
# UTMACH

## FACULTAD DE INGENIERÍA CIVIL

## CARRERA DE INGENIERÍA DE SISTEMAS

# APLICACIÓN DE LA MINERÍA DE DATOS EN LA COMERCIALIZACIÓN DE INSUMOS AGRÍCOLAS

RAMON BRITO ALEX ANDRES
INGENIERO DE SISTEMAS

MACHALA
2020


# UTMACH

## FACULTAD DE INGENIERÍA CIVIL

## CARRERA DE INGENIERÍA DE SISTEMAS

# APLICACIÓN DE LA MINERÍA DE DATOS EN LA COMERCIALIZACIÓN DE INSUMOS AGRÍCOLAS

RAMON BRITO ALEX ANDRES
INGENIERO DE SISTEMAS

MACHALA
2020


# UTMACH

## FACULTAD DE INGENIERÍA CIVIL

## CARRERA DE INGENIERÍA DE SISTEMAS

TRABAJO TITULACIÓN
PROPUESTAS TECNOLÓGICAS

# APLICACIÓN DE LA MINERÍA DE DATOS EN LA COMERCIALIZACIÓN DE INSUMOS AGRÍCOLAS

RAMON BRITO ALEX ANDRES
INGENIERO DE SISTEMAS

MAZÓN OLIVO BERTHA EUGENIA

MACHALA, 17 DE DICIEMBRE DE 2020

MACHALA
2020

# APLICACIÓN DE LA MINERÍA DE DATOS EN LA COMERCIALIZACIÓN DE INSUMOS AGRÍCOLAS

INFORME DE ORIGINALIDAD

| 0% | 0% | 0% | 0% |
|---|---|---|---|
| INDICE DE SIMILITUD | FUENTES DE INTERNET | PUBLICACIONES | TRABAJOS DEL ESTUDIANTE |

FUENTES PRIMARIAS

| | | | |
|---|---|---|---|
| Excluir citas | Activo | Excluir coincidencias | < 30 words |
| Excluir bibliografía | Activo | | |

**CLÁUSULA DE CESIÓN DE DERECHO DE PUBLICACIÓN EN EL REPOSITORIO DIGITAL INSTITUCIONAL**

El que suscribe, RAMON BRITO ALEX ANDRES, en calidad de autor del siguiente trabajo escrito titulado APLICACIÓN DE LA MINERÍA DE DATOS EN LA COMERCIALIZACIÓN DE INSUMOS AGRÍCOLAS, otorga a la Universidad Técnica de Machala, de forma gratuita y no exclusiva, los derechos de reproducción, distribución y comunicación pública de la obra, que constituye un trabajo de autoría propia, sobre la cual tiene potestad para otorgar los derechos contenidos en esta licencia.

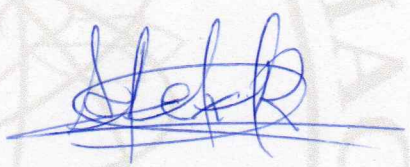
El autor declara que el contenido que se publicará es de carácter académico y se enmarca en las dispociones definidas por la Universidad Técnica de Machala.

Se autoriza a transformar la obra, únicamente cuando sea necesario, y a realizar las adaptaciones pertinentes para permitir su preservación, distribución y publicación en el Repositorio Digital Institucional de la Universidad Técnica de Machala.

El autor como garante de la autoría de la obra y en relación a la misma, declara que la universidad se encuentra libre de todo tipo de responsabilidad sobre el contenido de la obra y que asume la responsabilidad frente a cualquier reclamo o demanda por parte de terceros de manera exclusiva.

Aceptando esta licencia, se cede a la Universidad Técnica de Machala el derecho exclusivo de archivar, reproducir, convertir, comunicar y/o distribuir la obra mundialmente en formato electrónico y digital a través de su Repositorio Digital Institucional, siempre y cuando no se lo haga para obtener beneficio económico.

Machala, 17 de diciembre de 2020



RAMON BRITO ALEX ANDRES
0704655182

## DEDICATORIA

Dedico este trabajo principalmente a mi madre por ser mi fuente de inspiración diaria y por apoyarme durante todo mi proceso estudiantil a lo largo de mi vida, ya que gracias a ella he podido lograr todo lo que me he propuesto, celebrando conmigo los buenos momentos y dándome ánimo cuando lo he necesitado.

A mi hermano por siempre estar ahí cuando lo necesito y ayudarme a tomar decisiones cuando surgen problemas.

A todos mis familiares por aconsejarme en todo aspecto de mi vida y por siempre estar presente en los momentos difíciles.

**Alex Ramón**

## AGRADECIMIENTO

Agradezco primero a Jehová Dios por las bendiciones que me ha brindado, por la vida que me regala día a día y por cuidar siempre de todos nosotros.

A los docentes que han impartido su conocimiento a través de las exposiciones en los salones de clases y sus experiencias de vida en el mundo laboral.

A la Ing. Bertha Mazón quien ha sido mi tutora en el presente trabajo y me ha ayudado con todas las interrogantes que han ido surgiendo durante este período.

Finalmente agradezco a la Universidad Técnica de Machala por haberme brindado los recursos necesarios para poder crecer profesionalmente.

**Alex Ramón**

## RESUMEN

Debido a la gran cantidad de información generada diariamente en todo el mundo, es contradictorio pensar que existan empresas que sigan creyendo que los datos deben ser solo para almacenarlos en servidores o equipos especializados y no hagan uso de ellos para comprender su entorno, sin tener que invertir de manera excesiva. En este trabajo se muestra cómo realizar un análisis estadístico utilizando técnicas de minería de datos, que ayudan a visualizar patrones que a simple vista sería complicado detectar. El enfoque del proyecto se lo realiza en base a los datos de una comercializadora de insumos agrícolas, la cual, como muchas empresas en el país, no cuenta con un sistema gerencial que presente estadísticas sobre los datos de su propiedad.

La minería de datos es un conjunto de técnicas y modelos matemáticos que permiten analizar la información desde una perspectiva diferente, centrándose más que todo en extraer conocimiento oculto entre los datos, para ello es necesario hacer uso de una metodología que brinde una guía durante todo el proceso de desarrollo y ejecución de modelos. Para la elaboración de este proyecto se hizo uso de la metodología CRISP-DM especializada en este tipo de análisis para aumentar la probabilidad de éxito del proyecto, cuenta con seis fases que brindan un enfoque claro de proceso que se debe seguir para lograr el objetivo planteado.

La popularidad que ha ido adquiriendo la ciencia de datos en general, ha provocado que cada vez se creen más y mejores modelos estadísticos que permitan extraer más información a partir de los datos, es por ello que actualmente existen algoritmos predictores capaces de saber cuándo sucederá un evento conociendo las fechas exactas en que ha sucedido ese evento anteriormente. En el área de la comercialización, es muy importante conocer el estado financiero en que se encuentra una empresa, para tomar decisiones siempre en pro de la mejor continua, por esa razón en este proyecto, se hace especial énfasis en la predicción las ventas utilizando series temporales como ARIMA o Holt-Winters para encontrar el modelo que mejor se

adapta al negocio y predecir con certeza la economía futura de la empresa. Además, se utiliza otras técnicas como Reglas de asociación y la correlación para detectar patrones en otras áreas que permiten elaborar mejores estrategias. Es importante mencionar que estos análisis fueron realizados utilizando lenguajes de programación especializados en la estadística como Python y R, para agilizar el proceso de análisis y comparar que los modelos realizados se útiles no solo con los datos de prueba sino una vez sea puesto en producción y cuando se alimenten de más datos puedan ser capaces de ir aprendiendo y mejorando continuamente.

