

# Proyecto Integrador 2

## Predicción de eventos de poda en redes de distribución de energía

– Maestría en Ciencias de los Datos y Analítica –

Desarrollado por

Diana Lisette Arango Cañas  
Andrés Alonso Ocampo Dávila  
Juan Felipe Renza Chavarría  
Daniel Alonso Sánchez Echeverri

7 de junio, 2023

Tutores

- 1: Jose Lisandro Aguilar Castro
- 2: Paula Maria Almonacid Hurtado
- 3: Edison Valencia Díaz

# Índice

<b>1. Introducción</b>	<b>1</b>
<b>2. Marco teórico</b>	<b>2</b>
2.1. Series de Tiempo . . . . .	2
2.2. Aproximaciones por <i>machine learning</i> a las series de tiempo . . . . .	4
<b>3. Desarrollo metodológico</b>	<b>4</b>
3.1. Entendimiento del negocio . . . . .	4
3.2. Entendimiento de los datos . . . . .	5
3.3. Preparación de los datos . . . . .	7
3.4. Modelado . . . . .	17
3.5. Evaluación e interpretación . . . . .	18
3.6. Despliegue . . . . .	28
<b>4. Conclusiones</b>	<b>28</b>
<b>5. Anexos</b>	<b>29</b>
<b>Referencias</b>	<b>30</b>

# 1. Introducción

EPM es una empresa de servicios públicos domiciliarios de propiedad del Municipio de Medellín, que presta los servicios de energía eléctrica, gas por red, agua y saneamiento, en 123 municipios de Antioquia.

Como parte de los procesos de negocio, se incluyen las actividades de operación y mantenimiento de la infraestructura requerida para la prestación del servicio de distribución de energía en los diferentes niveles de tensión.

Una de las actividades primordiales es la poda de árboles juega un papel crucial para mantener la confiabilidad y la calidad del servicio en los sistemas de distribución de energía. La poda oportuna de árboles cerca de las redes de distribución es esencial para prevenir interrupciones y daños en el suministro eléctrico. En este contexto, surge la necesidad de desarrollar métodos precisos y eficientes para predecir la cantidad de trabajos de poda que se requerirán en un determinado período de tiempo, con el fin de identificar el recurso necesario para atender dichos trabajos y lograr un direccionamiento óptimo de los recursos disponibles.

El presente proyecto se centra en abordar esta necesidad a través de la aplicación de técnicas de predicción, utilizando datos históricos de trabajos de poda, información sobre precipitaciones en la zona y una variable que indica la disponibilidad del sistema. Se explorarán dos enfoques principales: uno basado en series de tiempo clásicas y otro utilizando métodos de machine learning aplicados a series de tiempo.

El primer enfoque, basado en series de tiempo clásicas, empleará técnicas tradicionales de análisis de series temporales, como ARIMA (Autoregressive Integrated Moving Average), Auto ARIMA y ARIMAX (ARIMA con variables exógenas). Se realizará un análisis exhaustivo de los datos históricos, identificando patrones estacionales, tendencias y posibles factores de influencia, como las precipitaciones y la disponibilidad del sistema. Estas técnicas permitirán construir un modelo que capture la dinámica de los trabajos de poda en función de las variables relevantes y realice predicciones precisas.

El segundo enfoque se basará en técnicas de machine learning aplicadas a series de tiempo. Aquí, se emplearán modelos como regresiones lineales, Random Forest, XGBoost y redes neuronales, que son capaces de capturar relaciones complejas y no lineales presentes en los datos de series temporales. Se utilizarán técnicas de preprocesamiento de datos, selección de características y entrenamiento de modelos

con el fin de obtener predicciones más precisas y ajustadas a las características específicas de los trabajos de poda en redes de distribución de energía eléctrica.

La combinación de ambos enfoques permitirá comparar y evaluar la efectividad de las técnicas clásicas y de machine learning en la predicción de la cantidad de trabajos de poda. Los resultados obtenidos en este proyecto tendrán aplicaciones prácticas en la gestión y planificación de los recursos necesarios para llevar a cabo la poda de árboles, buscando un direccionamiento óptimo de los recursos disponibles. Esto contribuirá a mejorar la eficiencia y la calidad del servicio en las redes de distribución de energía eléctrica, al permitir una asignación adecuada de los recursos y una respuesta más rápida y efectiva ante las necesidades de poda.

## 2. Marco teórico

### 2.1. Series de Tiempo

Una serie de tiempo es una secuencia de observaciones realizadas a una variable a lo largo del tiempo en intervalos constantes que pueden ser horarios, diarios, semanales, mensuales, anuales, etc. Estas observaciones se registran en momentos específicos y se presentan en orden cronológico. Además, estas observaciones pueden ser realizadas sobre una característica (univariante) o sobre varias características (multivariante).

Las series de tiempo presentan ciertas propiedades que influyen en los modelos utilizados para su análisis y predicción de valores futuros. Estas propiedades ayudan a caracterizar la serie de tiempo y a comprender su comportamiento. Algunas de estas características son:

- **Estacionaridad:** una serie de tiempo se considera estacionaria cuando se tiene una estabilidad a lo largo del tiempo de la media, la varianza y la estructura de covarianzas[1]. Con el objetivo de determinar si una serie de tiempo es estacionaria se usan típicamente dos pruebas, ADF (Augmented Dickey-Fuller Test) y KPSS (Kwiatkowski-Phillips-Schmidt-Shin).
- **Tendencia:** es la dirección a la que se mueven los datos de la serie de tiempo en el largo plazo. Puede ser creciente o decreciente.
- **Estacionalidad:** una serie de tiempo se considera estacional si presenta un comportamiento superpuesto que se repite a lo largo del tiempo.[1]

- **Autocorrelación:** Se define como una medida de la interdependencia lineal entre datos de una muestra que están separados por un número  $k$  de unidades de tiempo[2]. La autocorrelación proporciona información del número máximo de observaciones en el pasado (pasos atrás) que tienen influencia en la predicción del dato futuro.

Entender el comportamiento de la serie de tiempo y descubrir sus propiedades permiten aplicar los modelos que llevarán a resultados más precisos en la predicción del dato futuro. A continuación se incluye el detalle del modelo ARIMA, típicamente usado en la predicción de series de tiempo sin tendencia ni estacionalidad.

**ARIMA (Autoregressive Integrated Moving Average):** El modelo está compuesto por 3 características, primero se tiene el componente AR, el cual se refiere a la autocorrelación de los datos, lo que implica identificar cuantas observaciones del pasado son útiles para pronosticar el dato futuro. El componente I, se refiere a la diferenciación necesario en la serie de tiempo para suavizar las características de tendencia y estacionalidad que inducen ruido en la predicción. Finalmente, el componente MA, indica como el valor futuro puede predecirse a partir de una componente aleatoria de este momento y las componentes aleatorias anteriores. Es decir, como los errores residuales de la media móvil en datos pasados, influye en el dato futuro.

Cada una de estas tres características deben ser parametrizadas al momento de aplicar el modelo, pero encontrar estos valores puede llegar a ser dispendioso por la cantidad de combinaciones posibles. Para tratar con esta dificultad, se tiene el algoritmo AutoARIMA que realiza el barrido de las múltiples combinaciones de parámetros y calcula el ajuste del modelo a través diferentes métricas. Esto permite obtener el modelo parametrizado de tal forma que la predicción del valor futuro se ajusta adecuadamente a los datos reales.

El modelo ARIMA tiene la ventaja de combinar tres tipos de modelado en un solo, esto lo hace muy flexible y logra ajustarse bien a los datos [2]. Existen variaciones de este modelo que permiten incorporar variables independientes a las observaciones de la serie de tiempo, las cuales contienen información que puede influir en su comportamiento. Este modelo se conoce como ARIMAX y se utiliza ampliamente con el objetivo de mejorar las predicciones, ya que incrementa la información que caracteriza la serie de tiempo.

