacionada con la retención de agua en la atmósfera y su posterior liberación en forma de lluvias afectando gradualmente el nivel del río, por otro lado el río Risaralda al ser de menor tamaño y responder de manera más localizada a las condiciones climáticas presenta una correlación más débil entre su nivel y la precipitación, esto puede deberse a la rápida infiltración del agua en el suelo, la existencia de estructuras de regulación o la capacidad de drenaje del terreno lo que impide que

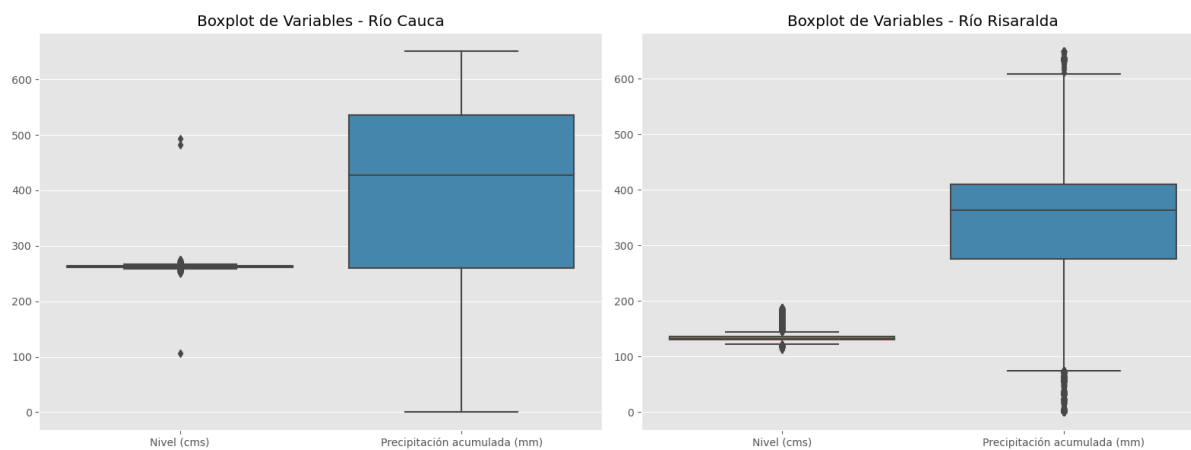
un aumento en la precipitación se refleje directamente en su caudal, por eso la importancia de contar con mediciones en otros puntos de los ríos.

## Figura 6

### *Histogramas de los Ríos Cauca y Risaralda*



Podemos observar que en el río Cauca los niveles están concentrados entre 260 y 270 cm, con algunos picos que podrían ser *outliers*, mientras que en el río Risaralda los niveles se concentran entre 130 y 140 cm, mostrando una menor dispersión comparado con el Cauca, podría suponerse que es por el tamaño y el recorrido de cada uno de ellos, en cuanto a la precipitación acumulada se puede observar que en ambos ríos están distribuidos de manera heterogénea con algunos picos extremos.

**Figura 7***Boxplots de los Ríos Cauca y Risaralda*

En los *boxplots* podemos evidenciar que ambas variables (nivel y precipitación) muestran valores extremos (*outliers*), por lo que se realiza una normalización de los datos reemplazando los outliers por la media de los valores de los datos vecinos, esto con el fin de evitar datos atípicos mediante el uso de otra función.

## Resultados del Modelado

Para este caso se implementó el Gradient Boosting (XGBoost) utilizando como variable a predecir el nivel de los ríos Cauca y Risaralda, luego de realizar un análisis de diferentes combinaciones de hiperparámetros utilizando la Búsqueda en Malla (Grid Search).