Una vez realizado con éxito los análisis propuestos es importante que estos se encuentren disponibles cuando sean necesarios, por ello se integró la información obtenida utilizando un sistema web desarrollado en Python con Django de tipo Dashboard o panel de control, en el que se compilan los gráficos generados y se brinda la facilidad al usuario de manipular los campos que requiera.

La intención de este trabajo es que más empresas en la provincia y el país, se interesen en el análisis de sus datos y de esa manera generar una cultura que ayude al país a generar mejores ingresos y desarrolle más su economía.

**Palabras claves:** Minería de datos, CRISP-DM, Series Temporales, Python, R, Comercialización de Insumos Agrícolas

## ABSTRACT

Due to the large amount of information generated daily around the world, it is contradictory to think that there are companies that continue to believe that data should only be stored on servers or specialized equipment and do not make use of it to understand their environment, without having to invest excessively. This work shows how to perform a statistical analysis using data mining techniques, which help to visualize patterns that would be difficult to detect at first glance. The approach of the project is carried out based on the data of an agricultural inputs marketer, which, like many companies in the country, does not have a management system that presents statistics on the data of its property.

Data mining is a set of mathematical techniques and models that allow information to be analyzed from a different perspective, focusing more than anything on extracting hidden knowledge among the data, for this it is necessary to use a methodology that provides guidance throughout the process of development and execution of models. For the development of this project, the CRISP-DM methodology specialized in this type of analysis was used to increase the probability of success of the project, it has six phases that provide a clear process approach that must be followed to achieve the proposed objective.

The popularity of data science in general has led to the creation of more and better statistical models that allow extracting more information from the data, which is why there are currently predictive algorithms capable of knowing when an event will happen. event knowing the exact dates that event has happened previously. In the area of marketing, it is very important to know the financial status of a company, to always make decisions in favor of the best continuity, for that reason in this project, special emphasis is placed on predicting sales using series temps such as ARIMA or Holt-Winters to find the model that best suits the business and to predict with certainty the future economy of the company. In addition, other techniques such as Association Rules and correlation are used to detect patterns in other areas that allow for better

strategies. It is important to mention that these analyzes were carried out using programming languages specialized in statistics such as Python and R, to speed up the analysis process and compare that the models made are useful not only with the test data but once it is put into production and When they are fed with more data, they may be able to continuously learn and improve.

Once the proposed analyzes have been successfully carried out, it is important that these are available when necessary, for this reason the information obtained using a web system developed in Python with Django of the Dashboard or control panel type was integrated, in which the graphics are compiled generated and the user is given the facility to manipulate the required fields.

The intention of this work is that more companies in the province and the country are interested in the analysis of their data and thus generate a culture that helps the country to generate better income and further develop its economy.

**Keywords:** Data Mining, CRISP-DM, ARIMA, Time Series, Python, R

# INDICE

DEDICATORIA ..... 1
AGRADECIMIENTO ..... 2
RESUMEN ..... 3
ABSTRACT ..... 5
GLOSARIO ..... 13
INTRODUCCIÓN ..... 15
1. CAPÍTULO I. DIAGNÓSTICO DE NECESIDADES Y REQUERIMIENTOS ..... 17
1.1. Ámbito de aplicación: Descripción del contexto y hechos de interés ..... 17
1.2. Establecimiento de requerimientos ..... 18
1.3. Justificación del requerimiento a satisfacer ..... 19
2. CAPÍTULO II. DESARROLLO DEL PROYECTO ..... 21
2.1. Definición del prototipo tecnológico ..... 21
2.2. Fundamentación teórica del prototipo ..... 22
2.2.1. Analítica de datos ..... 22
2.2.2. Ciencia de datos ..... 22
2.2.3. Aprendizaje automático ..... 23
2.2.4. Minería de datos ..... 23
2.2.5. Técnicas de clasificación ..... 24
2.2.5.1. Arboles de decisión ..... 24
2.2.5.2. Redes Neuronales Artificiales (ANN) ..... 25
2.2.6. Descriptivas ..... 26
2.2.6.1. Reglas de asociación ..... 26
2.2.6.2. Clustering ..... 27
2.2.6.3. Correlación ..... 28
2.2.7. Predictivas ..... 28

2.2.7.1. Series temporales ....28
2.2.7.2. Regresión lineal....31
2.2.8. Herramientas ....32
2.2.8.1. Python....32
2.2.8.2. *R* ....33
2.2.8.3. Jupyter ....34
2.2.8.4. RStudio ....35
2.2.8.5. Django....36
2.2.8.6. Pandas....37
2.3. Objetivos del prototipo ....37
2.3.1. Objetivo general ....37
2.3.2. Objetivos Específicos ....37
2.4. Diseño del prototipo ....38
2.4.1. Fase I. Comprensión del negocio....39
2.4.2. Fase II. Estudio y comprensión de los datos....40
2.4.3. Fase III. Análisis de los datos y selección de características ....42
2.4.4. Fase IV. Modelado....43
2.4.5. Fase V. Evaluación....44
2.4.6. Fase VI. Despliegue ....45
2.5. Ejecución y/o ensamblaje del prototipo ....45
2.5.1. Técnicas utilizadas ....45
2.5.1.1. Series temporales ....45
2.5.1.2. Preparación de los datos....45
2.5.1.2.1. ARIMA ....49

2.5.1.2.2. Holt-Winters ..... 53
2.5.1.2.3. Suavizado Exponencial ..... 55
2.5.1.2.4. Suavizado Exponencial y Autorregresión ..... 56
2.5.1.2.5. Suavizado Exponencial y ARIMA ..... 57
2.5.1.2.6. TBATS ..... 58
2.5.1.2.7. NNETAR ..... 58
2.5.1.3. Reglas de asociación ..... 59
2.5.2. Estadísticas ..... 64
2.5.2.1. Ventas ..... 64
2.5.2.2. Compras ..... 66
2.5.2.3. Clientes ..... 67
2.5.2.4. Productos ..... 67
*2.5.3. Dashboard* ..... 68
3. CAPÍTULO III. EVALUACIÓN DEL PROTOTIPO ..... 70
3.1. Plan de evaluación ..... 70
3.1.1. Evaluación de los modelos predictivos ..... 70
3.2. Resultados de la evaluación ..... 71
3.2.1. Evaluación de los modelos predictivos ..... 71
3.2.2. Comprobación de valores reales vs predichos ..... 72
CONCLUSIONES ..... 74
RECOMENDACIONES ..... 75
BIBLIOGRAFÍA ..... 76

