Análisis de Datos

Proyecto final

Accidentes Eléctricos en Colombia desde 2010 hasta 2019

Jou Jaramillo Uribe
Steven David Oviedo Herrera

Universidad de Medellín

Ingeniería Electrónica

2023-1

Introducción

En la actualidad, la electricidad juega un papel tan importante en la cotidianidad de los humanos casi como si de respirar se tratara, puesto que brinda desde iluminación hasta la capacidad de dar vida a diversos avances tecnológicos que impactan positivamente el bienestar de quienes cuentan con esta. A pesar de esto, su misma capacidad de sorprender la convierte en un agente altamente influyente en lo riesgosa y peligrosa que puede llegar a ser cuando no es correctamente manipulada, puesto que, de llegarse a sentir como un simple hormigueo, realizar un contacto directo puede generar espasmos o incluso que el mismo corazón deje de latir.

Al identificar esta problemática surge la idea de evaluar la posibilidad de analizar diversos casos de accidentes eléctricos en Colombia, con la finalidad de tener la capacidad de predecir nuevos eventos derivados de la mala manipulación o de fallas en esquemas de seguridad o de la misma maquinaría.

Al realizar una investigación se identifica un set de datos público sobre accidentes ocurridos en Colombia entre los años 2010 y 2019 en el sector eléctrico, el cual, discrimina con claridad factores como la fecha de lo sucedido, la ubicación departamental, la consecuencia médica de lo sucedido y algunos otros aspectos.

El objetivo principal del proyecto, será entonces que a partir del dataset seleccionado se obtenga un modelo para realizar predicciones de accidentes en el sector eléctrico del país que sirva como información para empresas y personas de interés con la cual fortalecer los métodos para prevenir los diversos riesgos causados por manipular la electricidad. Por lo mencionado, la metodología de trabajo se basó primeramente en realizar un entendimiento de negocio que diera paso a la preparación de los datos obtenido, un proceso de modelado, una evaluación y finalmente un despliegue con el fin de sacar conclusiones al respecto.

1. Entendimiento de negocio

Esta etapa se refiere a la predicción de accidentes anuales en Colombia de origen eléctrico. Para esto se realiza una comprensión y análisis de los accidentes ocurridos entre los años 2010 y 2019, lo cual nos darán una idea de la frecuencia con la que ocurren este tipo de eventos en el país.

a) Establecer objetivos

- La investigación acarrea el objetivo de hacer un reconocimiento mensual de los incidentes eléctricos ocurridos entre 2010 y 2019 para predecir e identificar patrones de dichos eventos para contar con información útil a la hora de crear medidas de seguridad activas y pasivas a los operadores de red por parte de empresas del sector, con lo cual se busca minimizar la frecuencia de ocurrencia de estos.

b) Evaluar la situación

- El número de accidentes en el país en el sector mencionado, tiene mucho que ver con la frecuencia en que se manipulan estos equipos que necesariamente necesiten una actualización o mantenimiento, esto depende mucho del sector del país en que se encuentre, como por ejemplo los equipos en sectores que estén a nivel del mar requerirán manteamientos más contantes.

  Por otro lado, también tiene una gran influencia en los incidentes ocurridos lo expuestos que estén estos puntos eléctricos para la sociedad, lo que abre la posibilidad de que personas mal intencionadas o incluso menores de edad que desconocen la gravedad y el peligro que representa someterse a este tipo de sitios acaben en un evento mortal

c) Preguntas que el modelo responderá

- ¿Se puede realmente predecir el número de eventos riesgosos del sector eléctrico de un determinado número de meses siguientes con la naturaleza de los datos seleccionados, teniendo en cuanta que estos fueron extraídos el periodo 2010-2019?
- ¿El modelo obtenido solo sirve para accidentes eléctricos o podría ser escalado para otras situaciones de sectores industriales distintos (automovilísticos, constructoras, robos, aseguradoras, fenómenos naturales, etc.)?
- ¿Qué porcentaje de datos deben ser asignados para entrenamiento y prueba con el finde obtener un mejor modelo predictivo?
- ¿Qué tan viable pueden ser las predicciones obtenidas y que porcentaje de confiabilidad representan estas?

