Pronóstico de ventas en kilos de un producto con ventas al por menor de una empresa de alimentos en Antioquia

Mateo Usme Valencia
Jorge Iván Rojas Díaz

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor
Jhon Jair Quiza Montealegre, Msc en Ingeniería - Telecomunicaciones

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2022

Cita	Usme Valencia y Rojas Díaz [1]
Referencia Estilo IEEE (2020)	[1] M. Usme Valencia y J. I. Rojas Díaz, "Pronóstico de ventas en kilos de un producto con ventas al por menor de una empresa de alimentos en Antioquia", Monografía, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, 2022.

Especialización en Analítica y Ciencia de Datos, Cohorte III.

I. RESUMEN EJECUTIVO

El objetivo principal es la predicción de las ventas en kilos por mes de los productos, esto para conocer con anticipación suficiente un estimado de la cantidad de materia prima necesaria para satisfacer la demanda futura y realizarse antes de las alzas de precios y negociar un precio de compra que permita generar rentabilidad con la posterior venta de los productos finales.

Los datos proporcionados se componen de las fechas de la factura por cada venta desde el 2010 hasta el 2022, los registros de ventas en kilos, el código de producto, las categorías y subcategorías de los productos. Inicialmente la estrategia fue realizar la predicción de los valores de kilos con un modelo supervisado de regresión utilizando variables exógenas, posteriormente las iteraciones se realizaron con otras herramientas donde se utilizó modelos de regresión univariables apoyados del componente de tendencia y estacionalidad de los datos de un producto en específico para obtener mejores resultados que en las primeras iteraciones una vez se comparan con datos reales contra los predichos por el modelo final.

El mejor resultado entre los modelos candidatos fue el del modelo ARIMA [1] el cual ofrece un pronóstico muy positivo en cuanto al problema de negocio que se desea afrontar. En una configuración óptima este tipo de modelos aprovechan muy bien el componente de estacionalidad en las series de tiempo y al tener esta ventaja los resultados suelen tener una precisión muy aceptable. El modelo ARIMA aprovecha la combinación de las propiedades Autorregresiva (AR), Integración (I) y Media Móvil (MA) para alcanzar el mejor resultado en los pronósticos de series de tiempo.

Link del repositorio de GitHub: www.github.com/MateoUsme/Esp_Analitica

II. DESCRIPCIÓN DEL PROBLEMA

Dadas las coyunturas actuales derivadas de la pandemia de la COVID-19, los impactos en la economía y las consecuencias en toda la cadena de abastecimiento que afectan la productividad de las organizaciones, se hace necesario mejorar la eficiencia en la planeación de las compras de las materias primas para la fabricación de los productos. El pronóstico del volumen de ventas evita un sobre abastecimiento o falta de materias primas que procuran el correcto funcionamiento de la empresa. Estos datos son de primordial importancia para el departamento de cadena de abastecimiento, que al conocer los lead time de los proveedores y demás características a tener en cuenta, pueden realizar una planificación ideal que apalanque procesos transversales dentro de la empresa.

Debido a que el proyecto tiene como fin pronosticar el comportamiento de ventas de un producto, se opta por tomar el que tiene las mayores ventas en kilogramos de los dos últimos años fiscales (el año fiscal comprende los meses entre octubre y septiembre de una multinacional del sector alimenticio). La variable kilogramos se prioriza dado que al realizar una explosión de las materias primas requeridas para la fabricación de sus productos se parte como insumo principal el volumen a producir. Con esta información y las condiciones de los proveedores, el departamento de abastecimientos podría realizar la planificación de las compras requeridas.

A este producto se le desarrollará y aplicará un modelo de pronóstico de las ventas en kilos, aprovechando la data contenida en la Data Warehouse que tiene el histórico de ventas desde el año 2010. La información (fecha, kilos y clasificaciones) se extraerá como una serie de tiempo y haciendo uso de los métodos estadísticos clásicos y de redes neuronales, determinar el modelo más ajustado a las características del negocio.

A. Problema de negocio

Las diferentes variables que forman parte de la cadena logística global de suministros, tienen un gran impacto en la forma como las empresas deben orientar sus esfuerzos en cuanto a su

abastecimiento para evitar tanto un sobre stock o falta de materias primas que lleguen a paralizar la cadena productiva por la escasez de una sola materia prima. Tanto la falta de materia prima como el exceso de inventario son factores críticos que impactan la empresa a nivel financiero debido a que estos constituyen un activo de la misma. Además, la materia prima requiere una rápida salida (dado que ese es el propósito de ser adquirida) de no ser así, se afectaría el equilibrio financiero de la misma. La materia prima forma parte del estado de resultados que se entraría a restar del costo de las mercancías disponibles con implicación directa en el capital de trabajo. [2]

Es por esto que la previsión y estimación de la demanda se ha vuelto fundamental en las empresas convirtiéndose en una tarea estratégica multidisciplinaria que conlleva mucho análisis. El Know - how de los encargados de la cadena de aprovisionamiento, que apoyados en herramientas de pronóstico más el conocimiento de las fluctuaciones de compras de los clientes por el departamento de ventas, determinan la planificación de compras más acertada de acuerdo a las diferentes temporadas de ventas teniendo en la compañía unos niveles de inventario apropiados.

La necesidad de la empresa requiere un tratamiento de la información ideal para realizar pronósticos mensuales: al ser una empresa que vende productos elaborados con materia prima obtenida de manera anticipada a precios que les permiten ser competitivos, el objetivo fundamental es tener una aproximación de cuánta cantidad se debe abastecer en materias primas antes de que los precios de las mismas suban y generen pérdidas o utilidades no suficientes para mantener el negocio. Aunque las ventas se realicen diariamente los encargos de materia prima se ejecutan en grandes lotes para economizar costos.

B. Aproximación desde la analítica de datos

Los modelos predictivos a emplear tienen por finalidad el pronosticar la demanda en kilos del producto con mayores ventas en kilogramos en los dos últimos años fiscales, con los datos que se encuentran en la Data Warehouse de ventas, aprovechando la estimación en diferentes escalas de

tiempo (años, meses, días o semanas) y que sirva como insumo principal para la planificación de compras de las materias primas requeridas.

C. Origen y caracterización de los datos

La empresa de donde provienen los datos es una multinacional perteneciente al sector de producción de alimentos, con especial interés en que sus productos sean seguros, saludables y de calidad, en un entorno global que monitorea cuidadosamente las tendencias alimentarias emergentes ayudando a los socios a ofrecer creaciones de vanguardia.

Aprovechando la Data Warehouse que tiene registros de ventas desde el 2010 se obtiene el producto con mayor ventas en kilogramos de los dos últimos años fiscales que incluye el periodo entre octubre y septiembre.

La información de la *Data Warehouse* tiene su origen en dos bases de datos transaccionales cuyo motor es Microsoft SQL Server. Los campos extraídos para este modelamiento son:

- Datos de las fechas con estacionalidad que comprende año, mes y día, pudiendo también obtener la semana.
- Códigos del producto del cual se extrae el código madre o bulk.
- Cantidad de kilogramos vendidos.
- Clasificaciones.
    - Categoría: productos complementarios o sustitutos (p. ej. dressing, seasoning, etc).
    - Drive: características complementarias que generan estrategias de mercadeo (p. ej. representation, resale, etc).
    - Subdrive: diferentes apartados que están contenidos en un Drive específico.
    - Segmento: características de los consumidores de acuerdo a las metas de la empresa y su público objetivo (p. ej. QSR, retail, etc.).
    - Subsegmento: diferentes apartados que están contenidos en un Segmento específico.