## INDICE DE TABLAS

Tabla 1: Comparativa entre Python y R.....34

Tabla 2: Dataset de ventas .....41

Tabla 3: Dataset de compras .....42

Tabla 4: Fórmulas de técnicas aplicadas .....43

Tabla 5: Extracción de datos.....46

Tabla 6: Código para obtener valores estadísticos del modelo .....51

Tabla 7: Tabla de frecuencias de productos.....60

Tabla 8: Reglas de asociación - Algoritmo Apriori.....61

Tabla 9: Establecer reglas en algoritmo Apriori.....63

Tabla 10: Resultados de aplicación de algoritmo Apriori.....63

Tabla 11: Métricas de evaluación de modelos predictivos.....70

Tabla 12: Resultados de aplicación de métricas .....71

Tabla 13: Comparación de predicciones con valores reales.....72

## INDICE DE GRÁFICOS

Gráfico 1: Prototipo del proyecto ..... 21
Gráfico 2: Técnicas de minería de datos ..... 24
Gráfico 3: Árbol de decisión ..... 25
Gráfico 4: Tendencias y Estacionalidad de las curvas ..... 30
Gráfico 5: Logotipo de Python ..... 33
Gráfico 6: Logotipo de R ..... 33
Gráfico 7: Logotipo de Jupyter ..... 35
Gráfico 8: Logotipo de RStudio ..... 36
Gráfico 9: Arquitectura Django MVT ..... 36
Gráfico 10: Fases de metodología CRISP-DM ..... 38
Gráfico 11: Total de ventas (2018 – 2020) ..... 47
Gráfico 12: Resultados prueba Dickey-Fuller ..... 48
Gráfico 13: Resultados prueba KPSS ..... 48
Gráfico 14: Serie diferenciada una vez ..... 49
Gráfico 15: Autocorrelación y Autocorrelación parcial ..... 50
Gráfico 16: Resultados estadísticos del modelo ARIMA ..... 51
Gráfico 17: Resultados estadísticos del modelo SARIMAX ..... 52
Gráfico 18: Predicción de ventas con ARIMA en Python ..... 52
Gráfico 19: Predicción de ventas con ARIMA en R ..... 53
Gráfico 20: Serie temporal Holt-Winters en R ..... 54
Gráfico 21: Serie temporal Holt-Winters en Python ..... 55
Gráfico 22: Predicción en R utilizando STLF ..... 56
Gráfico 23: Predicción en R utilizando STLM y Autorregresión ..... 57

Gráfico 24: Predicción en R utilizando STLM y ARIMA .....57
Gráfico 25: Predicción en R utilizando TBATS .....58
Gráfico 26: Predicción en R utilizando NNETAR .....59
Gráfico 27: Distribución de productos por cantidad .....61
Gráfico 28: Conjunto de ítems por frecuencia .....62
Gráfico 29: Estadística - Ventas diarias del mes de Agosto 2020 .....65
Gráfico 30: Estadística - Top vendedores (2018 – 2020) .....66
Gráfico 31: Estadística - Ventas vs Compras vs Utilidad por año .....66
Gráfico 32: Estadística - Top clientes (2018 - 2020) .....67
Gráfico 33: Estadística - Top productos por cantidad (2018 – 2020) .....68
Gráfico 34: Dashboard - Diseño .....69
Gráfico 35: Dashboard - carga y visualización de figuras .....69

## GLOSARIO

**CRISP-DM:** Metodología orientada a proyectos de minería de datos, consta de 6 fases que son flexibles según el trabajo de investigación.

**Machine Learning:** Aprendizaje Automático en español, conjunto de técnicas que forman parte de la ciencia de datos, cuya principal característica son modelos capaces de aprender de información pasada para eventos futuros.

**KNN:** Algoritmo del vecino más cercano, forma parte de las técnicas de minería de datos y sirve para estimar valores predictores.

**SVM:** Máquinas de vectores de soporte, forman parte de los algoritmos de aprendizaje supervisado.

**PCA:** Análisis de componentes principales, técnica de minería de datos que permite encontrar describir la correlación entre variables.

**CART:** Es un modelo predictivo perteneciente a los árboles de decisión utilizado mayormente en machine learning.

**KDD:** Extracción de conocimiento en base de datos, metodología de minería de datos, utilizada muy frecuentemente.

**SEMMA:** Metodología centrada en el reconocimiento de patrones sobre fuentes de datos, utilizada en proyectos de minería de datos.

**ASCII:** Tipo de codificación de los datos, utilizado comúnmente en dispositivos electrónicos.

**IDE:** Entorno de Desarrollo Integrado, hace referencia al sistema que permite desarrollar código en un determinado lenguaje.

**MVC:** Patrón de diseño y arquitectura de programación que divide la estructura de un sistema en Modelo, Vista y Controlador.

**Statsmodels:** Librería que contiene modelos predefinidos de series temporales y otras técnicas de minería de datos en Python.

**Sklearn:** Librería que posee múltiples modelos de minería de datos para Python.

**AIC:** Criterio de información de Akaike, es una métrica que evalúa la calidad de un modelo estadístico

## INTRODUCCIÓN

Hoy en día, los sistemas transaccionales se encuentran en las pequeñas, medianas y grandes empresas, incluso algunos micronegocios cuentan con un sistema automatizado para llevar bajo control sus cuentas, esto hace que cada vez existan más datos almacenados, y en la mayoría de las ocasiones simplemente sirven para realizar consultas o visualizar documentos contables. En Ecuador existe la misma tendencia antes mencionada, muchas compañías almacenan sus datos, pero en muy pocas ocasiones los analizan; como es el caso de la empresa que por motivos de privacidad llamaremos ABC, dedicada a la comercialización de insumos agrícolas, actividad que en el país es muy común, por ello, es necesario encontrar una ventaja competitiva frente al resto de competidores. La minería de datos es un paso más allá al análisis superficial de información, consiste en aplicar una serie de técnicas y modelos matemáticos para encontrar patrones en las operaciones que permitan predecir eventos.

El objetivo de este proyecto es analizar los datos históricos generados por la empresa ABC utilizando técnicas de minería y gráficos estadísticos para brindar información sobre el estado actual de la empresa y pronosticar datos futuros. Además, parte del objetivo principal es transformar el actual proceso de reportes estadísticos mediante herramientas ofimáticas y automatizarlo con un sistema integrado que presente los gráficos de manera instantánea. También, se pretende incentivar a las demás empresas de la provincia a estudiar sus datos y poco a poco generar una cultura en el país donde la tecnología sea una herramienta de ayuda en cualquier sector y los datos adquieran el valor que realmente poseen.

Para el desarrollo del proyecto se hace uso de lenguajes de programación especializados en la estadística como son Python y R, ambos cuentan con una gran comunidad y capacidad de análisis que agiliza la comprensión de los datos. Asimismo, se hace uso de la metodología CRISP-DM para guiar el ciclo de vida del proyecto, organizándolo en fases sencillas de comprender.

Este documento se encuentra dividido en tres capítulos, los cuales contienen lo siguiente:

El **Capítulo I** detalla la problemática por la cual surge la idea de realizar un análisis estadístico sobre los datos del sector agrícola, también se establecen los requerimientos que debe cumplir el proyecto junto con las herramientas y técnicas que se utilizarán para el desarrollo de la solución.

El **Capítulo II** está conformado por la parte teórica de algunos los elementos que engloban la minería de datos, también se encuentra el desarrollo y aplicación de los modelos matemáticos sobre los datos y la implementación de la metodología.

El **Capítulo III** se enfoca en comprobar que las técnicas utilizadas sean correctas, evaluando los modelos con métricas, además de parametrizar la usabilidad del Dashboard realizado.