d) Riesgos

- Se espera que los datos obtenidos sean confiables debido a que fueron extraídos de fuentes oficiales del gobierno colombiano y que adicionalmente, el hecho de que sean abiertos al público conlleva a que muchas personas puedan dar su retroalimentación con respeto a la opinión que tengan de lo útil que les ha sido toda la información publicada. Pero esto no quiere decir que no se realizará una correcta preparación del dato y un pre-estudio, analizando el contenido del dataset y verificar la coherencia de estos.
  Aun así, sabemos que existe la posibilidad de toparnos con errores humano debido a que se trata de un conjunto de datos extenso por lo que hay una gran probabilidad de que este esté afectado de cierta manera por malas digitaciones, entre otros.

2. Entendimiento de datos

Para realizar un correcto modelamiento y poder construir predicciones a futuro que estén realmente en concordancia con resultados esperados y con los datos que se seleccionan, es importante comenzar por hacer una exploración de los datos de manera que se comiencen a encontrar posibles errores, riesgos para no llegar al éxito del producto, plantear un plan de trabajo y saber las correcciones necesarias a hacer en una próxima etapa de tratamiento de los datos.

Para esto se emplea obtener la información del dataset obtenido, con lo cual entender la naturaleza de los datos (si son enteros, buleanos, Strings, entre otros), así como todas las columnas y clasificaciones dadas a estos. Se realiza entonces, empleando Python, librerías pandas y otras herramientas de programación, una extracción de información del dataframe (la manera en la que se agrupan los datos).

```
In [908]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3168 entries, 0 to 3167
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   IDENTIFICADOR_EMPRESA  3168 non-null   float64
 1   EMPRESA                3168 non-null   object
 2   FECHA                  3168 non-null   object
 3   ANIO                   3168 non-null   int64
 4   TRIMESTRE              3168 non-null   int64
 5   HORA                   3168 non-null   object
 6   COD_DANE               3168 non-null   int64
 7   DEPARTAMENTO           3145 non-null   object
 8   MUNICIPIO              3145 non-null   object
 9   CENTRO_POBLADO         3145 non-null   object
 10  UBICACION              3168 non-null   object
 11  SEXO                   3168 non-null   object
 12  EDAD                   3168 non-null   int64
 13  TIPO_IDENTIFICACION    3168 non-null   object
 14  VINCULADO_EMPRESA      3168 non-null   object
 15  TIPO_VINCULACION       3168 non-null   object
 16  GRADO_ESCOLARIDAD      3168 non-null   object
 17  TIEMPO_VINCULACION     3168 non-null   int64
 18  SECCION_EMPRESA        2957 non-null   object
 19  TIPO_LESION            3168 non-null   object
 20  ORIGEN_ACCIDENTE       3168 non-null   object
 21  CAUSA_ACCIDENTE        3168 non-null   object
 22  MEDIDAS                3006 non-null   object
dtypes: float64(1), int64(5), object(17)
memory usage: 569.4+ KB
```

Se encuentra entonces que múltiples columnas podrían ser descartadas como información no realmente útil para alcanzar el objetivo propuesto. Así mismo, se identifica que las columnas de Fecha y Departamento presentan alta relevancia en el dataset extraído, esto porque podemos establecer una serie de tiempo, a la par que la columna Departamento permite generar unas frecuencias de datos de forma mensual como se desea hacer esto.

Adicional a la información extraída, se realiza una descripción de los datos que permite obtener algunos valores estadísticos necesarios de marcar, tales como los valores mínimos, máximos, el promedio de los valores numéricos de los datos, entre otra información. De aquí se evidencia que hay datos de edad donde los valores mínimos están en 1 y los máximos en 99, lo cual es una gran sospecha puesto que no es tan normal que personas de 1 año o menos de vida tengan alta relación con el sector eléctrico, así mismo como tampoco personas de 99 años.

```
In [907]: data.describe()
```

Out[907]:

	IDENTIFICADOR_EMPRESA	ANIO	TRIMESTRE	COD_DANE	EDAD	TIEMPO_VINCULACION
count	3168.000000	3168.000000	3168.000000	3.168000e+03	3168.000000	3168.000000
mean	3165.964015	2015.402778	2.651199	3.630353e+07	30.848169	9238.954545
std	5833.239490	2.867237	1.095615	2.769699e+07	22.945206	4145.502366
min	480.000000	2010.000000	1.000000	0.000000e+00	1.000000	1.000000
25%	603.000000	2013.000000	2.000000	1.100100e+07	14.000000	11111.000000
50%	2249.000000	2016.000000	3.000000	2.341703e+07	30.000000	11111.000000
75%	2249.000000	2018.000000	4.000000	6.617000e+07	43.000000	11111.000000
max	44278.000000	2022.000000	4.000000	9.520000e+07	99.000000	11111.000000

De lo visualizado aquí, se procede a plantear un plan de ruta donde se debe hacer un tratamiento de los datos para eliminar columnas sin información útil y además, verificar que las edades sean correctas, así mismo como crear una serie de tiempo conforme a la evaluación mensual que se necesita hacer.

Se verifica también que la fecha está en un formato complejo puesto que está desordenada y no de forma estandarizada, por lo que a su hora de digitación se aplicaron varias maneras y símbolos de separación de fecha tales como “aaaa-mm-dd” o “dd/mm/aaaa”, esto es una corrección que se debe de hacer. Finalmente se logra identificar que los datos de departamento cuentan con algunos datos vacíos, para lo cual será necesario aplicar también una corrección de eliminación de datos nulos.

3. Preparación de los datos

Como se mencionó en el segundo apartado del proyecto, se debe realizar una preparación de datos para obtener información realmente útil, verificable, no inventada y que permita comprender la realidad tras los accidentes eléctricos sucedidos. Con lo cual, lo primero que se procede a realizar es eliminar las columnas de datos que no brindan información a emplear para realizar el modelamiento. Esto se observa en la siguiente imagen:

```
In [910]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3168 entries, 0 to 3167
Data columns (total 2 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   FECHA         3168 non-null   object
 1   DEPARTAMENTO  3145 non-null   object
dtypes: object(2)
memory usage: 49.6+ KB
```

Otra corrección importante es hacer la verificación de las fechas, para que estas queden en el mismo formato y agrupadas como son necesarias, así que lo primero es aplicarles una transformación de formato.

```
In [919]: data['FECHA']=pd.to_datetime(data['FECHA'], infer_datetime_format=True, format="%Y/%m/%d")
          data.head()
```

Out[919]:

	FECHA	DEPARTAMENTO
0	2015-09-05	CASANARE
1	2014-11-27	CAUCA
2	2015-10-02	LA GUAJIRA
3	2012-08-12	MAGDALENA
4	2016-11-04	LA GUAJIRA

El siguiente paso es hacer un conteo de los accidentes ocurridos por departamento, de esto se evidencia cuáles son los departamentos más influyentes en accidentes, así mismo como los menos influyentes, puesto que departamentos con muy pocos datos son partidarios a ser información errada o mal diligenciada.

```
In [912]: data['DEPARTAMENTO'].value_counts()

Out[912]: ATLÁNTICO              405
          VALLE DEL CAUCA        380
          ANTIOQUIA              312
          BOLÍVAR                309
          MAGDALENA              197
          CÓRDOBA                140
          BOGOTÁ, D.C.           128
          CESAR                  119
          SANTANDER              109
          RISARALDA              101
          PUTUMAYO                92
          LA GUAJIRA              87
          SUCRE                   86
          CAUCA                   84
          CALDAS                  83
          NARIÑO                  74
          TOLIMA                  69
          HUILA                   59
          NORTE DE SANTANDER      53
          META                    51
          CHOCÓ                   50
          QUINDÍO                 49
          CUNDINAMARCA            41
          BOYACÁ                  30
          CAQUETÁ                 20
          CASANARE                13
          GUAVIARE                 3
          ARAUCA                   1
          Name: DEPARTAMENTO, dtype: int64
```