De acuerdo a la fecha en que fueron efectuadas las ventas la información se tomará de dos bases de datos así :

1. Datos entre marzo de 2010 a septiembre de 2017 (Database 1)
2. Datos entre octubre de 2017 a la fecha actual (Database 2)

D. Métricas de desempeño

Los modelos predictivos se evaluarán con diferentes métricas de errores para determinar la calidad de la predicción, dichas métricas serán MAPE [3] y RMSLE [4] como un conjunto de valores que ayudarán a determinar el desempeño de cada modelo de una manera relativa y no con magnitudes reales según la escala de los datos predichos.

La métrica de negocio principal será la variabilidad porcentual en el precio de insumos sugerido por el modelo comparado con el precio total de los insumos que se hubiesen comprado sin el apoyo de algún modelo de predicción. También una estimación del incremento de utilidades por producto y su volatilidad del último año comparando el momento en el que el modelo empezó a influir en la toma de decisiones.

III. DATOS

A. Datos Originales

TABLA I
DATOS A UTILIZAR PARA REALIZAR EL PRONÓSTICO

Campo	Tipo de dato	Descripción
FechaFactura	String (yyyymmdd)	Fecha de la venta del producto
Año	Integer	Año de fecha de la venta del producto
Nombre del mes	String	Nombre del mes de la venta del producto
Nombre del día	String	Nombre del día de la venta del producto

Número de la semana	Integer	Número de la semana del año en que se realizó la venta del producto
Código	String	Código del producto vendido
Kilos	Float	Cantidad de kilos vendidos del producto
NombreCategoria	String	Nombre de la categoría a la que pertenece el producto vendido
NombreDrive	String	Nombre del drive a la que pertenece el producto vendido
NombreSubDrive	String	Nombre del Subdrive a la que pertenece el producto vendido
NombreSegmento	String	Nombre del Segmento a la que pertenece el cliente que adquirió el producto
NombreSubsegmento	String	Nombre del Subsegmento a la que pertenece el cliente que adquirió el producto

B. Datasets

A partir de los datos originales cargamos el modelo, se convierte la variable fecha a una periodicidad mensual, puesto que no tiene afectación considerable con datos atípicos en este rango de tiempo. Se realizó la codificación de todas las variables categóricas para ser utilizadas dentro del modelo como pesos y realizar pruebas con librerías tradicionales para machine learning.

Los datos principales utilizados en la primera iteración tienen un tamaño total de 24.1 MB, con un total de 304 800 registros y variables como la fecha de la factura del producto vendido, el grupo o código del producto, país y departamento donde se vendió el producto, categoría, drive, subdrive y la cantidad del producto vendido en unidades de kilos. Los datos utilizados en la segunda iteración son los mismos que se emplearon en la primera iteración. Sin embargo, estos datos, a diferencia de los primeros, se enfocan en un solo producto. El tamaño total del archivo es de 161 KB, con un total de 3 896 registros y variables como la fecha completa de la factura del producto vendido, la semana, el año y el número de la semana separados en diferentes variables, el código del producto y la cantidad vendida en unidades de kilos.

C. Descriptiva

La Figura 1 , 2 y 3 muestran la exploración de los datos originales en la primera iteración, la cual emplea el uso de todas las variables del dataset principal tanto categóricas como numéricas. La figura 4, 5, 6 y 7 presentan la exploración de los datos originales en la segunda iteración transformados a periodicidad mensual y filtrados a un solo producto debido a la necesidad del problema de negocio descrita en el literal anterior, además de que se halla estacionalidad en un producto individual.

	FechaFactura	GrupoProducto	NombrePais	NombreDepartamento	NombreCategoria	NombreDrive	NombreSubDrive	Kilos
0	2018-05-18	685325	COLOMBIA	BOGOTÁ, D.C.	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	117.6
1	2018-05-18	688435	COLOMBIA	BOGOTÁ, D.C.	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	16.0
2	2018-05-18	688435	COLOMBIA	ANTIOQUIA	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	16.0
3	2018-05-18	688443	COLOMBIA	CESAR	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	12.8
4	2018-05-18	688443	COLOMBIA	VALLE DEL CAUCA	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	192.0
...	...	...	...	...	...	...	...	...
304795	2021-10-30	685063	COLOMBIA	VALLE DEL CAUCA	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	136.0
304796	2021-10-30	685063	COLOMBIA	ANTIOQUIA	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	23.8
304797	2021-10-30	685063	COLOMBIA	HUILA	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	136.0
304798	2021-10-30	685398	COLOMBIA	MAGDALENA	SAUCES & DRESSINGS	CUSTOM CULINARY	ZAFRÁN	64.0
304799	2021-10-30	696005	COLOMBIA	RISARALDA	OTHER PRODUCTS	SIN CLASIFICAR	SIN CLASIFICAR	102.2

304800 rows × 8 columns

Fig 1. Exploración de los datos originales en la primera iteración: Dataset

Fig 2. Exploración de los datos originales en la primera iteración: Visualización por columna Drive

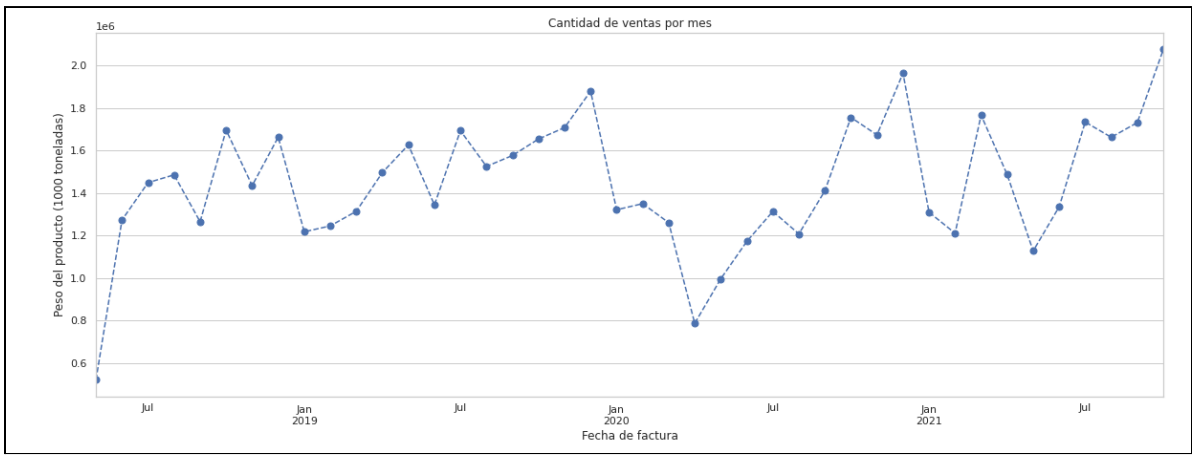

Fig 3. Exploración de los datos originales en la primera iteración: Serie de tiempo de las ventas en kilos mensuales