## 2.2. Aproximaciones por *machine learning* a las series de tiempo

Las aproximaciones tradicionales a las series de tiempo presentan tanto ventajas como desafíos, como se ha mencionado anteriormente. No obstante, el avance de la capacidad computacional ha propiciado la adopción de enfoques más sólidos, que posibilitan un análisis más riguroso de las series de tiempo.

Mediante el uso de técnicas de aprendizaje automático (Machine Learning), se han logrado importantes avances en el estudio de las series de tiempo. Por ejemplo, en el análisis de los mercados financieros, los algoritmos de aprendizaje automático han sido fundamentales para el análisis de las series de tiempo relacionadas con las fluctuaciones de las acciones. En este contexto, las redes neuronales han adquirido una relevancia significativa, gracias a los avances actuales en capacidad computacional que han permitido su implementación constante y precisa[3].

Existen otros contextos en dónde las series de tiempo también son útiles; por ejemplo, el control de los parámetros de plantas termodinámicas. En la literatura podemos ver el uso de técnicas de ensamble como XGBoost en dónde su aplicación permiten la generalización del análisis con unos costos computacionales más razonables, pues un inconveniente relevante de las redes neuronales suele ser que sus procesos de hiperparametrización y costo computacional suelen ser más altos[4].

Finalmente, podemos decir que las aproximaciones robustas basadas en Machine Learning han revolucionado el análisis de series de tiempo al proporcionar modelos más flexibles y sofisticados. Estas técnicas permiten una mejor captura de patrones complejos en los datos y ofrecen la posibilidad de realizar predicciones más precisas en diversos campos de aplicación.

## 3. Desarrollo metodológico

### 3.1. Entendimiento del negocio

Este proyecto integrador se desarrolla en Empresas Públicas de Medellín (EPM) con foco en el negocio de distribución energía, el cual es caracterizado por su alto dinamismo ya que está enmarcado en la atención de todo el departamento de Antioquia, dividido administrativamente en 8 regiones, cerca de 1050 circuitos, que a su vez están compuestos por más de 98.000 kilómetros de red. Esta cantidad de

redes de energía en su gran mayoría utiliza el tipo de construcción aéreo debido a su flexibilidad para llegar a zonas rurales y a sus bajos costos [5].

Este tipo de montaje tiene implícito algunos riesgos, ya que las redes de energía están expuestas a factores externos que inciden en su desempeño y pueden ocasionar fallas en la prestación del servicio, como el choque de vehículos con los postes de energía, la contaminación por suciedad (partículas de polvo u otros agentes), el contacto de árboles, animales, cometas u otros objetos con los cables de energía, etc. Lo que afecta la confiabilidad del sistema.

Garantizar estándares adecuados en la calidad del servicio prestado al usuario es una exigencia en la que está inmersa la empresa a través de unos indicadores y metas impuestos por el ente regulador gubernamental. Por lo tanto, es fundamental dirigir esfuerzos hacia el cumplimiento de estas metas y fortalecer constantemente la confiabilidad del sistema.

Para aportar a la confiabilidad del sistema de distribución se tiene el proceso de mantenimiento, el cual a través de diversas actividades busca evitar las fallas en los equipos que componen el sistema. Debido a la exposición que tienen las redes de energía y a la abundante vegetación que caracteriza al departamento de Antioquia, una de las actividades más relevantes de mantenimiento es podar los árboles que están cerca a los cables de energía, garantizando una distancia de seguridad que permita el crecimiento orgánico de estos individuos arbóreos sin generar contacto con las redes y por lo tanto una falla para el sistema. Destinar adecuadamente los recursos finitos con los que se cuenta para atender las necesidades de poda de los 98.000 kilómetros de red que tiene EPM no es una tarea trivial y su correcta asignación aporta a evitar fallas en el sistema y por lo tanto el cumplimiento de calidad de servicio a los usuarios de la empresa.

### **3.2. Entendimiento de los datos**

Para el desarrollo de este proyecto integrador se utilizarán como insumo principal las órdenes de trabajo registradas en el sistema de información EAM MAXIMO, el cual soporta la gestión de mantenimiento de los activos de las redes de distribución de energía.

Cada registro u orden de trabajo se origina al ocurrir un evento en la red de distribución que requiere de la atención de las cuadrillas de mantenimiento. Para el caso específico de este proyecto los eventos analizados se relacionan con trabajos de poda de especies vegetales que se encuentran cercanas a las redes y su infraestructura. Cada vez que es necesario realizar una actividad de poda de uno o varios árboles cercanos a la red, ya sea porque han generado una afectación del servicio o porque existe un riesgo inminente de que puedan llegar a generarla, se crea un registro en el sistema de información. Luego de que las actividades han finalizado, estas se documentan de forma detallada en la orden de trabajo y quedan almacenadas en la base de datos transaccional (RDBMS Oracle) del sistema de información.

Cada orden de trabajo, representada por un identificador único, contiene, entre otros datos, la fecha de ocurrencia del evento, la identificación del elemento o sección de la red intervenida, su ubicación geográfica y la parte del sistema de distribución la que pertenece (circuito) y el grupo responsable de la atención.

Figura 1: Elementos de datos en las órdenes de trabajo

Nombre de la columna	Descripción de la variable	Tipo de dato
<b>ID</b>	Identificador de la orden de trabajo	NUMBER
<b>FECHA</b>	Fecha de creación de la orden de trabajo	DATE
<b>PENDIENTE</b>	Actividad requerida (para la cual se genera la orden de trabajo)	VARCHAR
<b>ESTADO</b>	Estado de la orden de trabajo	VARCHAR
<b>CIRCUITO</b>	Circuito (sistema) sobre el cual se ejecuta el trabajo	VARCHAR
<b>ELEMENTO</b>	Subsistema sobre el cual se ejecuta el trabajo	VARCHAR
<b>NODO</b>	Componente (activo) sobre el cual se ejecuta el trabajo	VARCHAR
<b>CRITICIDAD_NUM</b>	Criticidad (prioridad) del circuito (sistema)	NUMBER
<b>COORD_LONG</b>	Coordenada (longitud) donde se ejecuta el trabajo	NUMBER
<b>COORD_LAT</b>	Coordenada (latitud) donde se ejecuta el trabajo	NUMBER
<b>REGION</b>	Región donde se ejecuta el trabajo	VARCHAR
<b>SUBREGION</b>	Subregión donde se ejecuta el trabajo	VARCHAR
<b>MUNICIPIO</b>	Municipio donde se ejecuta el trabajo	VARCHAR
<b>GRUPO_RESPONSABLE</b>	Grupo responsable de ejecutar el trabajo	VARCHAR



De esta forma, se cuenta con un registro histórico de los trabajos de poda realizados como parte del mantenimiento del sistema de distribución de energía: cuántos se han ejecutado, con qué frecuencia y en qué partes de la red.

Específicamente, para el alcance definido para el presente proyecto, se analizarán los datos correspondientes a trabajos de poda ejecutados en la región suroeste de Antioquia, en el periodo comprendido entre septiembre de 2020 y abril de 2023, lo que hace un total de casi 4200 registros.

Bajo la hipótesis de que la cantidad de trabajos de poda que se deben atender en un periodo dado se ve determinada en parte por las condiciones meteorológicas de la zona donde se ubican las redes, específicamente por la ocurrencia de lluvias de alta intensidad, se incorporarán en el análisis los datos correspondientes a las precipitaciones registradas en cuatro estaciones de monitoreo localizadas en el suroeste antioqueño. Estos datos, abiertos, son publicados por el IDEAM en el sitio [Consulta y Descarga de Datos Hidrometeorológicos](#) e incluyen entre otras las siguientes variables:

- Código de identificación de la estación de monitoreo
- Nombre de la estación de monitoreo
- Ubicación de la estación de monitoreo - departamento, municipio, latitud, longitud y altitud.
- Fecha del dato
- Valor de la precipitación total diaria (mm)

### **3.3. Preparación de los datos**

La preparación de los datos para este proyecto, se basa en la aplicación de los siguientes pasos, que parten desde la extracción de los datos de los sistemas transaccionales de EPM hasta llegar a una representación útil para el modelado.