De este proceso se eliminan entonces los datos de Guavare, Arauca, así mismo como los vacíos detectados, de esta manera hacemos validación que los datos restantes son correctamente los que necesitamos para el objetivo de predicción. Se menciona también que se hace el cambio de los datos marcados como Bogotá a Cundinamarca, el departamento correcto:

```
In [922]: data = data.sort_values(by='FECHA')
          data.drop([953], axis=0, inplace=True)
          data.drop([1061], axis=0, inplace=True)
          data.drop([698], axis=0, inplace=True)
          data.reset_index(drop=True, inplace= True)
          data = data.drop(data.index[2979:3161])
          data = data.drop(data.index[0:6])
          data = data[data['DEPARTAMENTO'].notna()]
          data.to_csv('depts.csv')
          data
```

Out[922]:

	FECHA	DEPARTAMENTO
6	2010-07-13	LA GUAJIRA
7	2010-07-13	RISARALDA
8	2010-07-13	ANTIOQUIA
9	2010-07-13	CÓRDOBA
10	2010-07-14	TOLIMA
...	...	...
2974	2019-10-26	MAGDALENA
2975	2019-10-28	BOLÍVAR
2976	2019-10-28	ATLÁNTICO
2977	2019-10-29	SUCRE
2978	2019-10-30	VALLE DEL CAUCA

2950 rows × 2 columns

Finalmente se hace el cambio de los datos de departamento a números y frecuencias que se puedan agrupar por fechas, lo cual se puede evidenciar de la siguiente imagen. Se debe recordar la importancia de que los datos anteriores a 2010 fueron eliminados, también los que iban más allá del 2020, puesto que evidentemente pandemia fue un gran influyente en la reducción de accidentes.

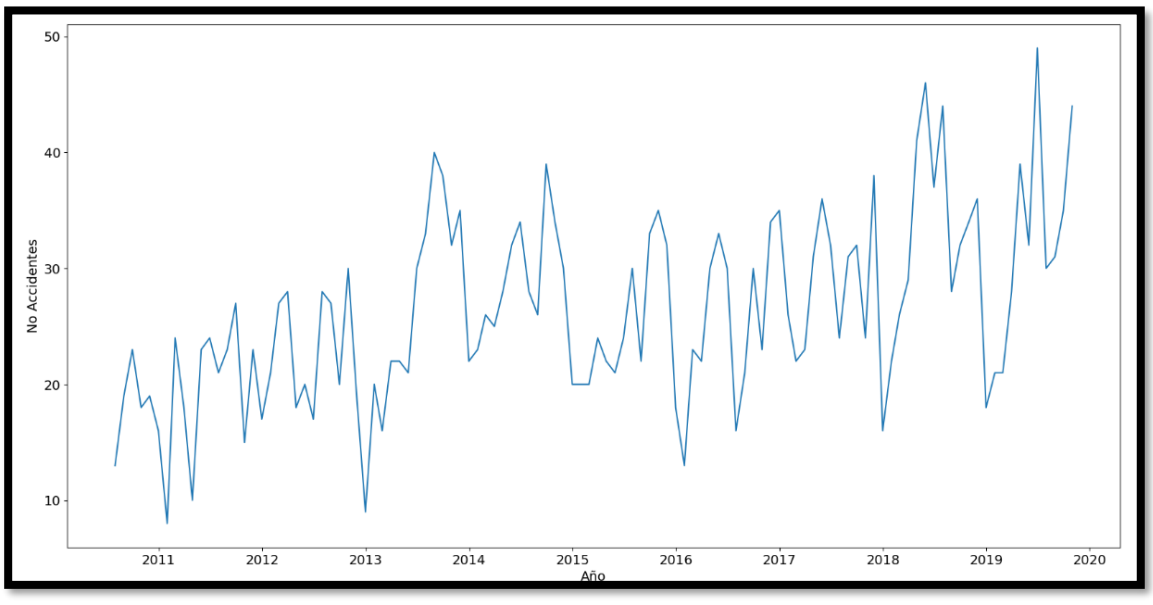
```
In [924]: data=data.set_index(['FECHA'])
          data = data.resample('M').sum()
          data.to_csv('numeros.csv')
          data
```

Out[924]:

FECHA	DEPARTAMENTO
2010-07-31	13
2010-08-31	19
2010-09-30	23
2010-10-31	18
2010-11-30	19
...	...
2019-06-30	49
2019-07-31	30
2019-08-31	31
2019-09-30	35
2019-10-31	44

112 rows × 1 columns

La gráfica que se muestra a continuación es el resultado de los datos obtenidos tras realizar una correcta manipulación de estos, así mismo como un acote, de esto se evidencia una tendencia creciente con el paso de los tiempos y algunos picos y valles importantes. Se hablará de esto más adelante.