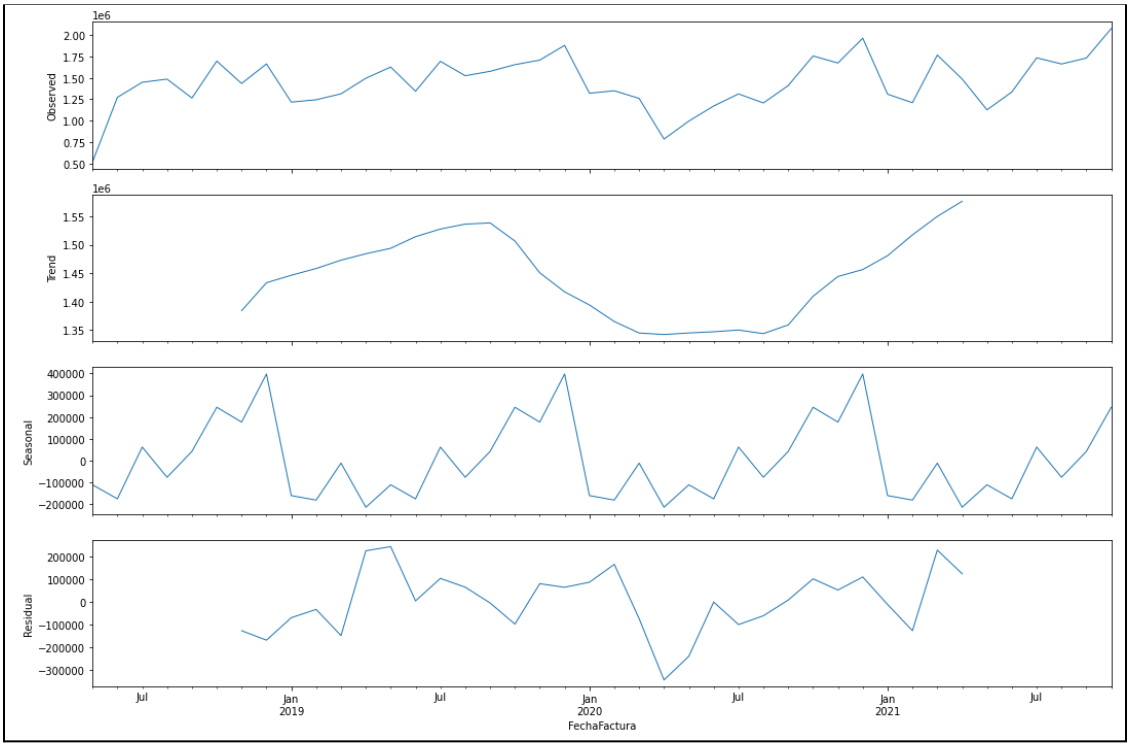

Fig 4. Descomposición de la serie de tiempo de la primera iteración en sus componentes de tendencia, estacionalidad y residual.

FechaFactura	Kilos
2018-11-30	922.6
2018-12-31	110247.4
2019-01-31	36202.6
2019-02-28	35353.0
2019-03-31	76954.6
2019-04-30	75514.6
2019-05-31	66313.0
2019-06-30	71885.8
2019-07-31	89266.6
2019-08-31	77401.0

Fig 5. Exploración de datos transformados en la segunda iteración: Dataset

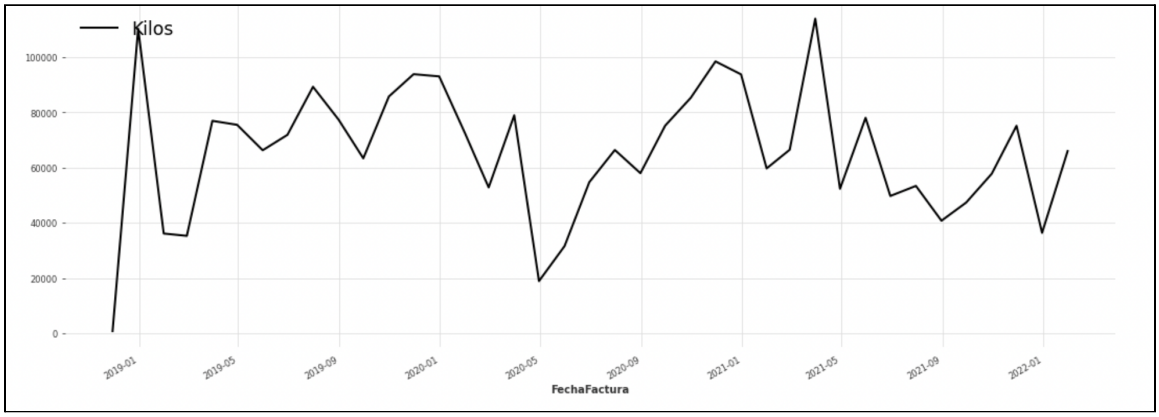

Fig 6. Exploración de datos transformados en la segunda iteración: Forma de la serie de tiempo

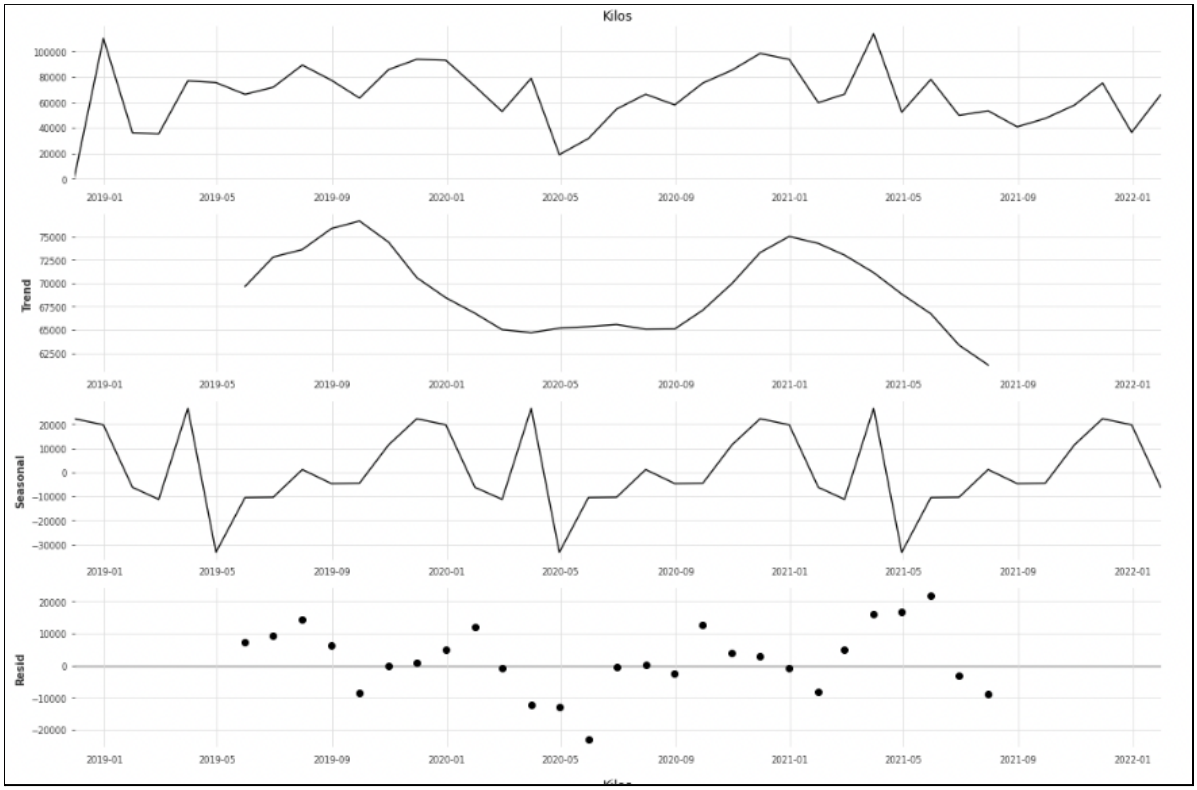

Fig 7. Descomposición de la serie de tiempo de la segunda iteración en sus componentes de tendencia, estacionalidad y residual.