#### **Extracción e integración de los datos**

Mediante consultas SQL a la base de datos Oracle del sistema transaccional EAM MAXIMO, se extraen los datos de las ordenes de trabajo de poda, los cuales incluyen,

como se mencionó anteriormente, la fecha de la actividad ejecutada, la identificación del sistema, subsistema y elemento al que se asocia la actividad así como los atributos que definen su ubicación geográfica. Estos datos se convierten en archivos planos (formato *.csv*) que se almacenan en un repositorio en la nube, para posteriormente ser procesados.

Para obtener el registro histórico de las precipitaciones ocurridas en el suroeste antioqueño, se realiza la consulta al sitio del IDEAM [Consulta y Descarga de Datos Hidrometeorológicos](#). De allí se descargan los datos para las siguientes estaciones de monitoreo ubicadas en la zona:

- BOLOMBOLO [26200130]
- FREDONIA [26200150]
- SANTA BARBARA [26190100]
- URRAO [11075020]

De esta manera se obtiene un archivo plano (formato *.csv*) que incluye principalmente la identificación de la estación que registro el dato, la fecha del registro y el valor de precipitación diaria. Estos datos también se almacenan en un repositorio en la nube para su procesamiento.

En etapas posteriores de la preparación de datos, la información procedentes de las dos fuentes consultadas se integrará y procesará haciendo uso de *scripts* de *Python*.

## Selección de las variables

Dado que la solución propuesta se enfoca en analizar los datos de los trabajos de poda como una serie de tiempo, de los conjunto de datos disponibles, ordenes de trabajo de poda y precipitaciones, se extraen en este punto las variables que permitirán, luego de ser transformadas, obtener la serie de tiempo sobre la cual se aplicarán los modelos.

Para el caso de las órdenes de trabajo se extrae la fecha de creación de la orden de trabajo (fecha de la actividad) y el identificador único de la orden de trabajo. Por su parte, para el registro histórico de precipitaciones, se crean cuatro subconjuntos de datos, uno para cada estación de monitoreo, con la fecha del registro y el valor de precipitación diaria como variables.

## **Limpieza de los datos**

En primer lugar se seleccionan (filtran) del conjunto de datos de las órdenes de trabajo aquellos registros correspondientes a actividades de mantenimiento relacionadas con trabajos de poda, ya sean de tipo preventivo o correctivo.

Teniendo en cuenta el contexto de negocio, se parte de la hipótesis de que la cantidad de eventos correspondientes a trabajos de poda depende en gran medida de la región geográfica: clima, tipo de vegetación, geografía, precipitaciones. Para realizar el análisis se elige inicialmente la región con una mayor cantidad de eventos, con el fin de disponer de una cantidad mayor de datos. Esta es la región Suroeste de Antioquia.

Por otra parte, teniendo en cuenta que en la etapa inicial de estabilización del sistema de información se produjo un subregistro de los trabajos de mantenimiento (incluyendo los trabajos de poda), se eliminarán los registros correspondientes a este periodo de estabilización con el fin de evitar sesgos en las estimaciones. Así, se determina que el periodo de análisis de los datos se iniciará en el mes de septiembre de 2020 y se extenderá hasta abril de 2023.

Además se eliminarán del conjunto de datos aquellos registros que tengan valores nulos correspondientes a la cantidad de trabajos de poda y a las precipitaciones en cualquiera de las estaciones de monitoreo seleccionadas, puesto que estas variables son las que se analizarán posteriormente y se ha determinado que los registros sin esta información no resultarán relevantes y que de hecho pueden llegar a generar sesgos en las predicciones generadas por los modelos.

## **Transformación de los datos**

Con el fin de hacer posible el modelado de los datos mediante el uso de series de tiempo, estos deben ser organizados de manera que se tenga una frecuencia determinada para el conteo de eventos (trabajos de poda). Para el caso de análisis se determina que la “frecuencia de muestreo” será semanal.

Así, se extrae la serie semanal, haciendo el conteo de eventos de trabajos de poda generados por cada semana en el periodo de análisis. Se obtiene de esta forma un conjunto de datos con 137 observaciones.

Figura 2: Serie de tiempo semanal trabajos de poda

Semana	Cantidad de trabajos de poda
2020-08-31/2020-09-06	28
2020-09-07/2020-09-13	23
2020-09-14/2020-09-20	30
2020-09-21/2020-09-27	30
2020-09-28/2020-10-04	23
...	...
2023-03-13/2023-03-19	34
2023-03-20/2023-03-26	30
2023-03-27/2023-04-02	35
2023-04-03/2023-04-09	17
2023-04-10/2023-04-16	31

Una transformación adicional de los datos permite mantener solamente la fecha inicial del rango definido (semana) como índice de la serie de tiempo, para facilitar el procesamiento de los datos.

Una transformación similar se realiza sobre los datos de precipitaciones registradas para las cuatro estaciones de monitoreo seleccionadas, de forma que se obtiene un conjunto de datos conformado por cuatro variables, cada una correspondiendo al valor total de precipitaciones acumuladas por semana registrado por una estación específica.

Figura 3: Serie de tiempo semanal precipitaciones

Semana	Estación Bolombolo	Estación Sta.Barbara	Estación Fredonia	Estación Urrao
2020-08-31	43.3	44.5	73.0	63.6
2020-09-07	34.3	24.6	124.0	16.8
2020-09-14	37.8	44.4	77.0	55.9
2020-09-21	71.7	39.9	67.0	12.7
2020-09-28	78.3	47.0	130.0	64.1
...	...	...	...	...
2022-12-19	0.0	9.4	24.0	28.0
2022-12-26	8.6	40.4	49.0	28.6
2023-01-02	21.8	89.6	132.0	31.0
2023-01-09	19.4	53.2	80.0	28.1
2023-01-16	12.9	13.0	15.0	6.3

Por otra parte, como se mostrará más adelante, en la serie de tiempo de los trabajos de poda se observa un componente de estacionalidad, con periodos cercanos a los tres meses. Con el fin de incorporar este comportamiento en el modelado, se extrae, a partir de la fecha en la serie de tiempo, el número del trimestre (1 al 4) y se agrega como variable predictora.

Finalmente, se han identificado situaciones donde se presenta una indisponibilidad del sistema de información donde se crean registran los trabajos, lo cual genera un comportamiento atípico en la serie de tiempo. Con el fin de evitar sesgos en las estimaciones, se crea una variable *dummy* que refleje la condición de indisponibilidad del sistema, al “señalar” los periodos donde han ocurrido este tipo de situaciones.

Ahora el conjunto de datos se compone de:

- la cantidad de trabajos de poda acumulados por semana
- las precipitaciones registradas por cada una de las estaciones de monitoreo en la zona de interés
- la condición de indisponibilidad del sistema de información donde se registran los eventos
- el trimestre correspondiente a la fecha de los registros

Figura 4: Conjunto de datos de la serie de tiempo con variables exógenas

Semana	Cantidad de trabajos de poda	Disp. del sistema de información	Estación Bolombolo	Estación Santa Barbara	Estación Fredonia	Estación Urrao	Trimestre_1	Trimestre_2	Trimestre_3	Trimestre_4
2020/08/31	28	0	43.3	44.5	73	63.6	0	0	1	0
2020/09/7	23	0	34.3	24.6	124	16.8	0	0	1	0
2020/09/14	30	0	37.8	44.4	77	55.9	0	0	1	0
2020/09/21	30	0	71.7	39.9	67	12.7	0	0	1	0
2020/09/28	23	0	78.3	47	130	64.1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...
2023/03/13	34	0	68	63	112	24.7	1	0	0	0
2023/03/20	30	0	0	19.8	13	3.1	1	0	0	0
2023/03/27	35	0	2.4	49.2	45	14.4	1	0	0	0
2023/04/3	17	0	5.4	76.5	26	1	0	1	0	0
2023/04/10	31	0	20.8	59.9	27	30.2	0	1	0	0

De igual forma, es necesario hacer una adaptación de los datos para poder aplicar técnicas de *machine learning* a la serie de tiempo. Esta debe reorganizarse en la forma de un conjunto de datos supervisado, es decir, una matriz en la que cada valor esta asociado a una ventana temporal conformada por número de rezagos que le preceden. Una vez que los datos se encuentran reordenados de esta forma, se puede entrenar un modelo para predecir el valor actual de la serie partir de los valores precedentes. Para este caso se ha definido usar tres rezagos como variables predictoras.