4. Modelación

Primeramente, en este proceso se realizaron 8 diferentes modelos entre ARIMA y SARIMA, esto con el fin, de evaluar distintos eventos y escenarios con el objetivo de hallar el mejor modelo que de cierta manera logre una mejor adaptación a los datos reales. Al realizar estos, uno a uno se iba haciendo un análisis y tratando de hacer una comparación entre ellos para seleccionar el que mejor se comportara con respecto a la descripción de los registros originales.

Es por esto que se iban realizando las gráficas de cada modelo frente a los datos reales, se hacía un análisis entre estas, pero nos dimos cuenta que de cierta manera, existía una similitud y no se lograba evidenciar una diferencia significativa la cual nos permita elegir un modelamiento como el mejor, es por esto, que también se llevó la extracción de las métricas de cada modelo. Estas están conformadas por el error promedio cuadrático (MSE)y el error promedio absoluto (MAE).

Todo este desarrollo se hizo con el único objetivo de tener las bases necesarias para tomar la decisión contundente al seleccionar el mejor modelo que describía los datos reales, pero al realizar la comparación de las métricas nos dimos cuenta que por medio del MSE era de cierta manera complejo interpretarlo y, por lo tanto, también era difícil tomar esta decisión.

Es por esto, que nuestra última opción fue apoyarnos en el error promedio absoluto (MAE) como parámetro a tener en cuenta para elegir el mejor modelo.

Para este caso, el mejor modelo conseguido según lo descrito y mencionado anteriormente fue el quinto modelo realizado, el cual se trata de un modelo ARIMA (Promedio móvil integrado autorregresivo). Esta técnica es un modelo estadístico que se utiliza para pronosticar el conjunto de datos que tiene una naturaleza de serie temporal. El modelo ARIMA no es más que una composición de tres modelos, los cuales son, autorregresión (AR), integración (I) y media móvil (MA).

Recordemos que los parámetros de un modelo ARIMA son (p,d,q), donde p es para laparte "AR", que representa cuántos rezagos se van a incluirán en el modelo, d es para la parte "I", la cual representa el número de veces se integró o se aplicó una diferencia en la serie. Y la q es para la parte "MA" que es el número de medias móviles de errores que se incluirán en la serie.

Teniendo, esto claro, los parámetros de nuestro modelo elegido son: (7,1,7) y sus respectivas métricas fueron las siguientes:

MSE: 33.7502200010534
MAE: 4.81573321759115

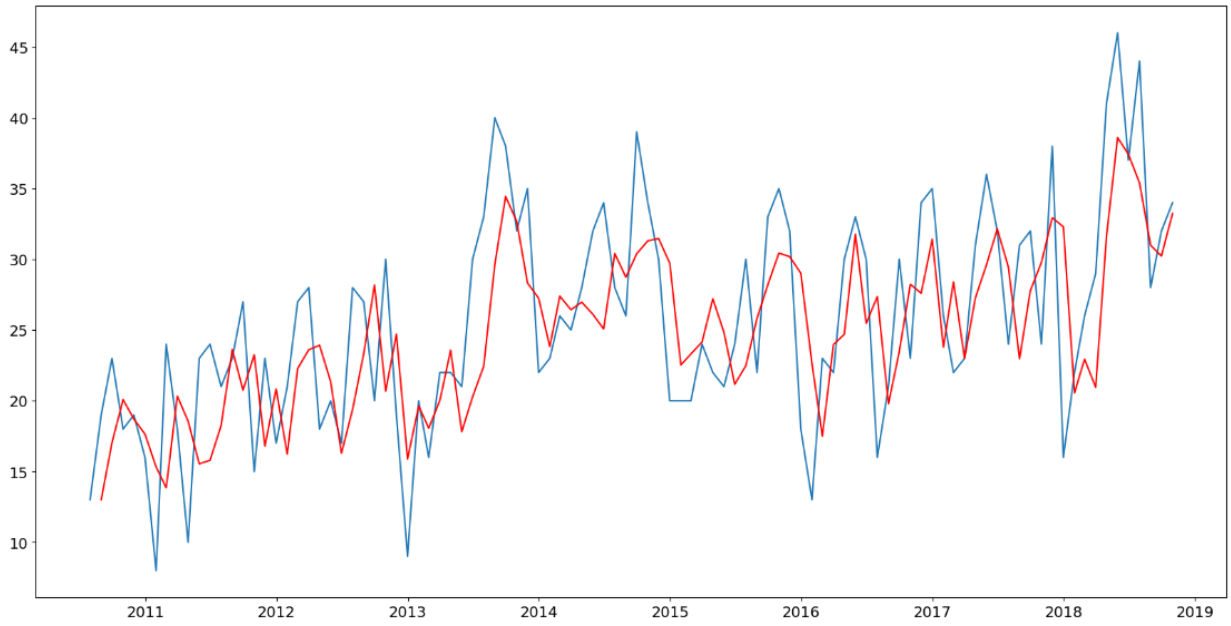


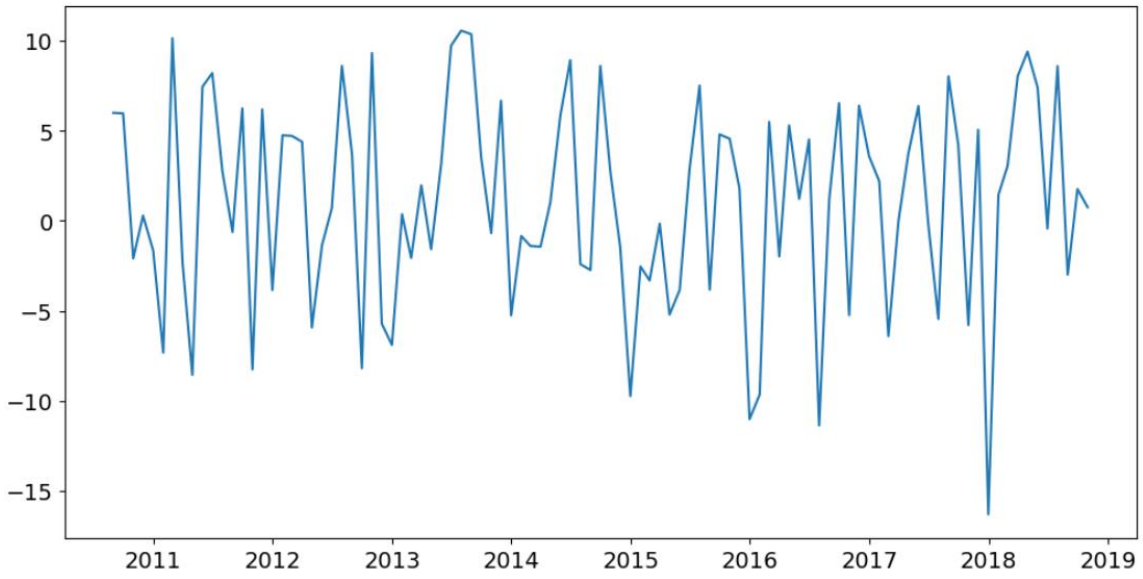
Ilustración: Grafica de los datos reales frente al mejor modelo encontrado [ARIMA (7,1,7)].

Como se puede observar, el modelo (línea roja) tiene un buen comportamiento frente a los datos reales (línea azul), pues se evidencia una gran similitud en los diferentes cambios y picos que hay en el transcurso de las gráficas. De cierta manera el modelo imita de muy buena forma y en gran parte a los registros reales, solo que tiene una característica peculiar y es que esta cuenta con menor amplitud.

5. Evaluación

Ya realizado la selección de nuestro mejor modelo, en este caso, 5to modelo ARIMA, procedemos a realizar la comprobación de sus métricas en el conjunto de prueba y de que sus residuos si tengan forma de ruido blanco.