IV. PROCESO DE ANALÍTICA

En esta sección se describen los procesos de análisis para el pronóstico del dataset del presente estudio.

A. Pipeline principal

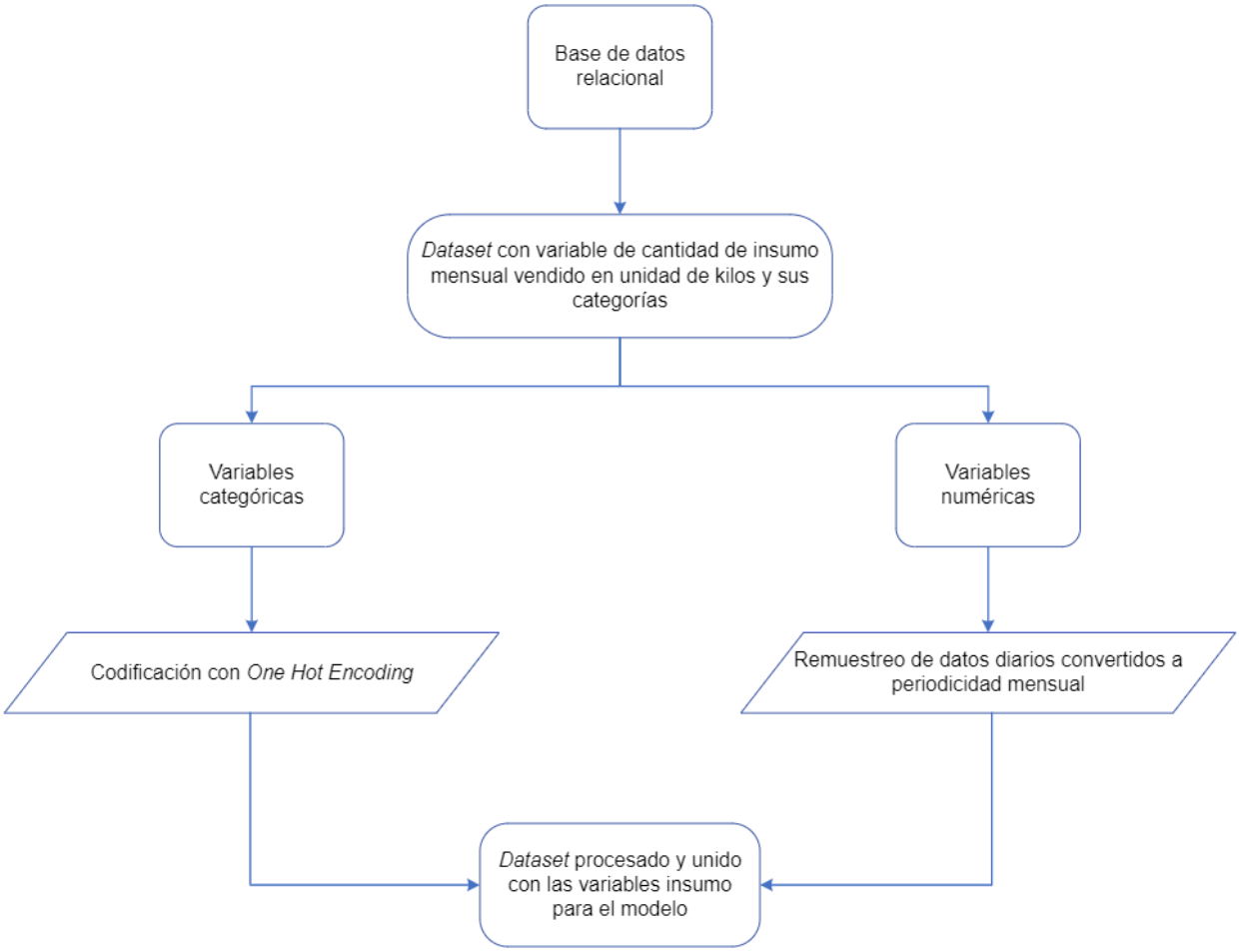

Fig 8. Metodología general de construcción del modelo de pronóstico

B. Preprocesamiento

En la primera iteración se realizó un procesamiento a los datos categóricos los que se codificaron por medio del método One Hot Encoding [5] buscando tener peso en todas las nuevas

columnas generadas por la codificación. Se revisó que las variables categóricas no tuvieran errores de ortografía o caracteres que causaran problemas al momento de tratar dichas variables.

En la segunda iteración se realizó un preprocesamiento de la variable “fecha” (datetime) para obtener el formato “yyyymmdd”. Luego, se realizó un remuestreo a la variable de “kilos” cambiando su periodicidad de diaria a mensual. Previendo que el dataset pudiera tener meses sin ventas, se suma 1 a las ventas para que la interpretación del modelo no generase errores con valores en cero (0) en los datos. Para las últimas iteraciones, el dataset se divide en datos de entrenamiento, validación y prueba. Esta separación permite optimizar el modelo y realizar las comparaciones de los resultados con las predicciones obtenidas. Se cuenta con un dataset donde se reflejan las ventas de un solo producto ya filtrado con un total de 39 registros de periodicidad mensual con valores de kilos totales vendidos en cada mes.

C. Modelos

Los modelos más adecuados a utilizar fueron los siguientes:

- **ARIMA:** El modelo Autorregresivo Integrado de Media Móvil se compone de distintos modelos para descomponer y determinar según los valores de cada modelo la disposición en el tiempo de los datos. Se compone de la parte Autorregresiva (AR) que se refiere al uso de valores pasados en la ecuación de regresión para la serie de tiempo, la componente de Integración (I) que utiliza la diferenciación de observaciones en la serie de tiempo para convertirla en estacionaria y la componente de Media Móvil (MA) que utiliza la dependencia entre una observación y un error residual con una ventana de tiempo elegida dentro del modelo.

  La configuración óptima utilizada para el modelo ARIMA fue AR(1), I(2), MA(1).

- **Theta [6]:** El modelo theta emplea la descomposición de la serie de tiempo en un conjunto de series de tiempo, el resultado de dicha serie son llamados líneas theta, y mantienen la media y la pendiente de los datos originales, pero no sus curvaturas. Su

proceso consiste en modificar las curvaturas de las series de tiempo ajustadas por el coeficiente theta el cual es un parámetro del modelo, este coeficiente es aplicado en la segunda derivada de las series de tiempo. Cada una de las líneas theta son extrapoladas de manera separada y las predicciones se combinan con pesos iguales para obtener el resultado final del modelo.

La configuración óptima empleada para este modelo hallada de manera iterativa con diferentes valores fue la resultante con un valor Theta de 0.153

- **Naive Drift + Seasonal [7]:** El modelo Naive y sus componentes Drift y Seasonal se conforman de la primera componente como la línea de tendencia de la serie de tiempo, y la segunda componente como el orden de la estacionalidad y la forma de esta misma estacionalidad.

  Este modelo obtiene la línea de tendencia y la forma de la componente estacional de la serie de tiempo para realizar su predicción, lo que la hace dependiente de una estacionalidad en los datos que se utilizan y desea pronosticar. Sin la estacionalidad, la capacidad de predicción de este modelo decae considerablemente.

- **Random Forest [8]:** Random Forest [Bosques Aleatorios] es una técnica de aprendizaje automático basada en el modelo de árboles de decisión, cada uno de estos árboles calcula un valor en base a los parámetros de los datos a través de las decisiones tomadas y el valor promedio de todos los resultados es el valor obtenido por el modelo.

  La configuración de este modelo empleada en el proyecto fue de 10 estimadores y una profundidad máxima en los árboles de decisión de 5 niveles.