# 1. CAPÍTULO I. DIAGNÓSTICO DE NECESIDADES Y REQUERIMIENTOS

## 1.1. Ámbito de aplicación: Descripción del contexto y hechos de interés

La venta de insumos agrícolas en la provincia es una de las actividades más comunes debido a la alta demanda generada por los bananeros y ganaderos del sector, un gran porcentaje del comercio no solo de la provincia sino del país es abarcado por esta labor agrícola. En este año 2020 según la Asociación de Exportadores de Banano del Ecuador (AEBE) la exportación de banano obtuvo un incremento del 19,3% en ventas, quedando solo por detrás del petróleo como la principal fuente de ingresos del país.

Además, según la provincia de El Oro produce aproximadamente el 25% de todo el banano del país, ubicándose en la segunda posición superado solo por la provincia de Los Ríos con aproximadamente el 32%. Estos datos revelan la relevancia del sector agrícola en el país y por ende la importancia que tiene la comercialización de los insumos para su producción y el impacto directo sobre la economía.

Aunque según los datos existe una gran demanda para el consumo de insumos agrícolas en el país, también existe una gran oferta por parte de pequeñas y medianas empresas, por esa razón es cada vez más laborioso para los proveedores elaborar nuevas estrategias de ventas que atraigan más clientes que generen beneficio para la empresa.

Hoy en día los datos se han convertido en un capital valioso para las empresas debido a que estos pueden brindar información de consumo, comportamiento y preferencias de los clientes, incluso con un buen análisis permiten a las empresas predecir el rumbo del mercado y planificar en base a ello. Por las consideraciones anteriores, es evidente que todas las empresas deberían optar por analizar sus datos, pero existen muchos factores que los detienen siendo el principal de ellos el dinero.

Como se menciona en [1], existen muchas formas de analizar los datos y buscar relaciones entre las estrategias aplicadas y su impacto en las ventas de productos en específico y la relación con otros productos. Entre estas técnicas destacan la agrupación, clasificación y correlación de los datos que junto con la información obtenida de las ventas producen información valiosa que puede ser utilizada para mejorar las ventas. Todas estas técnicas y tecnologías utilizadas son abarcadas dentro

de la minería de datos y requieren de una comprensión y manipulación de los datos de manera correcta.

## 1.2. Establecimiento de requerimientos

Debido a la alta oferta existente en el mercado de insumos agrícolas cada vez es más complejo para las empresas captar o incluso retener a los que ya han sido sus compradores por mucho tiempo, el comportamiento cambiante de los clientes, las ofertas por parte de otros proveedores, el marketing y estrategias de ventas implementadas por sus competidores son algunas de las razones por la que muchas empresas fracasan luego de tener varios años en el mercado.

Todo negocio debe tener algo que lo diferencie de los demás y si en caso no se conoce aún, se debe ir poco a poco esclareciendo para poder ser explotado y lograr que sea su punto central y su mayor ventaja frente a la competencia, y la mejor manera de comprender esa ventaja competitiva es a través del análisis de los datos.

Las empresas generan datos que pueden ser analizados y estructurados para generar información sobre el estado económico o social de la organización con respecto a su entorno, sin embargo, muy pocos son los gerentes que invierten en este tipo de procesos, ya sea por falta de conocimiento o por temor a que esto represente un gasto para la empresa antes que una inversión.

Ante un mundo globalizado como el que se vive en la actualidad, es de vital importancia analizar todos los datos de nuestro entorno y medir los factores que puedan beneficiar o perjudicar a las empresas para poder llevar un control y tomar medidas ante las posibles amenazas que puedan existir.

Ante todo, lo descrito anteriormente es de vital importancia comenzar a analizar los datos generados por las transacciones diarias y encontrar patrones que permitan encontrar las fortalezas y debilidades de las empresas para que estas puedan reforzar y mejorar sus procesos en busca de la mejora continua. Para ello, no es suficiente con realizar análisis estadísticos que muestren los datos de forma gráfica, sino que hace falta de un sistema de extracción, y procesamiento de los datos para convertirlos en información que ayude a las empresas a conocer sus puntos débiles, sus puntos más

fuertes y sobre todo el comportamiento de sus clientes a lo largo del tiempo intentar predecir las acciones que estos realizarán en el futuro.

Todo este proceso se realiza utilizando técnicas avanzadas con los datos en donde se incluyen, ciencia de datos, aprendizaje automatizado, estructuras de procesamiento de datos en información, junto con otros procesos que permiten que los altos gerentes de las empresas puedan visualizar una información de manera sencilla.

### 1.3. Justificación del requerimiento a satisfacer

Los datos obtenidos en la empresa ABC, actualmente son simplemente almacenados en bases de datos y guardados como respaldos sin hacer uso de la información para realizar análisis estadísticos continuos, cada cierto tiempo se suele utilizar herramientas de ofimática para graficar ciertas estadísticas que interesan a la gerencia, pero no permite tener los datos visibles en el momento que lo requieren sino que debe hacerse todo un proceso que tarda días para poder ser presentado y luego analizado.

Con este proyecto se pretende automatizar esos procesos y ayudar a encontrar patrones en las ventas y la distribución de productos que permitan mejorar los procesos internos para llevar un mejor control y asegurar que la empresa tome decisiones correctas basados en la información que se genera día a día en las transacciones registradas en el sistema contable.

Para ello es necesario hacer uso de técnicas de análisis de datos como lo es la minería de datos, la cual abarca una serie de técnicas y procedimientos que permiten junto con un buen análisis encontrar un modelo que se ajuste a los datos de la empresa y pueda mostrar e incluso predecir los eventos futuros basado en los comportamientos históricos de los clientes y el sector de adquisiciones.

El objetivo de este trabajo consiste en analizar los datos generados en la venta de insumos agrícolas de la empresa ABC utilizando técnicas de minería de datos para la mejora en la toma de decisiones y obtención de ventajas competitivas prediciendo el futuro de la empresa y su situación actual en torno al mercado.

Para lograr este objetivo es esencial hacer uso de una metodología para guiar el proyecto de forma ordenada y clara, por ello se ha escogido la metodología CRIPS-DM, la cual especifica claramente las etapas a seguir y los procesos que se deben realizar en cada etapa. Además, para realizar este proyecto se hará uso de lenguajes de programación orientados a la estadística como lo es R y Python con ayuda de librerías especializadas para este campo, junto con el framework Django para desarrollar un Dashboard que permita visualizar de manera sencilla los gráficos que se requiera.