Es muy importante que los residuos de un modelo ARIMA, al graficarlos, tengan forma de ruido blanco, ya que en caso de que no sea así no cumpliría con lo que dice la estadística, por lo que lo ideal es reajustar el modelo con otros hiperparametos.



Gráfica ruido blanco.

Recordemos que cuando de hacemos referencia a ruido blanco hablamos de una señal aleatoria que se caracteriza por el hecho de que sus valores de señal en dos tiempos diferentes no guardan correlación estadística. Como notamos en la gráfica, esta tiene un compartimento totalmente aleatorio por lo tanto podemos afirmar que el residuo si tiene forma de ruido blanco y por lo tanto también con esto se puede afirmar que nuestros residuos son independientes.

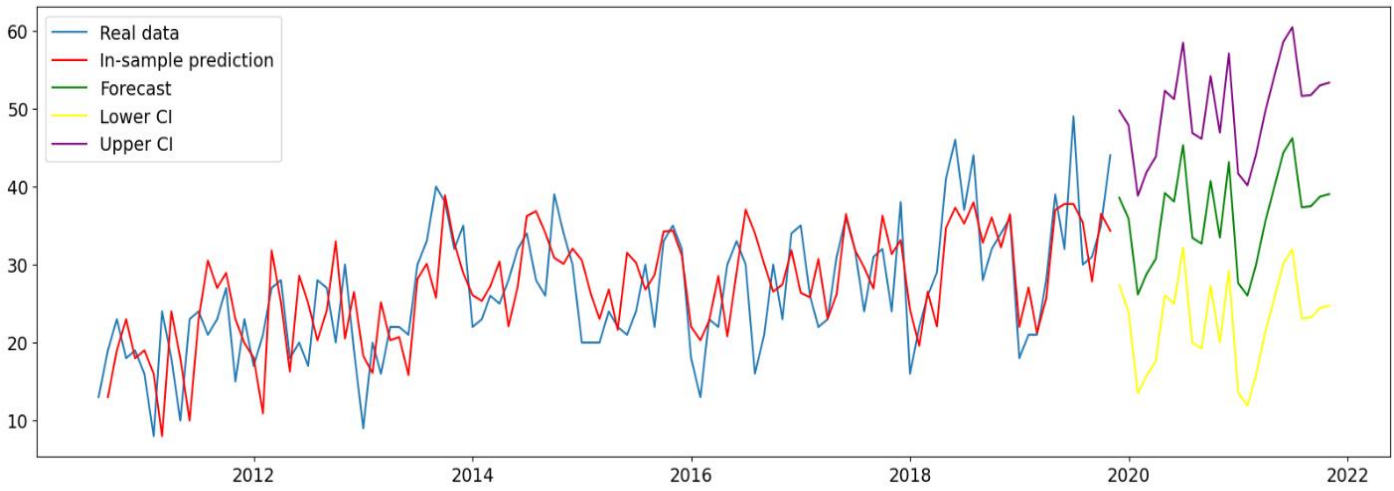
Ahora, al hallar el residuo medio vemos que este es de 1.0213994630855563. No tan cercana a cero como quisiéramos idealmente, pero al analizarlo conociendo muy bien nuestros datos y la escala de estos, podemos decir que tener un accidente eléctrico de error promedio es algo relativamente bajo. Es por esto que podemos continuar con este modelo a realizar las predicciones futuras.

Por último, hacemos uso del test de Ljung-Box, el cual sirve para comprobar si nuestros residuos son independientes. De cierta manera con el ruido blanco ya lo hemos comprobado, pero con este test vamos reafirmarlo. Recordemos que la hipótesis nula de este test es que los datos sí son independientes; la alternativa es que no lo son.

Al realizarlo obtenemos p-value de 0.909006, con esto no hay una suficiente evidencia para que podamos hacer caso omiso a la hipótesis nula de que los datos si son independientes. Dado que se ve que los residuos son independientes, y su media es cercana a 0, pero no tanto como se esperaba.