### Figura 8

#### *Determinación de los Mejores Hiperparámetros*

```
param_grid = {
    'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7],
    'subsample': [0.8, 1.0], 'colsample_bytree': [0.8, 1.0] }
xgb_model = XGBRegressor(random_state=42)
grid_search = GridSearchCV(estimator=xgb_model, param_grid=param_grid, scoring='r2', cv=5, n_jobs=-1, verbose=1)
```

Se determinó que la mejor combinación de hiperparámetros son:

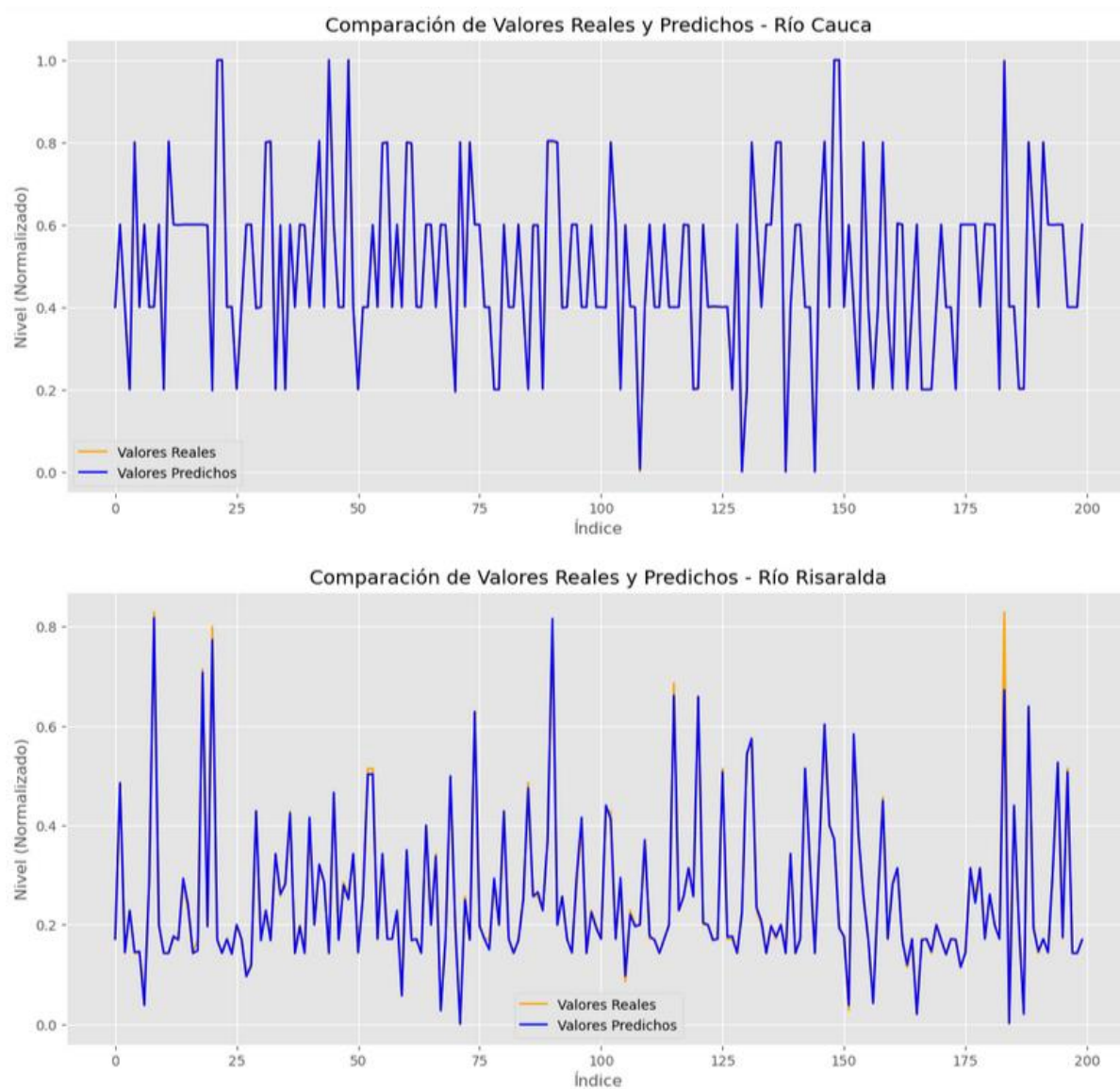
- Número de estimadores: 100
- Tasa de aprendizaje: 0.1
- Profundidad máxima: 5