Figura 5: Conjunto de datos para modelos de *machine learning*

Serie de tiempo

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Variable exógena

a	b	c	d	e	f	g	h	i	j
---	---	---	---	---	---	---	---	---	---

X

1	2	3	4	5	f	6
2	3	4	5	6	g	7
3	4	5	6	7	h	8
4	5	6	7	8	i	9
5	6	7	8	9	j	10

y

Finalmente, luego de realizar todas las transformaciones a los datos, se obtiene un conjunto de datos con 137 registros y 10 variables, incluyendo la variable objetivo,

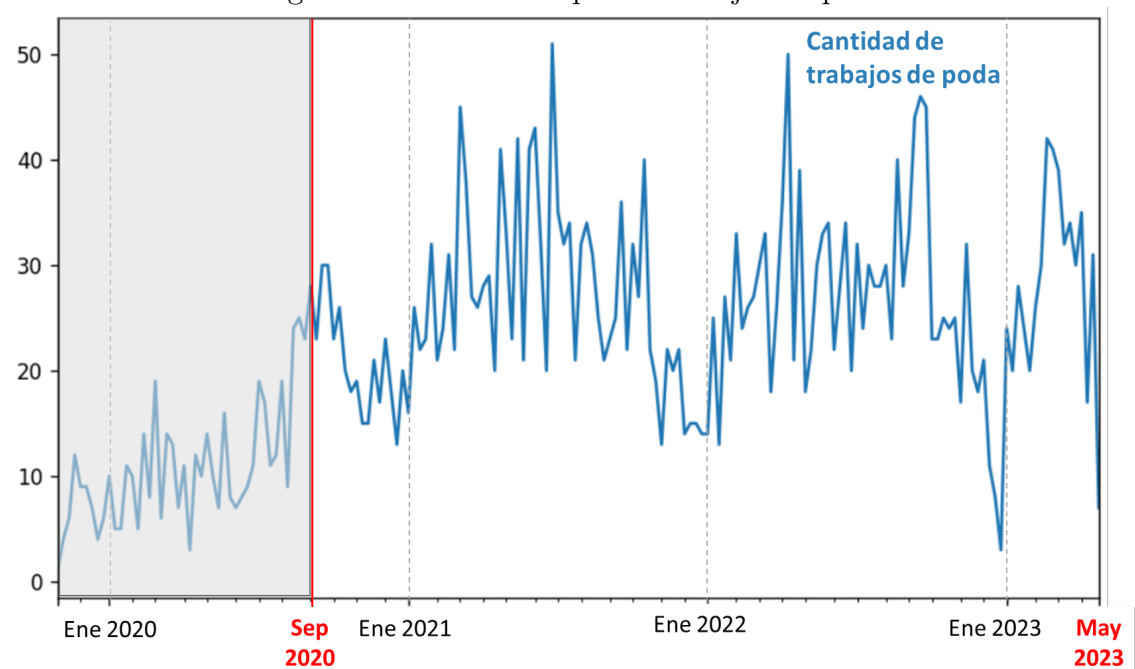
**cantidad de trabajos de poda.** Al adaptar este conjunto de datos a la forma necesaria para aplicar los modelos predictivos, este queda conformado por 134 observaciones, con 12 variables predictoras y la variable objetivo.

## Análisis exploratorio de los datos

El conjunto de datos que será analizado está compuesto por la serie de tiempo de los trabajos de poda, obtenida a partir del registro histórico de órdenes de trabajo como se expuso anteriormente.

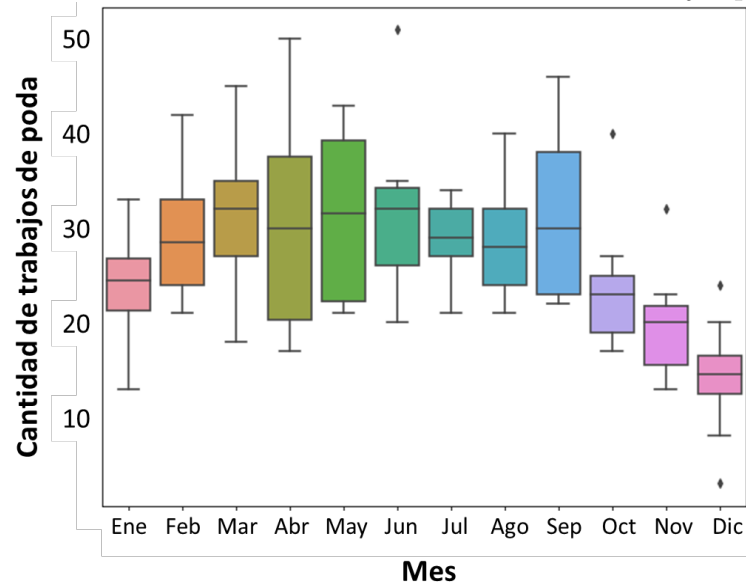
Inicialmente, la serie de tiempo está conformada por 182 registros, con periodicidad semanal, comprendidos en el periodo entre octubre de 2019 y abril de 2023. Al realizar una revisión de los datos, se observa una tendencia creciente en la cantidad de eventos registrados antes de septiembre de 2020. Esto se debe a un subregistro de los eventos en la etapa inicial de estabilización del sistema de información. Con el fin de evitar sesgos en las estimaciones se eliminarán los registros correspondientes a este periodo de estabilización con lo que el análisis se realizará sobre los 137 registros generados entre septiembre de 2020 y abril de 2023.

Figura 6: Serie de tiempo de trabajos de poda



Al evaluar la serie de tiempo, discriminando la cantidad de trabajos realizada en cada mes (sin tener en cuenta el año), se observa un comportamiento estacional que posteriormente es confirmado.

Figura 7: Estacionalidad de la serie - Cantidad de trabajos por mes



### Estacionaridad de la serie de tiempo

Se verifica la estacionaridad de la serie a través de las pruebas Augmented Dickey-Fuller test (ADF) y Kwiatkowski-Phillips-Schmidt-Shin (KPSS), obteniéndose los siguientes resultados:

- ADH Test:  $p - value = 0,0000864$
- KPSS test:  $p - value = 0,100000$

Para el caso del ADF test, la hipótesis nula  $H_0$  es que la serie de tiempo es no estacionaria. Así, dado que el  $p-value$  en el test es menor que el nivel de significancia 0,05, entonces se rechaza la hipótesis nula.

Para el caso del KPSS test la hipótesis nula  $H_0$  es la opuesta al ADF test, es decir que la serie es estacionaria. Así, dado que el  $p - value$  en el test es mayor que el nivel de significancia 0,05, entonces no se rechaza la hipótesis nula.

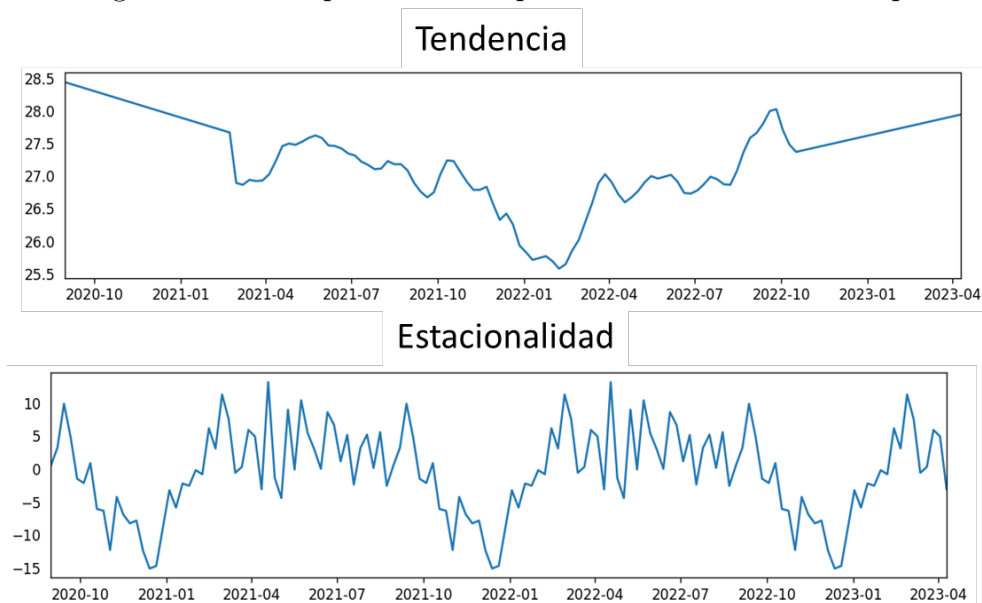
Luego, de acuerdo con los resultados conjuntos de las pruebas, la **serie es estacionaria**.