6. Despliegue

Se procede a generar predicciones a futuro con intervalo de confianza de alrededor del 95%, en un lapso de 24 meses después. Esto con el fin de tener una idea de cómo será el comportamiento de los accidentes eléctricos en Colombia al paso de 2 años después de 2019 que es el año hasta donde llegan nuestros datos.



	predicted_mean
2019-11-30	38.555769
2019-12-31	35.885043
2020-01-31	26.160769
2020-02-29	28.784896
2020-03-31	30.730252
2020-04-30	39.168383
2020-05-31	38.074500
2020-06-30	45.289069
2020-07-31	33.431844
2020-08-31	32.671668
2020-09-30	40.702642
2020-10-31	33.469241
2020-11-30	43.134920
2020-12-31	27.625831
2021-01-31	26.035072
2021-02-28	29.861418
2021-03-31	35.411532
2021-04-30	39.817849
2021-05-31	44.344356
2021-06-30	46.191046
2021-07-31	37.350265
2021-08-31	37.475626
2021-09-30	38.696604
2021-10-31	39.029989

Gráfica predicciones futuras.

Al observar la gráfica notamos unos limites superiores (violeta) e inferiores (amarillo) los cuales describen el comportamiento en los rangos máximos y mínimos en el que se espera que oscilen el número de accidentes eléctricos en Colombia en casos extremos, para los años en que se hizo la predicción.

Se espera que para los próximos 2 años desde 2019, haya un aumento significativo en el número de sucesos eléctricos en Colombia, esto lo podemos evidenciar en el comportamiento que describe la gráfica de color verde, ya que como vemos en esta existen unos picos los cuales están por encima de todos los que hay la grafica que describe el modelo (rojo), es por esto que se puede afirmar que para los próximos 28 meses en Colombia, haya un aumento con respecto a los años anteriores en cuanto a lo que accidentes en el sector eléctrico se trata.

Esta información puede ser de gran utilidad para empresas del sector en el país ya que de cierta manera pueden tomar cartas en el asunto y hacer lo posible para disminuir el número de accidentes tomando las respetivas prevenciones y precauciones los operarios del campo.

Por otro lado, también podemos evidenciar el numero de accidentes eléctricos que habrá en esos meses que se predijeron.

Como notamos en los registros mes a mes que fueron predichos, existen números de gran magnitud y mayores a los que encontrábamos en los datos reales que extrajimos desde un principio. Con esto, tenemos otro apoyo para afirmar que al cabo de 2 años existirá un aumento en esos accidentes eléctricos en el territorio nacional de Colombia.

Muchos de los casos que notamos al trabajar con este dataset es que se presentaban accidentes también por fuera del campo laboral, normalmente estos tenían mucho que ver con menores de edad. Es por esto esta información sería también de importancia en general para familias que tal vez vivan cerca de puntos en el que se encuentren dispositivos eléctricos de gran riesgo, esto con el objetivo de que padres de familia informen a sus hijos y eviten el contacto con sitios como estos.

Conclusiones

- Para el desarrollo de este proyecto la técnica que se usó para la modelación fue por medio de ARIMA ya que con esta se obtuvo el mejor modelado, pero esto no siempre será así, es por eso que también hay comprobar con la técnica SARIMA, así como también ensayar con distintos parámetros en estos, con el fin de probar diferentes modelos y hallar la mejor opción para un llevar a cabo un buen desarrollo de lo que se vaya a realizar.

- Para llegar a un correcto modelado se evidencia la importancia en las variaciones y errores obtenidos de hacer un correcto acote de los datos, así como también eliminar aquellos que estén vacíos o que no representen un aporte importante al objetivo a conseguir.

- En el transcurso y desarrollo de los modelos de predicción diseñados, se resumen y concluyen que estos pueden ser escalados y aplicados tanto a la cantidad de información que se necesite, siempre y cuando cumplan con ser claramente separados mes a mes sus frecuencias, así mismo como a los sectores industriales, empresariales y de otra naturaleza, donde se desee predecir la ocurrencia de un evento claramente discretizado o escogido, sean accidentes o eventos diversos, esto es, posibilidad de aplicación por parte de empresas de seguros con respecto a accidentes automovilísticos, empresas de sector alimenticio, altas cadenas industriales, entre otras.