D. Métricas

Las métricas son calculadas a partir de funciones de la librería DARTS [9] y Sklearn [10] según con cuál se realiza la predicción para mantener la integridad del código, se da una breve descripción de qué nos ofrece el resultado de cada una:

- **MAPE [4]:** El Mean Absolute Percentage Error [Error Porcentual Absoluto Medio] es una métrica que permite medir la magnitud del error de manera porcentual. La mayor ventaja de este indicador es su fácil interpretación pues el valor obtenido representa en porcentaje el error de la predicción comparado con los datos reales evitando obtener un valor cuando se no usa un porcentaje (MAE) muy grande o muy pequeño según la magnitud aunque porcentualmente el error sea el mismo.

  El MAPE se calcula con la siguiente fórmula:

$$MAPE = \frac{100}{N} * \sum_{i=1}^{N} \left| \frac{x_i - y_i}{x_i} \right| \qquad (1)$$

  Donde:

  - xi son los valores reales de la serie de tiempo.
  - yi son los valores predichos de la serie de tiempo.
  - N es el número de valores de la serie de tiempo.

- **RMSLE [5]:** El Root Mean Square Log Error [Error Logarítmico Cuadrático Medio] es una métrica que permite calcular de manera relativa las diferencias entre los valores predichos y los reales. La ventaja de este indicador que son de gran interés en este proyecto es que las magnitudes de los valores no afectan el resultado al calcularse de manera relativa obteniendo un resultado orientado a cuánto varían los resultados cada mes sin importar que en magnitud los de un año atrás sean cinco veces más pequeños que los actuales. Además el RMSLE penaliza en mayor medida a las predicciones que están por

debajo del valor real comparado con las predicciones por encima de este, dicha penalización es de interés ya que obtener un valor por debajo del real lleva a la posibilidad de quedar sin insumos en cierto momento del mes, mientras que estar un poco por encima puede ser tolerado siempre y cuando no se vean afectadas las ventas a los clientes.

El RMSLE se calcula con la siguiente fórmula:

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (log(x_i + 1) - (log(y_i + 1))^2} \quad (2)$$

Donde:

- xi son los valores reales de la serie de tiempo
- yi son los valores predichos de la serie de tiempo
- N es el número de valores de la serie de tiempo

Las métricas de negocio se construyen a partir de los datos que contengan series de tiempo. La predicción obtenida por el modelo seleccionado (p. ej. ARIMA, Naive, Theta, etc.) se selecciona de acuerdo a la precisión del mismo y el dato real de dicho producto para el mes que se pretende predecir. La elección del modelo se determina de acuerdo a:

- La variabilidad porcentual en el precio de insumos sugerido por el modelo comparado con el precio total de los insumos que se hubiesen comprado sin el apoyo de algún modelo de predicción
- La volatilidad del último año comparando el momento en el que el modelo empezó a influir en la toma de decisiones

V. METODOLOGÍA

En esta sección se explicará la metodología usada en el presente estudio.

A. Baseline

La primera iteración consistió en un análisis exploratorio de los datos iniciales para tener un conocimiento de las variables que componen el dataset y así seleccionar los elementos determinantes. Por último se codifican las variables categóricas al crear nuevas columnas por cada categoría individual y asignando un valor de uno (1) o cero (0) en cada fila indicando que la categoría pertenecía a dicha fila, esto para obtener un dataset con una estructura adecuada que constituyen el insumo del modelo.

En la primera iteración se realizó la carga del dataset. Posterior a esto, al visualizar el dataset a través de Pandas [11] y Plotly [12] es observable el comportamiento de las variables y la distribución de las mismas como se puede observar en la Figura 9.

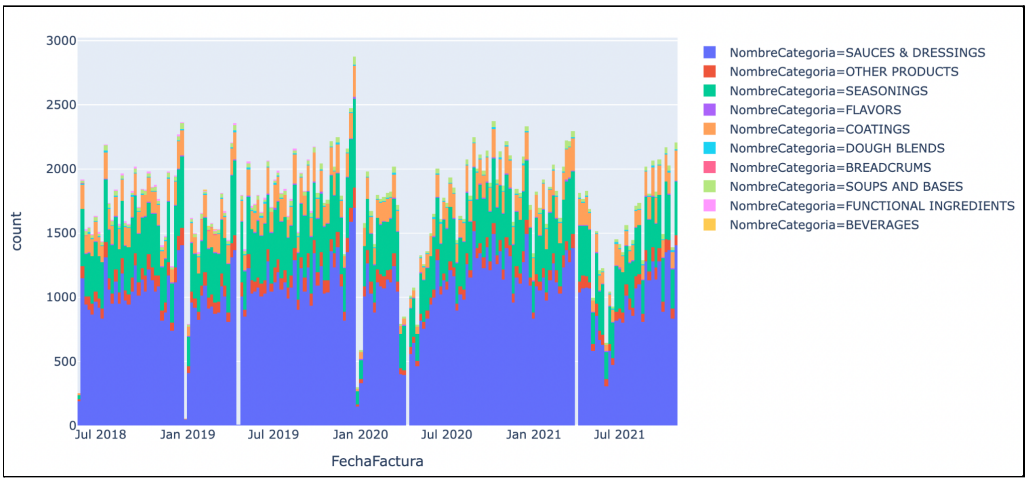

Fig 9. Visualización de dataset por categoría

Se puede apreciar en la Figura 6 que los productos más vendidos están en la categoría “Salsas y Aderezos” (en púrpura), seguido de “Condimentos” (en verde). Esta información se puede emplear para conocer la distribución de datos por categoría y así centrarnos en las que aquellas que presenten mayores registros y consecuentemente, poderles aplicar el método.

Una de las dificultades que se presentaron fue el tamaño del dataset procesado que serviría de insumo para el modelo de pronóstico. Dado que el tamaño del dataset excedió los topes de almacenamiento permitidos por GitHub, este se dividió en cuatro partes.

Para esta primera iteración donde se buscaba una exploración inicial del dataset se usó una regresión con Random Forest del cual se obtuvieron los siguientes resultados:

- MAPE = 4.15%
- RMSLE = 13.79%

Dichos resultados son positivos en cuanto a error dentro del marco de la primera iteración donde el modelo utilizado no tuvo una mayor optimización o personalización enfocada en los datos que se emplearon.

La segunda iteración estuvo enfocada en la serie de tiempo de un solo producto, para esta iteración el preprocesamiento consistió en tener inicialmente el producto al que se iba a aplicar los modelos de pronóstico y posteriormente realizar un remuestreo en los datos para tener un solo valor por mes, dicho valor resultante de la suma de los valores contenidos en ese mismo mes pues originalmente los datos estaban con una periodicidad diaria y la necesidad de la empresa se situaba en que una predicción de días no iba a ser suficiente para tener anticipadamente las materias primas a precios bajos y con una cantidad pronosticada adecuada.

B. Validación

En la primera iteración se realizó una partición de dataset (70% para entrenamiento y 30% para pruebas). En la segunda iteración se realizó una partición del dataset (85% para entrenamiento y 15% para pruebas) para tener suficientes datos a evaluar en las métricas seleccionadas.

C. Iteraciones y evolución

Las iteraciones siguientes se desarrollaron probando diferentes configuraciones de hiper parámetros en el modelo, el uso de distintos modelos para observar su eficacia en la tarea principal de pronóstico.