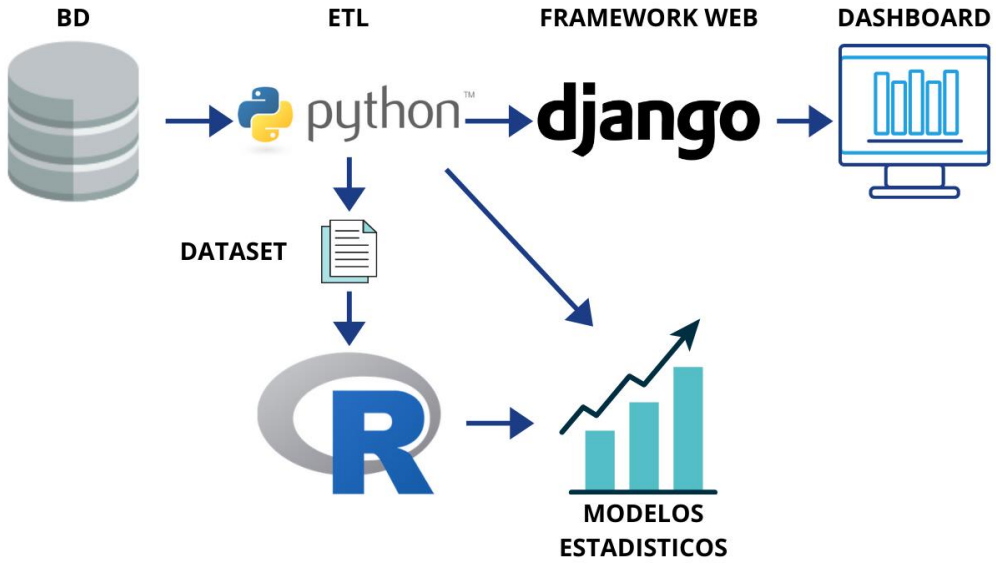
Al implementar este proyecto el concepto sobre los datos que el área administrativa de la empresa tiene actualmente cambiará, y ayudará a que poco a poco se le dé más importancia a la recolección de información, debido a que al analizar estos datos se obtiene mucha información no solo de la empresa sino también de los clientes e incluso del entorno en el que se encuentra.

# 2. CAPÍTULO II. DESARROLLO DEL PROYECTO

## 2.1. Definición del prototipo tecnológico

El proyecto consiste en un Dashboard desarrollado en Python utilizando el framework Django, donde muestre análisis descriptivos, predictivos y estadísticos realizados sobre los datos obtenidos de la venta de insumos agrícolas de la empresa de ABC.

*Gráfico 1: Prototipo del proyecto*




Fuente: Elaboración propia

Como se puede observar en el gráfico 1, para obtener los datos generados por el sistema contable se debe realizar un proceso de extracción y limpieza que será realizado a través de Python, se leerán los datos de una base de datos almacenada en el gestor PostgreSQL luego se procederá a analizar los datos procesados en R y Python para buscar la mejor opción y esa será la que se muestre al usuario que utilice el sistema.

El sistema estará conformado principalmente de 3 módulos generales, los cuales permitirán una interacción rápida y sencilla, esto módulos son los siguientes:

- Estadística: En este módulo se presentarán varias opciones en donde se podrá ver gráficos estadísticos que permitan visualizar las finanzas de la empresa.
- Descripción: Permitirá visualizar información relevante de los entes de la empresa como sus clientes, meses más productivos entre otros.

- Predicción: Contará con gráficos predictivos como series temporales y regresión lineal sobre las ventas de la empresa.

## 2.2. Fundamentación teórica del prototipo

### 2.2.1. Analítica de datos

El análisis de los datos es el conjunto de estrategias y procedimientos que pueden aplicarse sobre los datos en bruto para convertirlos en información que apoye los procesos de toma de decisiones basados en sus actividades anteriores [2], [3] .

Cade día que avanza es más común ver como las empresas utilizan el análisis de datos ya sea de manera empírica o con ayuda de entidades especializadas en ello, para medir su rendimiento, alcance, factibilidad y más aspectos que influyen directamente en sus ingresos y egresos. También utilizan estas técnicas para encontrar nuevos patrones de comportamientos de los entes de su entorno como los clientes, sus competidores e incluso factores ambientales que puedan representar un riesgo[4].

Un análisis de datos eficiente, según [5], debe tener en cuenta las siguientes características antes de ser puesto en marcha:

- Interpretable y escalable
- Capaz de utilizar todo tipo de datos
- Fácil de aplicar

### 2.2.2. Ciencia de datos

“El conjunto de principios fundamentales que apoyan y guían la extracción de información y conocimiento a partir de los datos”[6] es conocido como ciencia de datos. Por otro lado, debido a la gran cantidad de datos disponibles hoy en día, es normal que las industrias intenten explotar los potenciales disruptivos del análisis de datos y el aprendizaje automático y al combinarla con las distintas disciplinas como la medicina, informática, ingeniería entre otras, da como resultado la creación de una nueva disciplina conocida como ciencia de datos [5], [7].

La ciencia de datos es un campo multidisciplinario de la ciencia que se basa en métodos matemáticos, estadísticos, información y procesos informáticos con el fin de

generar nuevos conocimientos con los datos históricos disponibles[8]. Es importante recalcar que, aunque los términos información, conocimiento e ideas son parecidos, no significan lo mismo, pero si tienen una fuerte conexión entre ellos, y es vital saber diferenciales para el análisis [9], [10].

### 2.2.3. Aprendizaje automático

Se conoce como aprendizaje automático (machine learning) a la agrupación de métodos computacionales que crean conocimiento utilizando algoritmos inteligentes que pueden ser retroalimentados cada vez con más información [11].

Existen dos tipos de aprendizaje en machine learning, supervisado y no supervisado, el primero está centrado en la detección, clasificación y discriminación, estos tienen una serie de métricas que establecen límites entre las clases, definen el dominio de las muestras y dividen dichas muestras en función de las variables, algunos métodos conocidos de este tipo de aprendizaje son: vecino más cercano (KNN), máquina de vectores de soporte (SVM) y redes neuronales artificiales (ANN)[12]; algunos ejemplos de aprendizaje no supervisado en cambio son: análisis de componentes principales (PCA), análisis de conglomerados (CA) [11], [13].

### 2.2.4. Minería de datos

La minería de datos es un término muy amplio que abarca una serie de procesos cuya finalidad es descubrir patrones, semejanzas o vínculos entre los datos con el fin de que se conviertan en información valiosa al cohesionar de forma ordenada dichos datos. Según [14] la minería de datos puede darse con un aprendizaje supervisado, no supervisado o por refuerzo aunque estás formas de extracción y clasificación de datos se conoce comúnmente como aprendizaje automático o machine learning (ML) [15], [16].

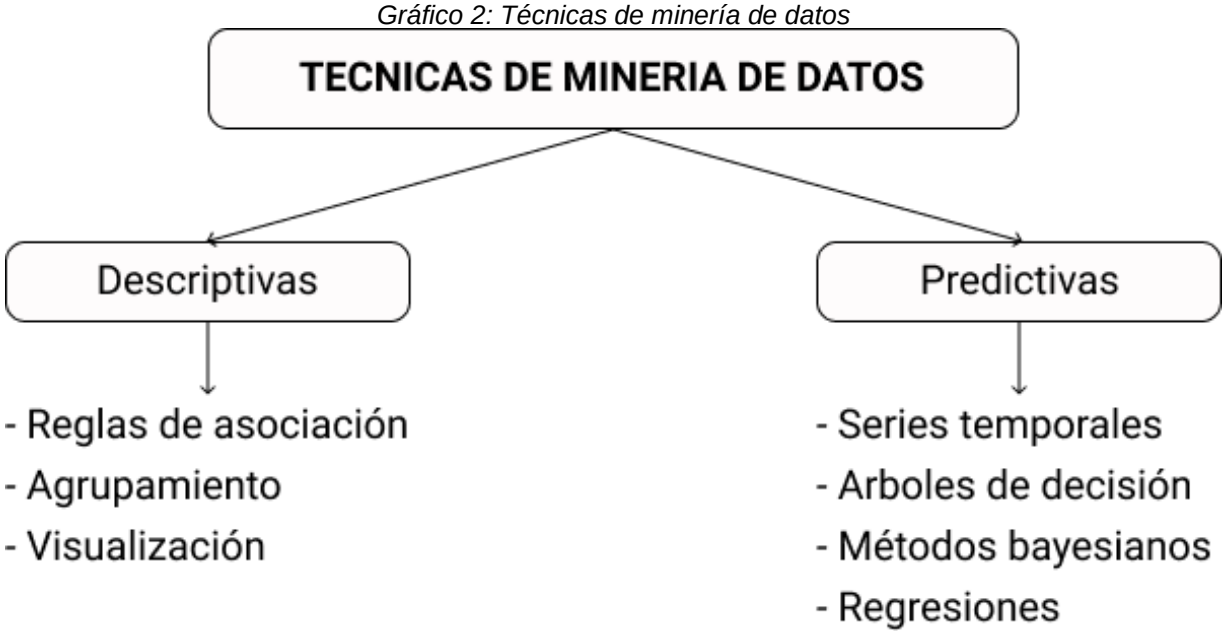
La minería de datos que se utiliza generalmente se divide en 4 tipos de conocimiento según [17]:

- Generalizado: Aquel que describe las características de las categorías de manera abstracta.
- Relacionado: Muestra las dependencias entre los eventos

- Clasificación: Refleja características comunes y diferentes entre los datos.
- Predictivo: Predice datos futuros basado en datos históricos o pasados.

En el gráfico 2 se muestra la división común que se realiza en los tipos de minería de datos.

*Gráfico 2: Técnicas de minería de datos*




Fuente: Elaboración propia

### 2.2.5. Técnicas de clasificación

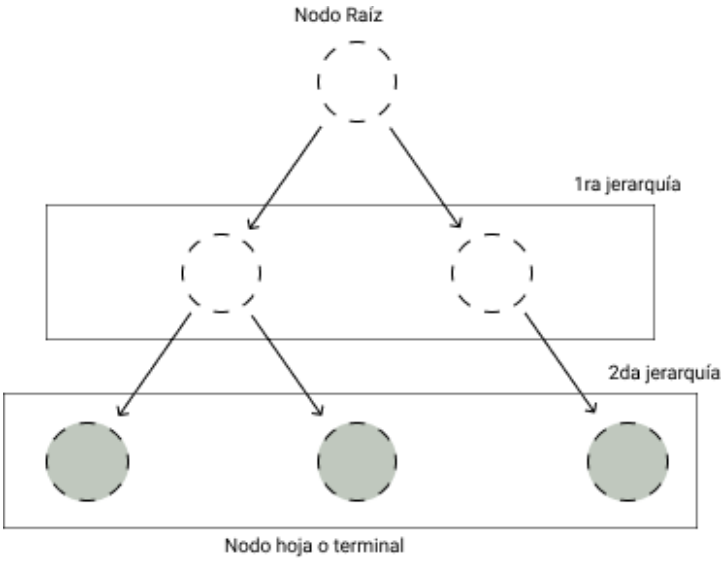
#### 2.2.5.1. Arboles de decisión

Un árbol de decisión es uno de los algoritmos más comunes utilizados para la selección de características, utilizado ampliamente en machine learning y la minería de datos. Existen varios métodos para la construcción de árboles entre los más utilizados se encuentras ID3, C4.5 y CART. El algoritmo C4.5 es en realidad una mejora de ID3 que permite seleccionar características más definidas, y CART permite el manejo de características con más valores que los dos algoritmos anteriores [18].

Los árboles de decisión derivados del proceso de minería de datos supervisado, también pueden actuar como modelos predictivos aplicando las características y estructura adecuada. Estos árboles por lo general basan sus decisiones mediante la clasificación de objetos, los niveles de sus atributos o los criterios fijados en cada nodo; en donde cada clasificación de un objeto crea una nueva rama, la división producida

se le llama nodo y al nodo principal se le conoce como raíz de árbol, los últimos nodos llevan el nombre de nodo terminal u hoja. Todos los nodos deben estar conectados con el nodo raíz ya sea directamente o a través de nodo intermedios, a los niveles de los nodos se le conoce como jerarquía y pueden existir tantas como sea necesario [19]. En el gráfico 3 se muestra la estructura y componentes de un árbol de decisión.

*Gráfico 3: Árbol de decisión*




Fuente: Elaboración propia

### 2.2.5.2. Redes Neuronales Artificiales (ANN)

Existe una inmensidad de redes neuronales, para efectos de este trabajo se hablará sobre las redes neuronales artificiales debido a que son las que más se apegan a la investigación de este trabajo. Este tipo de redes neuronales pertenecen al grupo de aprendizaje automático supervisado ya que para poder clasificar o predecir los valores solicitados se debe introducir un conjunto de observaciones con valores concretos, para poder entrenarlas. Las redes artificiales tienen como principal ventaja que pueden manejar grandes conjuntos de datos comúnmente con valores de entrada relativamente imprecisos y generar robustos [20].

Las ANN asemejan su comportamiento a la hora de aprender a la de un cerebro, utilizan modelos matemáticos no lineales para replicar este comportamiento debido a la simplicidad y flexibilidad de los mismos. Estas redes están compuestas generalmente de dos componentes principales, las neuronas o también llamadas

nodos y las interconexiones o pesos. Asimismo, su funcionamiento es relativamente sencillo, las neuronas son las encargadas de procesar la información entrante, mientras que los pesos son conectados entre las diferentes neuronas [21].

Una de las redes neuronales más comunes de este tipo son las redes neuronales perceptrón multicapa (MLP). Las MLP constan de tres componentes básicos: capa de entrada, una o varias capas ocultas y capa de salida. Aunque su estructura parezca sencilla, la forma en que trabajan en bastante compleja, en la capa de entrada poseen el mismo número de neuronas que la cantidad de parámetros ingresados; el número de neuronas de la capa de salida corresponderá exclusivamente a la definición del modelo y sus valores de probabilidad oscilarán entre 0 y 1; por otro lado en las capas ocultas es donde ocurre todo el procesamiento de los datos, estas capas son las encargadas de conectar las neuronas de la capa de entrada, realizar el proceso solicitado y enviar esa información a la capa de salida, es por esa razón que el número de neuronas y capas ocultas que se utilice influirá directamente en el rendimiento y precisión del algoritmo [20], [21].

### 2.2.6. Descriptivas

#### 2.2.6.1. Reglas de asociación

Las reglas de asociación son utilizadas comúnmente para determinar los niveles de asociación existentes entre varios factores, dando como resultado evidencia de probables relaciones de causa y efecto. Esta técnica de la minería de datos se la utiliza para encontrar patrones ocultos en una gran cantidad de datos, y ha tenido grandes casos de éxito en diversos campos como mercados, sitios web, recomendación de ofertas y más [22].

Para medir su efectividad en el desarrollo de un modelo existen diversas técnicas que se pueden utilizar como el chi-cuadrado, prueba de bondad de ajuste, prueba de independencia y la prueba de homogeneidad [23].

Existen las reglas de asociación redundantes las cuáles se pueden diferenciar entre positivas y negativas, este tipo de reglas han tenido gran presencia actualmente en la minería web, sistemas de recomendación y detección de intrusos. Existen formas de validar si una regla es válida o no, por ejemplo, el umbral final es menor que el umbral

definido por el usuario lo mismo aplica con el porcentaje de confianza del modelo definido [24].