Y se obtuvieron los siguientes resultados:

**Tabla 2**

#### *Resultados del Modelado en XGBoost de los Ríos Cauca y Risaralda*

Rio	Cauca	Risaralda
MAE (Entrenamiento)	0.00056	0.00322
MAE (Prueba)	0.00060	0.00337
RMSE (Entrenamiento)	0.00138	0.00597
RMSE (Prueba)	0.00286	0.00665
R <sup>2</sup> (Entrenamiento)	0.99996	0.99884
R <sup>2</sup> (Prueba)	0.99998	0.99853
Validación Cruzada	0.99976	0.99595

**Figura 9***Comparación Entre Valores Reales y Predichos*

*Nota.* Se muestran los primeros 200 valores, en cuanto al nivel del río, se ha escalado (normalizado) con valores entre “0” y “1” esto con el fin de que el modelo funcione mejor y pueda comparar con mayor facilidad los valores reales y los predichos.

Como se observa en la Figura 9, los valores predichos presentan un alineamiento casi perfecto con los valores reales confirmando la capacidad del modelo para generar predicciones de mucha confianza, por lo que se considera que no es necesario realizar modificaciones a los parámetros del modelo ni tampoco utilizar otro como ARIMA.

**Tabla 3**

*Resultados del Test de ADF de los Ríos Cauca y Risaralda*

Métrica	Río Cauca	Río Risaralda
Estadístico ADF	-15.8443	-4.9832
P-Valor	9.6117e-29	2.39575e-05
Valor Crítico (1%)	-3.4305	-3.4305
Valor Crítico (5%)	-2.8616	-2.8616
Valor Crítico (10%)	-2.5668	-2.5668
¿Es Estacionaria?	Si	Si

De acuerdo con los resultados del test ADF donde se observa que tanto el estadístico como el p-valor son muy bajos se confirma que las series de tiempo de ambos ríos son estacionarias, aunque la estacionariedad es muy importante para ciertos modelos tradicionales de series de tiempo, en el caso de XGBoost esta característica ayuda a la ingeniería de características y el modelado respaldando su uso en este contexto.

### **Evaluación del Desempeño**

Este modelo cuenta con una alta precisión en donde el  $R^2$  es superior al 99% lo que demuestra que los modelos explican casi toda la variabilidad de los niveles de los ríos, además cuenta con un error absoluto y un error cuadrático medio muy bajas mostrando así un ajuste excelente a los datos, si es importante tener en cuenta el estar realizando un análisis continuo de los nuevos datos para poder ajustar el modelo y evitar que pierda su precisión.

## Conclusiones

El análisis exploratorio y la limpieza de los datos permitieron identificar patrones en las variables que afectan los niveles de los ríos, además la eliminación de valores atípicos y la creación de características temporales, como rezagos y medias móviles mejoraron la calidad de los datos utilizados en el modelado lo que ayudó bastante en la detección de tendencias y la evaluación de la estacionalidad mediante la prueba de Dickey-Fuller Aumentada.

Los modelos de aprendizaje automático desarrollados para los ríos Cauca y Risaralda mostraron un desempeño casi perfecto con coeficientes de determinación superiores al 99%, aunque estos resultados podrían generar dudas acerca de un posible sobreajuste se realizaron validación cruzada y análisis de error en los datos de prueba en los cuales se comprobó que los errores son realmente bajos tanto en entrenamiento como en prueba, además la evaluación con distintos hiperparámetros permitió seleccionar la mejor configuración lo que apoya la alta precisión en la predicción de los niveles de los ríos y su aplicabilidad en escenarios reales.