## Descomposición de la serie

Al realizar la descomposición multiplicativa y aditiva de la serie se puede verificar, por ambos modelos, que esta tiene una componente estacional pero no una componente de tendencia. A modo de referencia, se presenta a continuación el resultado de la descomposición multiplicativa de la serie.

Figura 8: Descomposición multiplicativa de la serie de tiempo

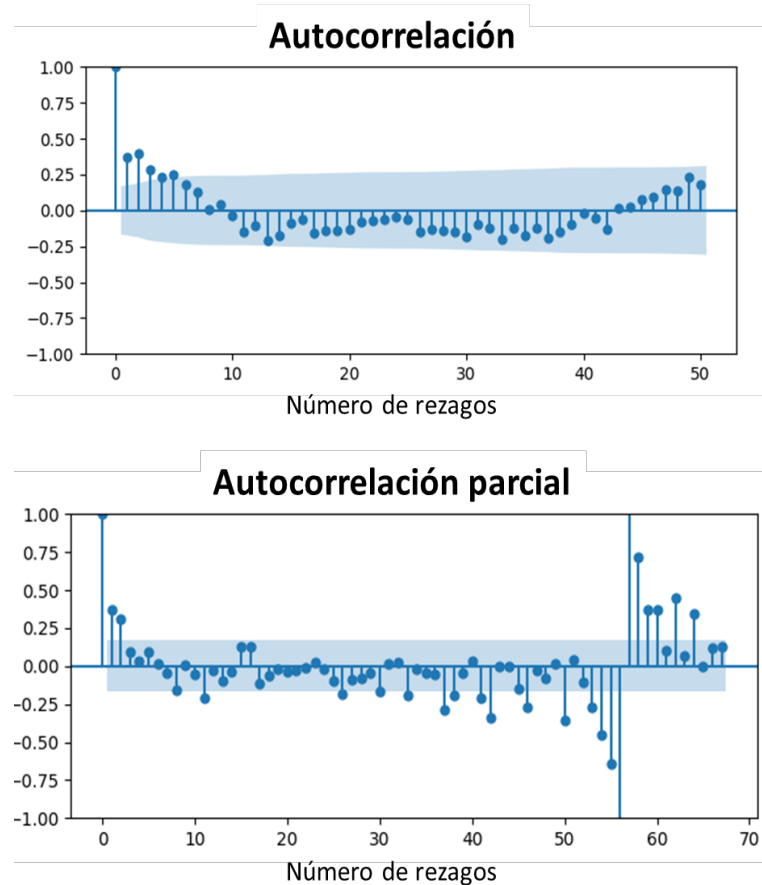


Comparando el resultado de ambos métodos, se identifica que la descomposición multiplicativa logra remover de mejor manera el componente estacional de la serie, y por lo tanto se selecciona esta descomposición para continuar con el análisis de la serie de tiempo.

## Funciones de autocorrelación y autocorrelación parcial

Se calculan las funciones de autocorrelación y autocorrelación parcial para la serie de tiempo original y para la serie sin el componente estacional usando descomposición multiplicativa. Se presentan a continuación los resultados para la serie de tiempo original (sin desestacionalizar).

Figura 9: Funciones de autocorrelación y autocorrelación parcial de la serie de tiempo



Como se observa en la imagen anterior, para la serie de tiempo original las funciones de autocorrelación y autocorrelación parcial indican que los valores previos de la serie (rezagos) pueden ayudar a predecir el valor actual. Se observa una autocorrelación de la serie con tres rezagos y una autocorrelación parcial con dos rezagos. Con estos datos se podrá determinar el orden de los procesos en un modelo ARIMA ( $AR = 2$  y  $MA = 3$ ).

Por su parte, se determinó que al transformar la serie de tiempo, eliminando su componente estacional, se pierde la relación de la serie con sus rezagos.

De acuerdo con los resultados anteriores, se determina realizar el análisis y modelado de la serie de tiempo con datos en su formato original, sin realizar sobre estos ninguna transformación, con el fin de mantener la capacidad predictiva de la serie con base en sus valores previos.

### 3.4. Modelado

El proceso de modelado de datos se inicia una vez completada la transformación de los datos. En este punto se procedió a la división de los mismos en conjuntos de entrenamiento y prueba. El proceso de modelado se llevó a cabo en dos enfoques diferentes. En primer lugar, se realizó la aproximación mediante el uso de modelos clásicos de series de tiempo, incluyendo modelos ARIMA, Auto ARIMA y ARIMAX. Se procedió a realizar la siguiente aproximación en el modelado mediante el uso de técnicas más sólidas para el análisis de series de tiempo. Es importante destacar que en esta aproximación se incorporaron variables exógenas, entre las cuales se incluye una variable binaria que representa los períodos de inactividad del sistema de recolección de registros, así como cuatro variables adicionales relacionadas con los datos de precipitación de cuatro estaciones ubicadas en la región suroeste de Antioquia.

En el modelado de las series de tiempo con métodos clásicos, los modelos ARIMA requieren los siguientes tres parámetros, los cuales fueron determinados a partir de las gráficas de autocorrelación y autocorrelación parcial detalladas en el numeral 3.3 (Preparación de los datos):

- **AR (autoregressive):** Este parámetro indica la cantidad de valores pasados (rezagos) que son relevantes para predecir el valor futuro. Esta información es obtenida al analizar la gráfica de autocorrelación parcial, para este modelo se determina usar un valor de 2 rezagos.
- **I (Integrated):** Este parámetro indica la diferenciación que se debe aplicar a la serie de tiempo con el objetivo de suavizar aspectos característicos de la misma como la tendencia y estacionalidad. Para este modelo se determina un valor de cero debido a la falta de tendencia y estacionalidad de la serie de tiempo en análisis.
- **MA (Moving Average):** Este parámetro indica como el valor futuro puede predecirse a partir de la componente aleatoria de datos anteriores. Este valor se obtiene al analizar la gráfica de autocorrelación, para este modelo se determina usar un valor de 3 rezagos.

Por otra parte, se abordó el problema usando técnicas de Machine Learning. En este sentido se llevó a cabo el modelado utilizando cuatro métodos : regresión lineal,

*Random Forest*, *XGBoost* y Redes Neuronales. Los últimos tres incluyeron un proceso de hiperparametrización con el objetivo de identificar los mejores parámetros para cada modelo.

Finalmente, se obtuvieron los resultados de cada uno de los modelos y con base en estos se toma aquel con el que se tuvo mejores resultados con base en las métricas elegidas para la evaluación. A continuación se muestran los resultados obtenidos.

### 3.5. Evaluación e interpretación

Posterior al modelado de los datos, se realizó la evaluación de los modelos implementados usando series de tiempo clásicas y técnicas de Machine Learning. A continuación se realiza una descripción breve de los resultados obtenidos en las diferentes aproximaciones.