La primera iteración consistió en utilizar todo el dataset tanto con las variables categóricas y numéricas sin distinguir entre productos para realizar un testeo de pronóstico general de ventas de toda la compañía, en la Figura 4 de la sección III tenemos la visualización de los componentes de dicho dataset donde no se encuentra un patrón de estacionalidad a resaltar, por esto mismo se decide realizar el Augmented Dickey-Fuller (ADF) Test [13] que nos permite por medio de una prueba de hipótesis verificar si la estacionalidad es estadísticamente significativa en la serie de tiempo. Para nuestro caso la hipótesis nula la cual sugiere que la serie de tiempo no tiene una estacionalidad importante se rechaza siempre que el valor P sea menor a 0.05.

```
Serie de tiempo con dataset original, primera iteración

ADF Statistic: -1.768323
p-value: 0.396298
Critical Values:
        1%: -3.646
        5%: -2.954
        10%: -2.616
```

Fig 10. ADF Test para dataset de la primera iteración

El valor P es de 0.396 (ver Figura 10) lo cual no nos permite rechazar la hipótesis nula y la serie de tiempo de la primera iteración no contiene una componente importante de estacionalidad que pueda ser aprovechada por algunos modelos de pronóstico de series de tiempo.

Para esta primera iteración el modelo utilizado fue un Random Forest Regressor, que teniendo como insumo el dataset original entero sin filtro de producto obtuvo un puntaje de precisión del 99% en el conjunto de entrenamiento y de validación, estos resultados no permiten asegurar que sea un buen modelo y es más propenso a que haya overfitting.

Por su parte, para la segunda iteración el ADF Test presentó el siguiente resultado:

```
Serie de tiempo con dataset preprocesado, segunda iteración

ADF Statistic: -6.478034
p-value: 0.000000
Critical Values:
        1%: -3.616
        5%: -2.941
        10%: -2.609
```

Fig 11. ADF Test para dataset de la segunda iteración

El valor P para el dataset preprocesado de la segunda iteración es muy pequeño (ver Figura 11) y nos permite rechazar la hipótesis nula afirmando que hay una componente de estacionalidad estadísticamente significativa con la serie de tiempo del producto con código 688706.

Parte del estudio se dedicó al rastreo de métricas y modelos de predicción que pudieran ser de utilidad para el dataset seleccionado. Entre las opciones encontradas se seleccionó la librería DARTS (Librería especializada en la manipulación y predicción de series de tiempo).

La segunda iteración se basó en el uso de esta librería DARTS: el preprocesamiento, los modelos regresivos de series de tiempo y sus configuraciones óptimas, además de los buenos resultados en los últimos pronósticos, permitieron afianzar el uso de esta herramienta en contraste con métodos más tradicionales.

Algunos de los modelos candidatos tuvieron un proceso para encontrar los parámetros más óptimos con la serie de tiempo usada en la segunda iteración. En el caso del modelo Theta se realizaron pruebas con su parámetro en un rango de 40 valores entre -2 y 2 como se observa en la Figura 12 obteniendo el valor Theta con menor error siendo este 0.153:

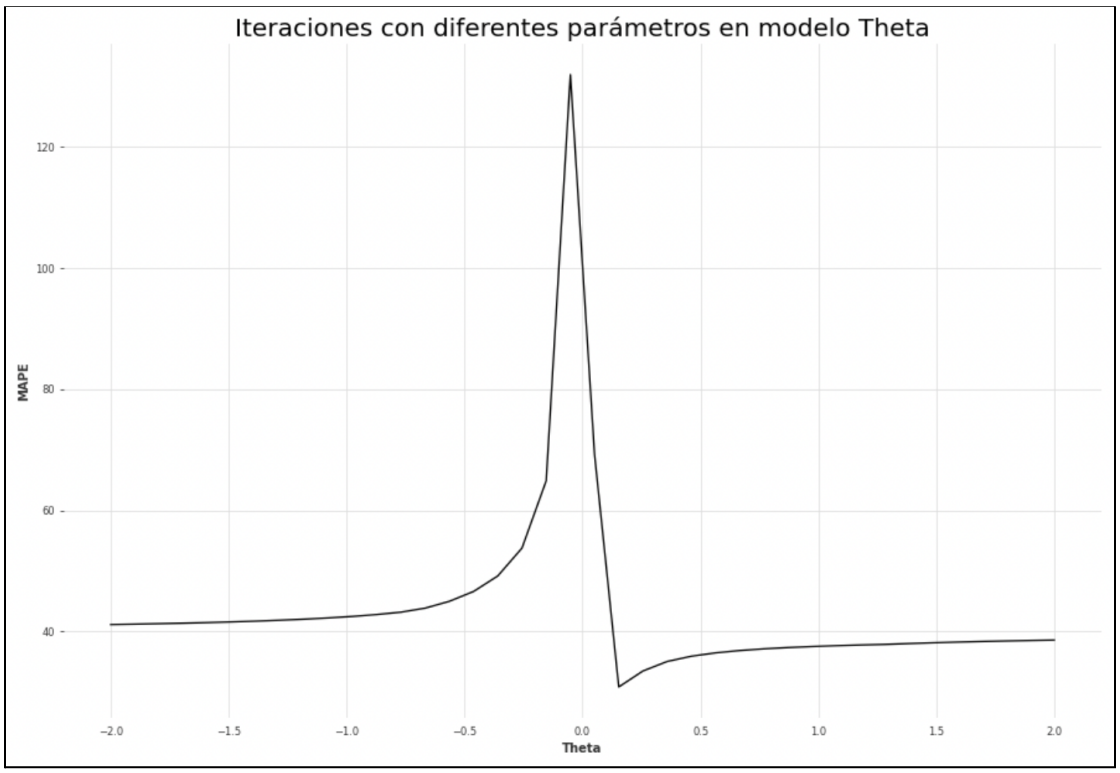

Fig 12. Error MAPE con diferentes valores de Theta

Así mismo el modelo ARIMA tuvo su optimización de los tres parámetros que lo componen por medio del uso de la función AutoARIMA de la librería DARTS que busca la combinación que mejor resultados de pronóstico ofrece para nuestro conjunto de datos como se observa en la Figura 13, el resultado arrojado es que el modelo ARIMA óptimo fue de AR(1), I(2) y MA(1):

```
ARIMA(1,2,1)(0,1,0)[4]             : BIC=inf, Time=0.09 sec
ARIMA(0,2,0)(0,1,0)[4]             : BIC=652.667, Time=0.01 sec
ARIMA(1,2,0)(1,1,0)[4]             : BIC=631.651, Time=0.04 sec
ARIMA(0,2,1)(0,1,1)[4]             : BIC=624.707, Time=0.06 sec
ARIMA(0,2,1)(0,1,0)[4]             : BIC=625.904, Time=0.05 sec
ARIMA(0,2,1)(1,1,1)[4]             : BIC=627.875, Time=0.09 sec
ARIMA(0,2,1)(0,1,2)[4]             : BIC=627.676, Time=0.07 sec
ARIMA(0,2,1)(1,1,0)[4]             : BIC=625.928, Time=0.05 sec
ARIMA(0,2,1)(1,1,2)[4]             : BIC=630.716, Time=0.16 sec
ARIMA(0,2,0)(0,1,1)[4]             : BIC=643.164, Time=0.03 sec
ARIMA(1,2,1)(0,1,1)[4]             : BIC=619.987, Time=0.06 sec
ARIMA(1,2,1)(1,1,1)[4]             : BIC=623.234, Time=0.10 sec
ARIMA(1,2,1)(0,1,2)[4]             : BIC=623.214, Time=0.09 sec
ARIMA(1,2,1)(1,1,0)[4]             : BIC=621.421, Time=0.07 sec
ARIMA(1,2,1)(1,1,2)[4]             : BIC=625.550, Time=0.15 sec
ARIMA(1,2,0)(0,1,1)[4]             : BIC=628.485, Time=0.05 sec
ARIMA(2,2,1)(0,1,1)[4]             : BIC=620.081, Time=0.10 sec
ARIMA(1,2,2)(0,1,1)[4]             : BIC=625.624, Time=0.08 sec
ARIMA(0,2,2)(0,1,1)[4]             : BIC=625.055, Time=0.08 sec
ARIMA(2,2,0)(0,1,1)[4]             : BIC=621.786, Time=0.06 sec
ARIMA(2,2,2)(0,1,1)[4]             : BIC=623.755, Time=0.19 sec
ARIMA(1,2,1)(0,1,1)[4] intercept   : BIC=623.014, Time=0.20 sec

Best model:  ARIMA(1,2,1)(0,1,1)[4]
```