La validación del desempeño de los modelos confirmó su efectividad en la predicción de niveles de los ríos con resultados consistentes tanto en pruebas individuales como en validaciones cruzadas, esto permite demostrar la capacidad de los modelos para generalizar en diferentes contextos y garantizar predicciones muy confiables, además los gráficos comparativos entre valores reales y los valores predichos comprobaron la alta calidad del ajuste y la capacidad de los modelos para capturar patrones complejos sin un sobreajuste.

Es importante resaltar que este proyecto responde a la necesidad urgente de mitigar el riesgo de inundaciones en el municipio de La Virginia, una región que históricamente ha sido vulnerable a este tipo de eventos, en donde los resultados demostraron la capacidad del modelo para identificar tendencias y patrones estacionales que fundamentan la predicción precisa de los

niveles de los ríos, esto se refleja en los altos coeficientes de determinación y la validación cruzada exitosa lo que evidencia que a pesar de utilizar datos externos en la fase actual, el enfoque es robusto y adaptable para incorporar en el futuro información propia de las estaciones, en este sentido el trabajo permitirá dar continuidad a un sistema de gestión integral y proactivo que permita mitigar el riesgo de inundaciones, no sólo de este municipio sino de muchos que en la actualidad sufren por no contar con herramientas afectando así a muchas comunidades.

## Recomendaciones

Si bien para el desarrollo de este proyecto se utilizaron datos externos suministrados por el Cuerpo de Bomberos a través de una implementación reciente, es importante el que este tipo de modelos puedan contar con fuentes de datos propias a través de Sistemas de Alertas Tempranas provistas de variables que permitan una robustez en la detección en tiempo real y preciso con el fin de predecir de manera efectiva una inundación con suficiente antelación para poder tomar las medidas preventivas necesarias en aras de ayudar a las comunidades.

Si bien el modelo arrojó un resultado sobresaliente es importante poder estar entrenando el modelo de manera continua ya que estos pueden verse afectados por factores externos, además de la poca información con el que se entrenó el modelo actual ya que con mayores periodos de tiempo ayudaría a capturar patrones estacionales más completos.

Sería interesante contar con más información de los ríos en otros puntos ya que esto permite contar con más variables y poder mejorar la capacidad predictiva de los modelos, en lo posible contar con otras variables como velocidad y dirección del viento.

Realizar capacitaciones y simulaciones para los responsables de la gestión del riesgo en el municipio usando los resultados obtenidos para fortalecer la confianza y que a partir de allí se puedan tomar decisiones.

### Referencias Bibliográficas

- Arreguín-Cortés, F. I., López-Pérez, M., & Marengo-Mogollón, H. (2016). *Las inundaciones en un marco de incertidumbre climática*. Tecnología y Ciencias del Agua. <https://www.scielo.org.mx/pdf/tca/v7n5/2007-2422-tca-7-05-00005.pdf>
- Buitrago, L. (2015). *Sistema de Alerta Temprana—SAT: Protocolo de Comunicación y Actuación*. FAPI. <https://www.fapi.org.py/wp-content/uploads/2017/02/Protocolo-SAT.pdf>
- Congreso de Colombia. (2012). Ley 1523 de 2012. *Por la cual se adopta la política nacional de gestión del riesgo de desastres y se establece el sistema nacional de gestión del riesgo de desastres y se dictan otras disposiciones*. Diario Oficial No. 48.411. Recuperado de <https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=47141>
- Di Baldassarre, G., Schumann, G., & Bates, P. D. (2009). *A technique for the calibration of hydraulic models using uncertain satellite observations of flood extent*. Journal of Hydrology, 367(1-2), 276-282. <https://doi.org/10.1016/j.jhydrol.2009.01.020>
- Domínguez-Calle, E., & Lozano-Báez, S. (2014). Estado del arte de los sistemas de alerta temprana en Colombia. *Revista De La Academia Colombiana De Ciencias Exactas, Físicas Y Naturales*, 38(148), 321–333. <https://doi.org/10.18257/raccefyn.132>
- Londoño Giraldo, L. A., & Ocampo Molano, L. M. (2017). *Análisis de Vulnerabilidad Social con Respecto a la Amenaza de Inundación en el Municipio de La Virginia Risaralda*. <https://repositorio.utp.edu.co/server/api/core/bitstreams/5a83e662-c27c-4317-a9a3-e59f2ae61506/content>
- Naciones Unidas (2022). Objetivos de Desarrollo Sostenible. <https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>