#### Series de tiempo clásicas

Se implementaron modelos clásicos de series de tiempo, como modelos ARIMA, Auto ARIMA y ARIMAX (incluyendo variables exógenas). A continuación se puede observar la serie de tiempo con la predicción obtenida posterior a la aplicación de los modelos:

- **Modelo ARIMA:** En los resultados de este modelo se observa que la predicción no logra una representación adecuada de los datos reales, entregando una función completamente lineal sin la captura de los picos en la serie de tiempo. Las diferencias entre el valor real y la predicción en algunos casos llega a ser de 18 unidades.

Figura 10: Resultados del modelo *ARIMA*



- **Modelo Auto ARIMA:** En la evaluación de este modelo se determinan los mejores parámetros a través de un algoritmo, donde se llega a los valores  $AR(1)$ ,  $I(0)$ ,  $MA(1)$ . En este modelado se obtienen unos errores en la predicción que se evalúa a través de las métricas  $RMSE=9.9$  y  $MAE=8.6$ .
- **Modelo ARIMAX:** Este modelo se evalúa con las componentes  $AR(1)$ ,  $I(0)$ ,  $MA(1)$  obtenidas como los mejores parámetros determinado en el algoritmo AutoARIMA y adicionando las variables exógenas disponibilidad del sistema, cantidad de precipitaciones en la zona y el trimestre del año. Como parte de la inclusión de las variables exógenas al modelo, se realizó la prueba de hipótesis a dichas variables para determinar cuales de ellas son estadísticamente significativas, en la figura 11 se observa que para las variables trimestre\_1, trimestre\_2, trimestre\_3 y trimestre\_4, el p valor es menor a 0.05 y por lo tanto aportan información para la predicción del modelo.

Figura 11: Variables exógenas

SARIMAX Results						
=====						
Dep. Variable:	CANT_EVENTOS	No. Observations:	125			
Model:	SARIMAX(1, 0, 1)	Log Likelihood	-417.258			
Date:	Tue, 06 Jun 2023	AIC	858.516			
Time:	01:58:21	BIC	892.456			
Sample:	08-31-2020	HQIC	872.304			
	- 01-16-2023					
Covariance Type:	opg					
=====						
	coef	std err	z	P> z	[0.025	0.975]
-----						
DUMMY	-11.6290	9.724	-1.196	0.232	-30.688	7.430
bolombolo	0.0242	0.027	0.889	0.374	-0.029	0.078
santabarbara	0.0297	0.032	0.932	0.351	-0.033	0.092
fredonia	-0.0054	0.020	-0.263	0.792	-0.045	0.035
urrao	0.0057	0.030	0.192	0.848	-0.053	0.064
trimestre_1	24.6461	2.224	11.083	0.000	20.287	29.005
trimestre_2	28.1744	2.236	12.602	0.000	23.792	32.556
trimestre_3	26.9738	2.288	11.787	0.000	22.489	31.459
trimestre_4	18.4295	2.301	8.008	0.000	13.919	22.940
ar.L1	0.7207	0.674	1.069	0.285	-0.601	2.042
ma.L1	-0.6635	0.701	-0.947	0.344	-2.037	0.710
sigma2	46.4422	6.028	7.705	0.000	34.628	58.256
=====						
Ljung-Box (L1) (Q):	0.13	Jarque-Bera (JB):	14.17			
Prob(Q):	0.72	Prob(JB):	0.00			
Heteroskedasticity (H):	1.24	Skew:	0.76			
Prob(H) (two-sided):	0.48	Kurtosis:	3.65			
=====						

Los resultados obtenidos al incorporar dichas variables exógenas, denotan una mejoría en la predicción al lograr representar en mejor medida las variaciones de la serie de tiempo, sin embargo se mantiene una brecha importante entre el valor de la predicción y el valor real que en algunos casos llega a ser de 16 unidades. Adicionalmente, se logró mejorar las métricas de RMSE= 9.2 y MAE=7.6.

Figura 12: Resultados del modelo *ARIMAX*



Como métricas para comparar los resultados de los diferentes modelos se tomó el RMSE (Root Mean Squared Error) y el MAE (Mean Absolute Error). A continuación se muestran los cuadros comparativos para ambas métricas, en las cuales se puede observar que los mejores resultados se dan para el modelo ARIMAX al incluir todas las variables exógenas de disponibilidad del sistema, precipitaciones y trimestre.

Figura 13: Comparativo entre modelos: Métrica *RMSE*

Modelo	RMSE en datos de prueba	Observaciones
ARIMA (2,0,3)	9.635	Datos de la serie de tiempo únicamente
AUTOARIMA (1,0,1)	9.909	Datos de la serie de tiempo únicamente
ARIMAX	11.154	Datos de la serie de tiempo más variable dummy de disponibilidad del sistema
<b>ARIMAX</b>	<b>9.232</b>	Datos de la serie de tiempo, más variable dummy, <b>más datos de precipitaciones, más variable trimestre.</b>

Figura 14: Comparativo entre modelos: Métrica *MAE*

Modelo	MAE en datos de prueba	Observaciones
ARIMA (2,0,3)	8.422	Datos de la serie de tiempo únicamente
AUTOARIMA (1,0,1)	8.673	Datos de la serie de tiempo únicamente
ARIMAX	9.896	Datos de la serie de tiempo más variable dummy de disponibilidad del sistema
<b>ARIMAX</b>	<b>7.641</b>	Datos de la serie de tiempo, más variable dummy, <b>más datos de precipitaciones, más variable trimestre.</b>

## Técnicas de Machine Learning

Se implementaron modelos de machine learning aplicados a series de tiempo. Los modelos aplicados fueron: Regresión Lineal, Random Forest, XGBoost y finalmente, Red Neuronal. A continuación, se muestran los resultados.

- **Modelo Regresión Lineal:** Se realizó el entrenamiento del modelo haciendo uso de las variables provenientes de la serie de tiempo además de las variables exógenas, 4 de ellas provenientes de las estaciones de precipitación y las últimas 4 variables que representantan los trimestres del año. Después de aplicar la regresión lineal se obtuvieron los siguientes resultados.

Figura 15: Resultados del modelo *Regresión Lineal*

OLS Regression Results						
Dep. Variable:	y	R-squared:		0.376		
Model:	OLS	Adj. R-squared:		0.320		
Method:	Least Squares	F-statistic:		6.698		
Date:	Tue, 06 Jun 2023	Prob (F-statistic):		4.45e-08		
Time:	18:54:28	Log-Likelihood:		-409.61		
No. Observations:	122	AIC:		841.2		
Df Residuals:	111	BIC:		872.1		
Df Model:	10					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	14.5771	2.598	5.611	0.000	9.429	19.725
x1	4.5983	4.209	1.093	0.277	-3.742	12.938
x2	6.7910	4.353	1.560	0.122	-1.834	15.416
x3	0.6461	4.514	0.143	0.886	-8.300	9.592
bolombolo	3.3529	3.652	0.918	0.361	-3.883	10.589
santabarbara	2.8678	5.361	0.535	0.594	-7.756	13.491
fredonia	1.0354	4.610	0.225	0.823	-8.100	10.171
urrao	0.8834	4.699	0.188	0.851	-8.429	10.195
trimestre_1	4.7864	1.308	3.658	0.000	2.194	7.379
trimestre_2	6.3148	1.792	3.523	0.001	2.763	9.866
trimestre_3	5.7551	1.630	3.530	0.001	2.524	8.986
trimestre_4	-2.2792	1.155	-1.973	0.051	-4.568	0.009
Omnibus:	9.282	Durbin-Watson:		2.018		
Prob(Omnibus):	0.010	Jarque-Bera (JB):		9.143		
Skew:	0.628	Prob(JB):		0.0103		
Kurtosis:	3.469	Cond. No.		6.98e+15		

La regresión lineal revela información importante sobre el modelo entrenado y su capacidad para predecir eventos de poda. El uso de variables provenientes de la serie de tiempo y las variables exógenas, que incluyen datos de las estaciones de precipitación y los trimestres del año, permite capturar la influencia de factores tanto temporales como ambientales en la variable de interés.

Al examinar los resultados de la regresión lineal, se observa que solo los trimestres 1, 2 y 3 resultaron ser estadísticamente significativos. Esto sugiere que hay una relación significativa entre estos trimestres y la cantidad de eventos de poda. La falta de significancia en el trimestre 4 podría indicar que este período del año no influye de manera significativa en la variable de interés.