Fig 13. Optimización del modelo ARIMA con AutoARIMA

El modelo de Naive Drift + Seasonal no tiene una configuración de parámetros pues extrae la componente de tendencia y estacionalidad de la serie de tiempo para realizar sus pronósticos, finalmente el modelo de Random Forest no tuvo una optimización profunda pues a este modelo se le definió los parámetros e hiper parámetros con el tanteo de diferentes combinaciones terminando con la que mejor resultado arrojó siendo este un número de estimadores de 10 y una profundidad máxima de cada árbol de decisión de 5.

D. Herramientas

En el proyecto se usaron principalmente las siguientes herramientas, todas en el lenguaje Python 3:

Procesamiento de datos y visualización:

- Sklearn.
- Pandas.
- Numpy [14].
- Matplotlib [15].
- Plotly.

Análisis estadístico, métricas y modelos de predicción:

- Statmodels [16].
- Sklearn.
- DARTS

.

VI. RESULTADOS

A. Métricas

Los resultados de las últimas iteraciones (ver Tabla II) permitieron obtener las siguientes métricas para los modelos elegidos en el proyecto.

TABLA II
RESULTADO DE LAS MÉTRICAS CON LAS PREDICCIONES POR MODELO

Métrica	Theta	ARIMA	Naive	Random Forest
MAPE	30.861	26.535	28.161	43.369
RMSLE	0.384	0.312	0.327	0.411
Tiempo (seg.)	0.042	1.956	0.041	0.070

B. Evaluación cualitativa

Desde los resultados obtenidos (ver Tabla II), el modelo con menor error en general es ARIMA. De los tres modelos restantes el que, según los errores, también tuvo buen desempeño fue el Naive. La Figura 14, presenta, de forma gráfica, los resultados del pronóstico de los modelos anteriores al contrastarlos con datos reales.

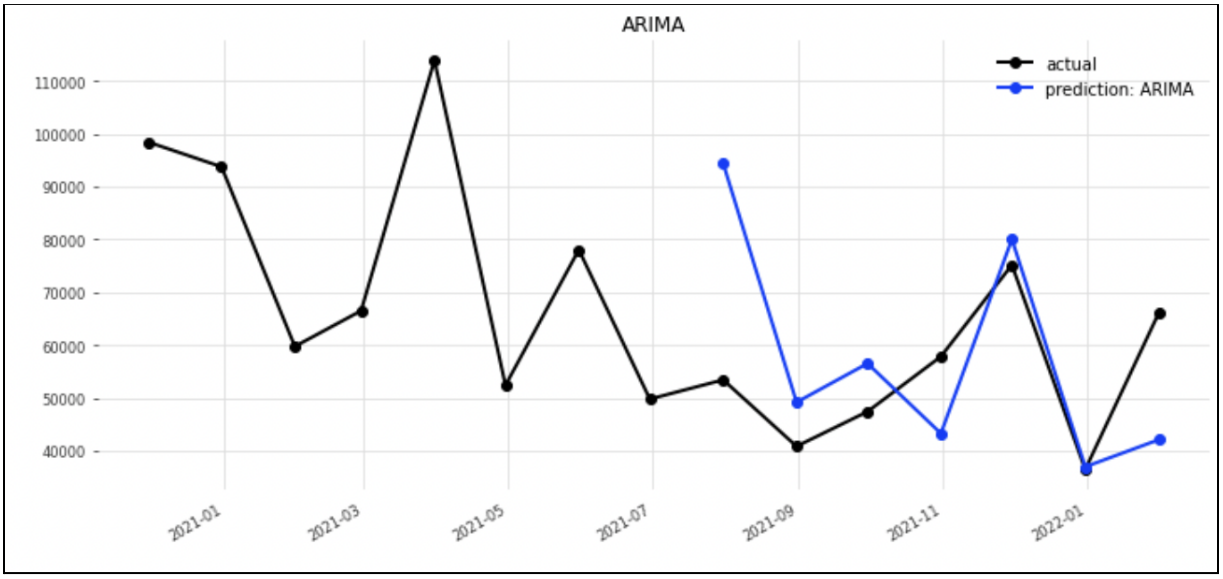

Fig 14. Resultado del pronóstico con el modelo ARIMA.

La predicción con el modelo de ARIMA ofrece un resultado bastante positivo. Sin embargo, al contrastarlo con el modelo Naive con sus componentes Drift y Seasonal, el cual en métricas de error estaba muy poco alejado de los resultados del primero, observamos lo siguiente en la Figura 15.

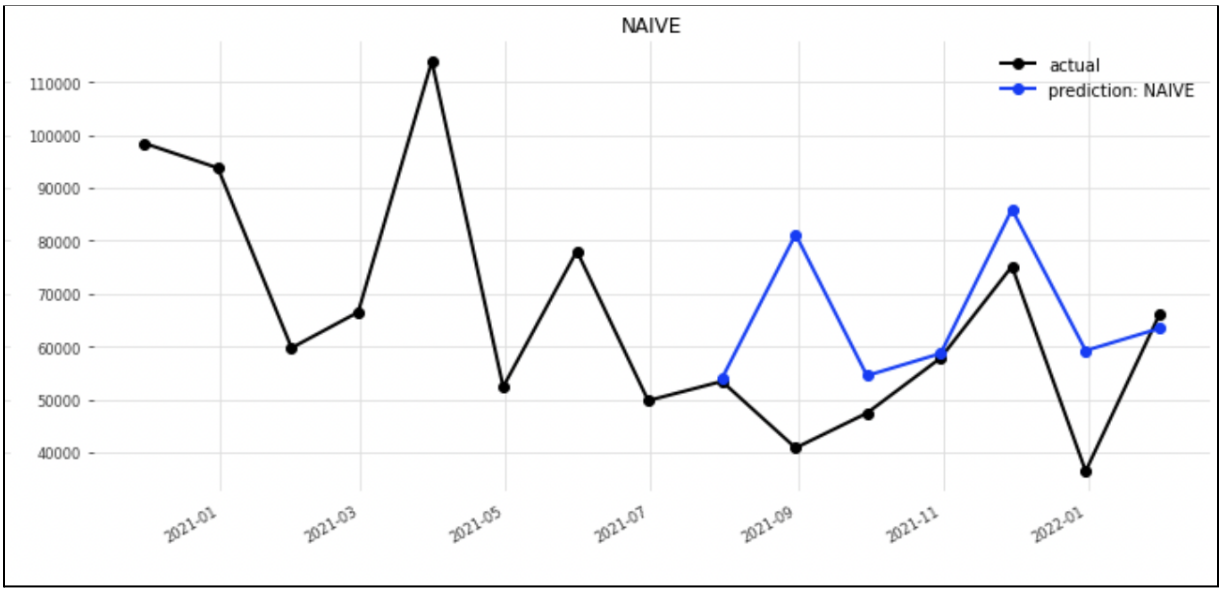

Fig 15. Resultado del pronóstico con el modelo NAIVE.