- Salinas Díaz, A. M., & Patiño Rodríguez, J. J. (2024). *Caracterización de la Vulnerabilidad Social por Amenaza de Inundación en el Barrio San Carlos Ubicado en el Municipio de la Virginia, Risaralda*. Universidad Católica de Manizales.  
[https://repositorio.ucm.edu.co/bitstream/10839/4400/1/SalinasD%C3%ADazAndr%C3%A9sMauricio\\_2024\\_EPRAD.TG.pdf](https://repositorio.ucm.edu.co/bitstream/10839/4400/1/SalinasD%C3%ADazAndr%C3%A9sMauricio_2024_EPRAD.TG.pdf)
- Sedano, K., Carvajal, Y., & Ávila, Á. (2013). *Análisis de aspectos que incrementan el riesgo de inundaciones en Colombia*. Luna Azul. <https://doi.org/10.17151/luaz.2013.37.13>
- Smith, K., & Ward, R. (1998). *Floods: Physical Processes and Human Impacts*. John Wiley & Sons. [https://doi.org/10.1002/\(SICI\)1096-9837\(199912\)24:13%3C1261::AID-ESP22%3E3.0.CO;2-%23](https://doi.org/10.1002/(SICI)1096-9837(199912)24:13%3C1261::AID-ESP22%3E3.0.CO;2-%23)
- Sutherland, J., & Schwaber, K. (2013). *The Scrum Guide*. Scrum.org. <https://www.scrum.org/>
- UNGRD (2022). *Inundaciones. Sistema Nacional de Información para la Gestión del Riesgo de Desastres*. <https://www.gestiondelriesgo.gov.co/snigrd/pagina.aspx?id=144>
- Urrutia, M. (2012). *Colombia: Manual Comercial e Industrial*. Banco de la República.  
[https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/9310/LBR\\_2012-4.pdf](https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/9310/LBR_2012-4.pdf)
- Zhou, L., & Kang, L. (2023). *A comparative analysis of multiple machine learning methods for flood routing in the Yangtze River*. *Water*, 15(1556). <https://doi.org/10.3390/w15081556>

## Apéndices

### Apéndice A

#### *Autorización de Uso de Datos de los Afluentes Cauca y Risaralda*



## **CUERPO DE BOMBEROS VOLUNTARIOS LA VIRGINIA-RISARALDA**

NIT. 891.408.932 – 5  
Personería jurídica Res. 465 del 3 de DIC/1963  
Cra 8a # 10 – 50 tel. 3682090 cel.311 354 82 81 alarma 119  
E-mail: bomberosvirginia2021@gmail.com



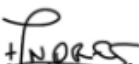
### **CUERPO DE BOMBEROS VOLUNTARIOS DE LA VIRGINIA**

**NIT: 891.408.932-5**

#### **AUTORIZA A:**

CARLOS HERNAN BERMÚDEZ BETANCUR identificado con la C.C. 18.609.846 de La Virginia Risaralda, estudiante del programa académico de Especialización en Ciencia de Datos y Analítica de la Universidad Nacional Abierta y a Distancia - UNAD para que pueda obtener y utilizar la información requerida en cuanto a los datos obtenidos de los puntos de los ríos Cauca y Risaralda para el desarrollo del proyecto de grado.

Se firma en La Virginia Risaralda el 30 de enero de 2025.

  
Elkin Andrés Ramírez Herrera  
Jefe operativo  
Cuerpo de bomberos voluntarios la virginia Rda.  
Cel 3207225436

**VALOR, ABNEGACION Y DISCIPLINA**