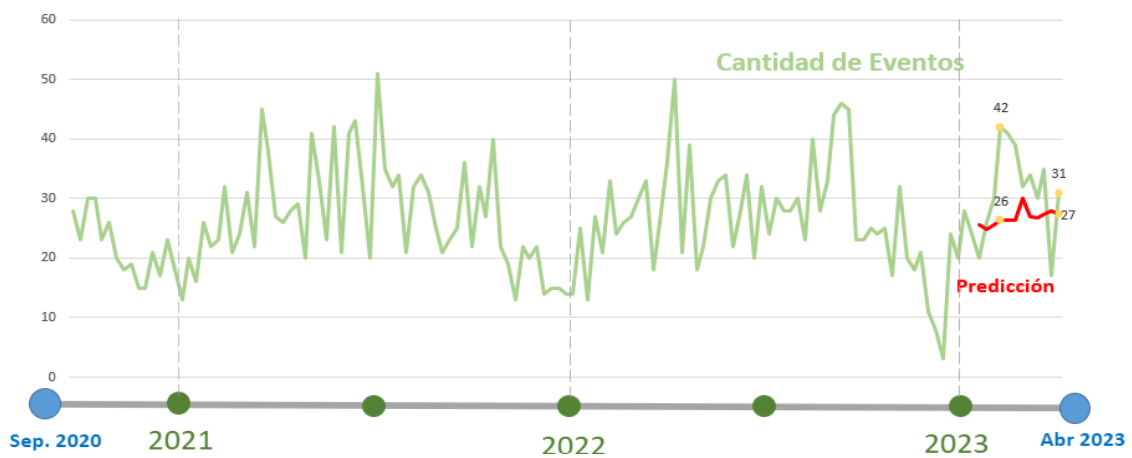
Además, se calcularon los errores en el conjunto de entrenamiento y el conjunto de prueba para evaluar la precisión del modelo. El Root Mean Square Error



(RMSE) en el conjunto de prueba fue de 6,95, mientras que en el conjunto de predicción fue de 8,18. Estos valores representan la discrepancia promedio entre los valores reales y los valores predichos por el modelo. Dado que los errores son relativamente bajos, podemos inferir que el modelo tiene una capacidad razonable para predecir eventos de poda.

- **Modelo Random Forest:** A continuación se puede ver la serie de tiempo con el modelo Random Forest en donde la línea roja representa la predicción.

Figura 16: Resultados del modelo *Random Forest*



La predicción anterior pasó por un proceso de ajuste con el objetivo de obtener mejores resultados. Se realizó una hiperparametrización, a través de la ejecución de la función gridsearch la cuál hace la comparación de los resultados obtenidos con cada configuración presentada haciendo uso de backtesting. Así, los parámetros que se validaron en el proceso se pueden ver en la siguiente figura, encontrando en la última columna los parámetros seleccionados en el proceso de hiperparametrización.

Figura 17: Resultados de la hiperparametrización *Random Forest*

• lags	3, 4, 5	<b>3</b>
• n_estimators	50, 100, 200	<b>50</b>
• criterion	squared_error, friedman_mse	<b>squared_error</b>
• max_depth	3, 5, 7	<b>3</b>

Los resultados del modelo Random Forest revelan el proceso de ajuste realizado para mejorar la precisión de las predicciones. Se llevó a cabo una hiperparametrización utilizando la función GridSearch, que permite comparar los resultados obtenidos con diferentes configuraciones mediante el uso de back-testing.

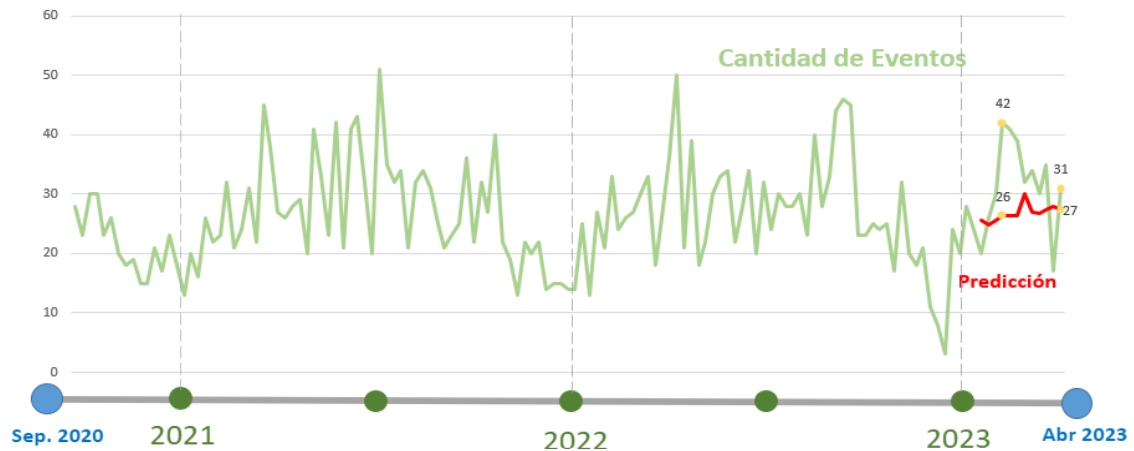
La figura anterior muestra los parámetros que se buscaron ajustar durante el proceso. Estos incluyen el número de lags, el número de estimadores, la profundidad máxima y el criterio de error. La hiperparametrización tiene como objetivo encontrar los valores óptimos para estos parámetros, lo que mejorará la capacidad del modelo para realizar predicciones precisas.

Los resultados finales muestran que los valores óptimos encontrados fueron de 3 lags, 50 estimadores, el criterio de error *squared\_error* y una profundidad máxima de 3. Estos parámetros representan la configuración que proporciona el mejor rendimiento del modelo en términos de precisión predictiva.

Para evaluar la precisión del modelo ajustado, se utilizó la métrica de Root Mean Square Error (RMSE). El valor de RMSE obtenido en el conjunto de prueba fue de 8.70. Esta métrica representa la discrepancia promedio entre los valores reales y los valores predichos por el modelo. Un valor de RMSE más bajo indica una mayor precisión en las predicciones, por lo que un valor de 8.70 indica un buen rendimiento del modelo en la predicción de los eventos de interés.

- **Modelo XGBoost:** En un proceso similar al realizado en el Random Forest, en el modelo XGBoost se obtuvo la serie de tiempo en donde la línea roja nuevamente representa la predicción.

Figura 18: Resultados del modelo *XGBoost*



En este modelo también se realizó un proceso de hiperparametrización haciendo uso de la función `gridsearch` con los parámetros que se pueden ver a continuación, encontrando en la última columna los parámetros seleccionados en el proceso de hiperparametrización.

Figura 19: Resultados de la hiperparametrización *XGBoost*

• <code>lags_grid</code>	[3, 4, 5]	<b>[1,2,3,4]</b>
• <code>n_estimators</code>	50, 100, 200	<b>100</b>
• <code>criterion</code>	mean_squared_error	<b>mean_squared_error</b>
• <code>max_depth</code>	3, 4, 5	<b>4</b>

El análisis de los resultados del modelo *XGBoost* revela un proceso de hiperparametrización realizado utilizando la función `GridSearch`. Se buscaron los mejores valores para los parámetros del modelo, que se presentan a continuación.

**Lags:** En este caso, se encontró que los mejores resultados se obtuvieron utilizando lags en los períodos 1, 2, 3 y 4. Esto significa que el modelo considera la información histórica de los eventos de interés en los cuatro períodos anteriores para realizar las predicciones.

Número de estimadores: Se determinó que la cantidad óptima de estimadores en el modelo XGBoost es de 100. Los estimadores representan los árboles de decisión individuales utilizados en el algoritmo de boosting.