Ambos modelos dan buenos resultados pero en el ejercicio de elegir un modelo a implementar y desplegar en la empresa el modelo ARIMA es el que mejor puede aportar valor para solucionar el problema de negocio.

C. Consideraciones de producción

El modelo de predicción será utilizado por dos áreas: (i) la de ventas, para ir ajustando el cumplimiento frente a los presupuestos, y (ii) la de abastecimiento, para el plan de compras de las materias primas que se requieran.

La información con la cual contarán será:

- Pronóstico de ventas en kilos.
- Información histórica del producto a pronosticar.
- Presupuesto del producto.
- Gráficas de valores históricos y presupuestos.

El área de ventas se encuentra dividida en gerencias dependiendo del tipo de producto (cárnica, cuentas globales, etc.). Estas gerencias tienen sus comités, los cuales analizan el comportamiento de sus productos y ventas de los mismos. El análisis del comportamiento de los productos y ventas lo efectuarán basados en el comportamiento histórico, presupuestado y pronosticado. Con esta información, el área de ventas realizará los ajustes en caso de ser necesario para lograr el objetivo, ya sea mediante estrategias de mercadeo, promociones, etc.

Planeación de compras cuenta con un grupo que se reúne para estimar la demanda teniendo en cuenta estos resultados. Las variables a tener en cuenta en la toma de decisiones son: valor del dólar, tipo de proveedor (nacionales o extranjero), lead time, etc. con esta información procederán a evaluar los factores y realizar los ajustes a su balanced scorecard y tomar la mejor decisión apoyados con estos datos y la experticia desarrollada en su día a día laboral.

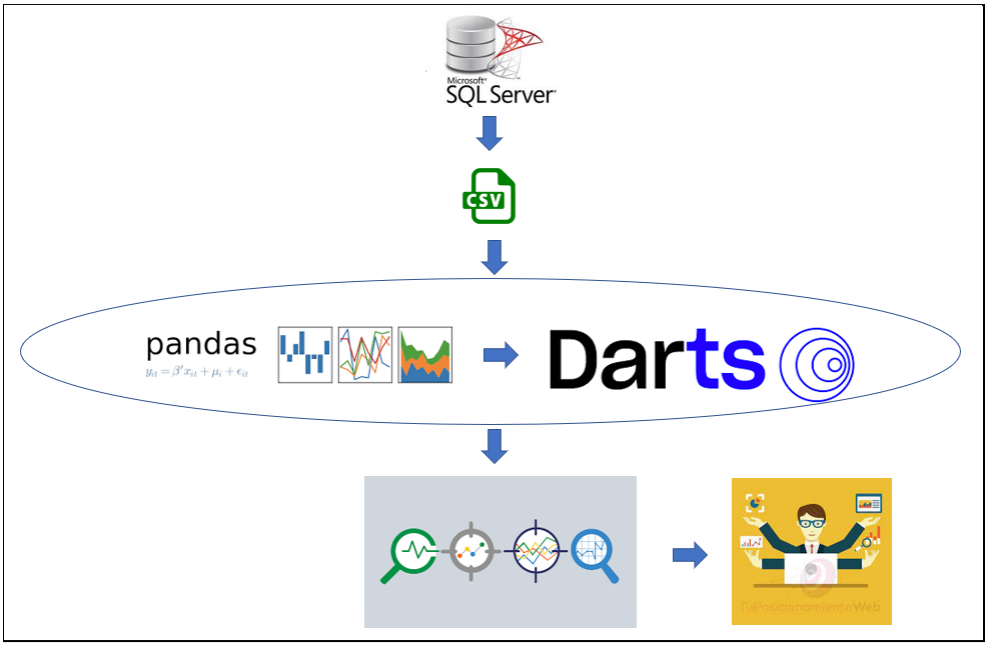

Fig 16. Flujo del proceso de modelamiento

VII. CONCLUSIÓN

Los modelos de pronóstico tienen un gran impacto en la empresa por sus aportes en la automatización de procesos que generan información más compacta y con ello un menor tiempo para la visualización del comportamiento de las ventas de los productos en el tiempo.

La aplicabilidad de este modelo requiere, por parte de parte de las personas implicadas, una actitud de asumir el modelo como una herramienta de apoyo para la toma de decisiones.

Igualmente se hace necesaria la creación de grupos de trabajo interdisciplinarios con el apoyo del área de informática para la administración del modelo y la extracción de la información. Asimismo, se sugiere generar un cambio de la cultura organizacional que está acostumbrada a utilizar recursos y software tradicionales para el análisis de ventas hacia estos nuevos métodos.

Se hace necesario, tras observar la funcionalidad de este modelo, el incluir variables geográficas y analizar sus comportamientos.

REFERENCIAS

[1] S. Prabhakaran, ARIMA Model – Complete Guide to Time Series Forecasting in Python. 2021. [En línea]. Disponible en: https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python

[2] Timothy Gallagher y Joseph Andrew, Administración Financiera, Segunda Edición, México: Prentice Hall, 2001.

[3] E. Lewinson, Choosing the correct error metric: MAPE vs. sMAPE. 2020. [En línea]. Disponible en: https://towardsdatascience.com/choosing-the-correct-error-metric-mape-vs-smape-5328dec53fac

[4] S. Saxena, What's the Difference Between RMSE and RMSLE?. 2019. [En línea]. Disponible en: https://medium.com/analytics-vidhya/root-mean-square-log-error-rmse-vs-rmlse-935c6cc1802a

[5] Scikit Learn, One Hot Encoder. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[6] G. Dudek, Short-term load forecasting using Theta method, Czestochowa University of Technology, Czestochowa, 2019.

[7] Darts, Baseline Models. [En línea]. Disponible en: https://unit8co.github.io/darts/generated_api/darts.models.forecasting.baselines.html

[8] N. Beheshti, Random Forest Regression. 2022. [En línea]. Disponible en: https://towardsdatascience.com/random-forest-regression-5f605132d19d

[9] J. Herzen, F. Lässig, S. Giuliano, T. Neuer, L. Tafti, G. Raille, T. Van Pottelbergh, M. Pasieka, A. Skrodzki, N. Huguenin, M. Dumonal, J. Kościsz, D. Bader, F. Gusset, M. Benheddi, C.

Williamson, M. Kosinski, M. Petrik, G. Grosch, Darts: User-Friendly Modern Machine Learning for Time Series, Unit8, 2021.

[10] Scikit Learn Machine Learning in Python. [En línea]. Disponible en: https://scikit-learn.org/stable/

[11] Pandas Python Data Analysis Library. [En línea]. Disponible en: https://pandas.pydata.org/

[12] Plotly: The front end for ML and data science models. [En línea]. Disponible en: https://plotly.com

[13] Statsmodels, Stationarity and detrending (ADF/KPSS). [En línea]. Disponible en: https://www.statsmodels.org/devel/examples/notebooks/generated/stationarity_detrending_adf_kpss.html

[14] Numpy. [En línea]. Disponible en: https://numpy.org

[15] Matplotlib — Visualization with Python. [En línea]. Disponible en: https://matplotlib.org

[16] Statsmodels, Statisticals models, hypothesis tests, and data exploration. [En línea]. Disponible en: https://www.statsmodels.org/dev/index.html