Las reglas de asociación son técnicas que permiten descubrir relaciones entre una gran cantidad de datos, en donde los conjuntos de elementos poseen dos lados, el lado de la mano izquierda (LHS) o *X* y el lado de la mano derecha (RHS) o *Y*. Las reglas de asociación son aquellas que cumplen con los valores del soporte mínimo y la confianza mínima. El soporte de una regla de asociación es la proporción con que una regla aparece en el conjunto de datos, mientras que la confianza es la frecuencia con que el ítem aparece en el conjunto de datos [25] [26].

Uno de los obstáculos de las reglas de asociación es la capacidad de cómputo frente a la cantidad de datos, debido a que incluso en un pequeño conjunto de datos la cantidad de reglas que se crean son muchas [27]. Por esa razón, es que con el pasar del tiempo se han creado nuevos métodos para encontrar reglas de asociación que faciliten el procesado de datos.

Uno de los valores determinantes en una regla se asociación es el valor de Lift, debido a que este valor permite detectar si la confianza de un conjunto de datos con respecto a otro ítem aumenta o simplemente es el complemento del conjunto de datos, mientras más alejado del 1 mejor será esa regla, caso contrario indicaría que los productos no tienen correlación según la regla aplicada [26].

#### 2.2.6.2. Clustering

La agrupación es una de las técnicas más utilizadas en el análisis de datos, ya que permite ver y descubrir comportamientos comunes en los datos, además, es útil para generar parámetros de otro tipo de algoritmos de predicción como redes neuronales o sistemas difusos. Hoy en día, los clústeres tienen usos diversos en el descubrimiento de patrones, detección de fallos o en los sistemas de recomendación [28].

Uno de los algoritmos de clustering más comúnmente utilizado es K-medias, este algoritmo está basado en la distancia de sus elementos y es factible de utilizar debido a su bajo costo en procesamiento computacional; uno de los requisitos para utilizar este método es tener conocimiento del número de grupos que se va a identificar que generalmente se le denomina *k*. Existen otros clústeres más complejos que reclasifican

la información en micro-clusteres y luego de un procesamiento agrupan la información en los clústeres finales requeridos. El algoritmo de CluStream también separa la información en micro-clusteres pero además permite una actualización incremental de los mismos, siendo más eficiente en la agrupación de datos [28].

En [29] proponen un nuevo algoritmo que denominan minería de secuencias de clúster (CSM) donde consideran el orden de ocurrencia de los eventos y los intervalos de tiempo como medidas de agrupamiento. Este algoritmo consta de tres pasos generales que son: Generar los candidatos, evaluar los patrones seleccionados y excluir patrones superpuestos.

#### 2.2.6.3. Correlación

La correlación entre dos variables en términos estadísticos es el valor de la covarianza estandarizada, obtenida normalizando la covarianza con la desviación estándar de cada variable. El coeficiente de correlación varía su resultado de 0 a 1, siendo 1 la correlación perfecta de los datos y 0 la no existencia de relación entre los datos [30].

El coeficiente de correlación es utilizado como una comprobación previa antes de aplicar algún método de machine learning o minería de datos, ya que permite conocer si los datos a analizar pueden poseen alguna relación o siguen algún patrón que pueda ser utilizado.

Según el diccionario Merrian-Webster define a la correlación como una relación entre las variables matemáticas que tienden a variar o asociarse de una manera no esperada por casualidad, un ejemplo de esto se observa en la medicina, donde los signos vitales presentas correlaciones de diversos tipos [31].

### 2.2.7. Predictivas

#### 2.2.7.1. Series temporales

Una serie temporal es una sucesión de datos ordenados de manera cronológica, cuya finalidad es en base a los datos de entrada generar una predicción de los mismos a un intervalo de tiempo definido.

Existen muchos modelos de series temporales que se utilizan para predecir, a continuación, se detallan algunos de los más conocidos.

#### 2.2.7.1.1. ARIMA

ARIMA fue presentado por Box y Jenkins, es utilizado para modelar y analizar series temporales en múltiples campos [32]. un modelo es autorregresivo si su variable endógena se explica en el tiempo basado en los valores anteriores, agregando un valor de error. Un modelo Autorregresivo (AR) se denota como AR(p) donde *p* es el número de retrasos necesarios para encontrar $Y_t$, $\varepsilon_t$ representa el error. Un modelo de media móvil (MA) plantea $Y_t$ como una función de términos independientes y errores basados en los términos pasados, representados como MA(q), donde *q* es el número de retrasos incluidos en término del error. Sin embargo, aunque estos modelos pueden ser utilizados de manera independiente, estos modelos utilizan valores de los parámetros p y q muy altos, por ellos, se creó ARIMA que es una combinación de ambos modelos de manera que optimiza el uso de parámetros utilizándolos, haciéndolo ideal para todo uso; este modelo se representa de la forma ARIMA(p,d,q), donde p y q representan a los modelos AR y MA y d representa el orden de integración [1].

Box y Jenkins plantean una metodología a utilizar al implementar ARIMA la cual según [20] es:

1. Realizar la prueba Augmented Dickey-Fuller para probar la estacionariedad.
2. Seleccionar el modelo tentativo.
3. Escoger los parámetros aplicables al modelo.
4. Medir el ruido blanco utilizando la prueba Ljung-Box.
5. Analizar el modelo propuesto.

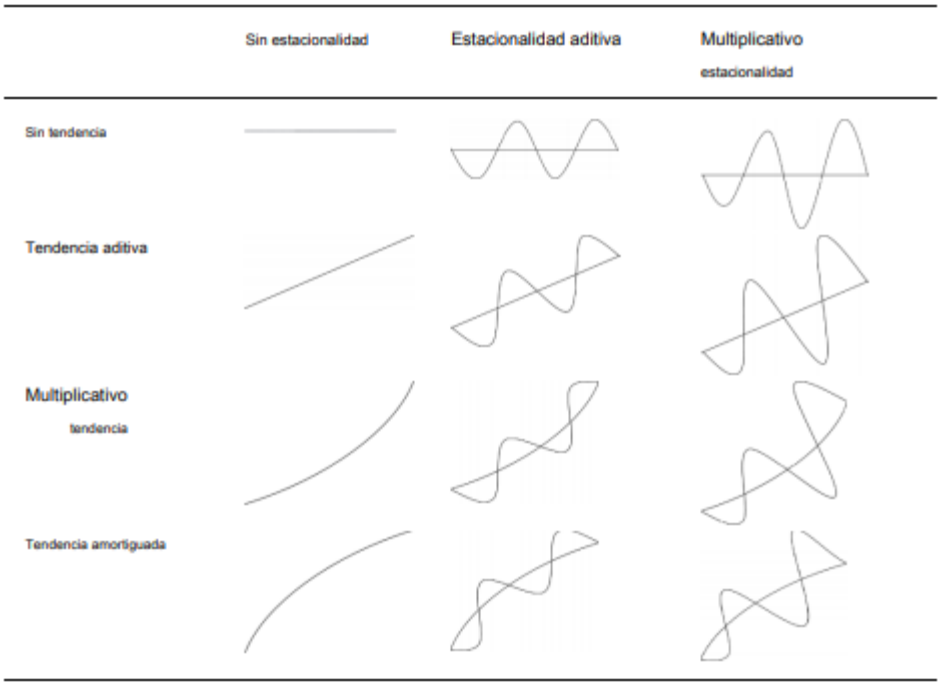
#### 2.2.7.1.2. SARIMA

Un modelo SARIMA trata a los datos de manera estacionaria utilizando la diferenciación estacional del orden apropiado. Al igual que ARIMA está compuesto de varias modelos, en este caso son los siguientes: Autorregresión (AR), Integración (I), Media móvil (MA), Autorregresión estacional (SAR), Integración estacional (SI), y media móvil estacional (SMA). La notación utilizada en este modelo es SARIMA (p, d, q) x (P, D, Q) [m], donde m es el periodo estacional [21].