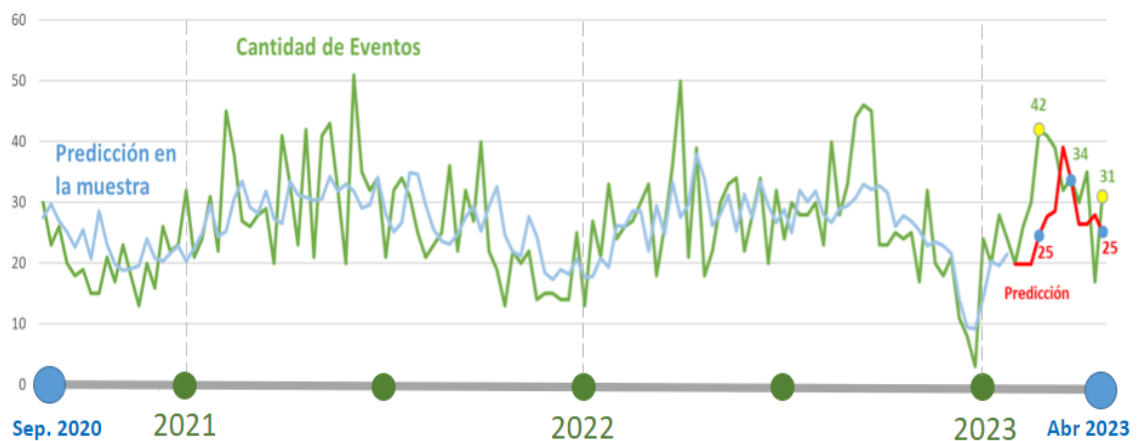
Profundidad máxima: La profundidad máxima se estableció en 4, lo que significa que cada árbol de decisión en el modelo XGBoost puede tener un máximo de 4 divisiones.

Es interesante destacar que, a pesar de tener parámetros diferentes, el modelo XGBoost produjo resultados muy similares a los obtenidos con el modelo Random Forest. Esto puede indicar que ambos modelos son efectivos para capturar las relaciones entre las variables y realizar predicciones precisas en el contexto del problema analizado.

Para evaluar la precisión del modelo XGBoost, se calculó el error de predicción en el conjunto de prueba utilizando la métrica de Root Mean Square Error (RMSE). El valor de RMSE obtenido fue de 8.21, lo que indica la discrepancia promedio entre los valores reales y los valores predichos por el modelo. Además, se calculó el Mean Absolute Error (MAE), que representa la diferencia promedio absoluta entre los valores reales y los valores predichos, y se obtuvo un valor de 6.68.

- **Modelo Redes neuronales:** A continuación se puede ver la serie de tiempo con el resultado de la aplicación del modelo Red neuronal, en donde la línea roja representa la predicción, esta logra identificar la tendencia de los datos reales y en algunos puntos se logran tener brechas pequeñas de 3 unidades de diferencia y en un punto de la misma es incluso de cero.

Figura 20: Resultados del modelo con *Redes Neuronales*



La predicción pasó por un proceso de ajuste con el objetivo de obtener mejores resultados. Para esto se realizó una hiperparametrización a través de la ejecución de la función gridsearch, la cuál hace la comparación de los resultados obtenidos con cada configuración presentada haciendo uso de crossvalidation. En la siguiente figura se pueden ver los parámetros que fueron validados en el proceso mencionado, encontrando en la última columna los parámetros seleccionados en el proceso de hiperparametrización.

Figura 21: Hiperparametrización *Red Neuronal*

• capa 1	32, 64	<b>32</b>
• capa 2	128, 256	<b>256</b>
• criterio	mae, mse	<b>mse</b>
• optimizador	Adam, RMSprop	<b>Adam</b>

Al igual que para los métodos clásicos, se utilizó el RMSE (Root Mean Squared Error) y el MAE (Mean Absolute Error) como métricas para comparar los resultados de los diferentes modelos. A continuación se muestran los cuadros comparativos para ambas métricas, en las cuales se puede observar que los mejores resultados se dan para el modelo basado Redes Neuronales con un RMSE de 7.77 y un MAE de 6.35.

Figura 22: Comparativo entre modelos ML: Métrica *RMSE*

Modelo	RMSE en datos de prueba	Observaciones
Regresión Lineal	8.175	$R^2$ ajustado = 0.320
<u>Random Forest</u>	8.694	{ <u>'criterion'</u> : 'squared_error', ' <u>max_depth</u> ': 5, ' <u>n_estimators</u> ': 50, <u>lags</u> : 5}
<u>XGboost</u>	8.214	{ <u>'learning_rate'</u> : 0.05, ' <u>max_depth</u> ': 4, ' <u>n_estimators</u> ': 100, <u>lags</u> : 4}
<b>Red Neuronal</b>	<b>7.772</b>	{ <u>'optimizer'</u> : adam, ' <u>hidden_units_l1</u> ': 32, ' <u>hidden_units_l2</u> ': 256,}

Figura 23: Comparativo entre modelos ML: Métrica *MAE*

Modelo	MAE en datos de prueba	Observaciones
Regresión Lineal	6.452	$R^2$ ajustado = 0.320
Random Forest	7.079	{'criterion': 'squared_error', 'max_depth': 5, 'n_estimators': 50, lags: 5}
XGboost	6.684	{'learning_rate': 0.05, 'max_depth': 4, 'n_estimators': 100, lags: 4}
<b>Red Neuronal</b>	<b>6.350</b>	{'optimizer': 'adam', 'hidden_units_l1': 32, 'hidden_units_l2': 256,}

### 3.6. Despliegue

Como alcance del proyecto integrador no se tiene planteado realizar un despliegue productivo del modelo, se busca que el ejercicio permita establecer la pertinencia de una futura implementación. Los hallazgos y conclusiones del proyecto servirán como base para tomar decisiones informadas sobre la viabilidad y los beneficios de aplicar el modelo de predicción de eventos de poda en redes de distribución de energía en el entorno real de EPM.

## 4. Conclusiones

- El modelo de Red Neuronal entrega una estimación de la cantidad de eventos de poda que se deben atender en la semana siguiente con RMSE de 7.77 lo cual se considera aceptable, teniendo en cuenta el contexto de negocio, aportando así a una planeación óptima de los recursos en campo.
- Las variables exógenas de precipitación de las estaciones Bolombolo, Fredonia, Santa Barbara y Urrao no resultaron estadísticamente significativas para los modelos clásicos de series de tiempo. Por su parte la variable “Trimestre” resultó estadísticamente significativa y aporta a la capacidad predictiva del modelo.
- La metodología desarrollada en este proyecto integrador entrega predicciones que se desvían en ocho unidades del valor real de la cantidad de trabajos de

poda por ejecutar en una semana, por lo cual se considera que la metodología puede ser replicada en las otras regiones del departamento de Antioquia.

## 5. Anexos

- Repositorio GitHub del proyecto. [Proyecto\\_Integrador\\_2](#)

## Referencias

- [1] Sanchez, Daniel Peña: *Análisis de series temporales*. Alianza Editorial S.A., 2010.
- [2] Kolker, Alexander: *Forecasting Time Series*. En *HEALTHCARE MANAGEMENT ENGINEERING: WHAT DOES THIS FANCY TERM REALLY MEAN? - THE USE OF OPERATION MANAGEMENT METHODOLOGY FOR QUANTITATIVE DECISION-MAKING IN HEALTHCARE SETTINGS*, Springer Briefs in Health Care Management and Economics, páginas 79–90. 2012, ISBN 978-1-4614-2067-5.
- [3] Alzaman, C.: *Forecasting and optimization stock predictions: Varying asset profile, time window, and hyperparameter factors*. Systems and Soft Computing, 5, 2023.
- [4] Siqueira-Filho, E.A., M.F.A. Lira, A. Converti, H.V. Siqueira y C.J.A. Bastos-Filho: *Predicting Thermoelectric Power Plants Diesel/Heavy Fuel Oil Engine Fuel Consumption Using Univariate Forecasting and XGBoost Machine Learning Models*. Energies, 16:29–42, 2023.
- [5] Sallam, Abdelhay Ahmed y Om Parkash Malik: *Electric distribution systems*. John Wiley Sons, Inc, 2019, ISBN 9781119509318.