#### 2.2.7.1.3. Holt Winters

La principal fortaleza de los modelos Holt-Winters es su cálculo de índices estacionales que, aplicado en un conjunto de datos de alta estacionalidad, presenta predicciones robustas y precisas. Puede ser clasificado en cuatro tipos principales: multiplicativo, aditivo, modificado y extendido según el comportamiento de los datos. Otra característica importante es que para realizar predicciones precisas no necesita una gran cantidad de datos [33]. Holt-Winters es parte de los modelos de suavizado exponencial, llamado también suavizado exponencial triple, se basa principalmente en la tendencia y estacionalidad y error de los datos; ideal para una serie de datos lineales [34]. Las características de una curva, en base al análisis de una serie temporal se visualiza en el gráfico 4.

*Gráfico 4: Tendencias y Estacionalidad de las curvas*




Fuente: Referencia [34]

### 2.2.7.1.4. Suavizado exponencial

Los modelos de suavizado exponencial son conocidos por ser robustos en las predicciones automáticas, además, son simples, confiables y estables lo que lo hace ideal para realizar grandes predicciones. Estos modelos analizan las series temporales, como un total del nivel, la tendencia y la estacionalidad, también poseen

otras características como saber si son aditivas, multiplicativas, lineales o no lineales [35].

Estas técnicas son tan flexibles que pueden ser utilizadas junto con otras para poder alcanzar una mayor precisión, por esa razón, existen varios artículos científicos en el que se proponen métodos innovadores que mezclan los modelos de suavizado exponencial, con teorías propias u otras técnicas conocida [36].

#### 2.2.7.1.5. TBATS

TBATS es modelo que fusiona la estacionalidad trigonométrica, transformación Box-Cox, errores ARMA, componentes estacionarios y tendencia, una mejoría del modelo BATS por la adición de la estacionalidad, el mejor criterio para medir la efectividad del modelo es el AIC. Estos modelos son conocidos por predecir conjuntos de datos que poseen múltiple estacionalidad, es capaz de predecir incluso con una cantidad mínima de datos dando un mejor rendimiento que BATS, en parte también debido a que está basado en la transformación de Fourier [37].

#### 2.2.7.1.6. NNAR

Las Redes Neuronales Autorregresivas han ido en aumento en los últimos tiempos por utilizar redes neuronales para realizar cálculos sobre las series temporales. Utilizan una estructura multicapa, con varias entradas y salidas; su notación se la realiza de la siguiente manera: NNAR(p, P, k) los cuales indican los números de capas que tiene las redes neuronales. En el lenguaje R la función obtiene el nombre de NNETAR [38].

### 2.2.7.2. Regresión lineal

La regresión lineal tiene como objetivo establecer la relación existente entre una variable de entrada o independiente y una variable de salida o dependiente. Esta técnica ha sido aplicada con éxito en diversas como las finanzas o la manufacturación de productos. Ha sido implementado en tantas áreas que han hecho necesario la creación de nuevas formas de regresión lineal, debido a las desviaciones provocadas por la imprecisión de los datos.

La forma más común de medir el éxito de un modelo de regresión lineal es obteniendo el valor del error cuadrado general entre los datos observados y los estimados, aunque

esto presenta dos inconvenientes principales: los supuestos estadísticos sólidos y la capacidad de predicción frente al ajuste excesivo [39].

### 2.2.8. Herramientas

#### 2.2.8.1. Python

Actualmente existen dos lenguajes de programación que de cierta manera compiten por ser el “mejor” en el campo de la ciencia y análisis de datos; estos lenguajes son Python y R, y según la problemática que se quiera solucionar se deberá escoger entre uno u otro.

Python es un lenguaje de programación creado por Guido van Rossum en los años 80 como resultado de un proyecto de hobby. Su logo se puede ver en el gráfico 6. Posee tres características que lo hacen destacar de los demás lenguajes como: Interpretado, escritura dinámica, alto nivel; al decir que es un lenguaje interpretado significa que las instrucciones escritas son directamente ejecutadas sin pasar por un proceso de conversión de todo el código al lenguaje de máquina, esto es una ventaja frente a otros lenguajes, aunque, sin duda las dos características más importantes que lo hacen un referente en el campo de la ciencia es su escritura dinámica y de alto nivel, debido a que el código escrito es fácilmente entendible en el dialecto común y sus variables no dependen de un tipo de dato específico [40].

Otra de las fortalezas que posee Python en el campo del análisis de datos, es la gran cantidad de librerías de estadística creadas por la comunidad o entes privados. Si bien es cierto R es un lenguaje desarrollado pensando exclusivamente en el campo de los datos, cada vez se crean nuevas librerías para abarcar más campos de análisis para Python. Un ejemplo podemos ver en [41], donde los autores desarrollan métodos exclusivos para realizar análisis estadísticos en el entorno hidrológico. Así como el anterior, cada día se desarrollan más herramientas que potencias el lenguaje y mejoran su rendimiento en más sectores.

*Gráfico 5: Logotipo de Python*


**Fuente:** Referencia [42]

Python posee una consola preinstaladas que permite desarrollar instrucciones sencillas de manera rápida, esto ayuda en el proceso de aprendizaje para comprender el funcionamiento del lenguaje, además que es útil para comprender errores de compatibilidad y conocer características del lenguaje.

La versión de Python con la que desarrolló este proyecto es la 3.7.9, que fue lanzada públicamente el 17 de agosto del 2020.

#### 2.2.8.2. *R*

R es un lenguaje y entorno de programación que permite analizar y graficar los datos de manera estadística, posee una gran variedad de métodos que permiten desarrollar técnicas como series temporales, regresiones, clustering y más de manera sencilla y sin utilizar muchas líneas de código como en otro lenguajes [43].

*Gráfico 6: Logotipo de R*


Fuente: Referencia [43]

Cuenta con una gran comunidad de desarrolladores que brinda actualizaciones periódicas y crea paquetes cada vez mejores y más potentes. Además, R cuenta con

un gran número de alojamientos alrededor del mundo para agilizar la descarga de paquetes para sus usuarios, esto lo hace a través de la Red Completa de Archivos R o CRAN por sus siglas en inglés.

El lenguaje R es muy versátil ya que puede ser utilizado para todo tipo de trabajo estadístico, además, con la ayuda de paquetes puede ser utilizado para programar interfaces web y trasladar los análisis realizados a un sistema dinámico con toda la potencia que brinda el entorno [44].

*Comparativa*

A continuación, en la tabla 4 se plantea algunas de las características más importantes para elegir conocer que lenguaje utilizar dependiendo del trabajo que se quiera realizar.

*Tabla 1: Comparativa entre Python y R*

| **Python** | **R** |
|---|---|
| Software Libre | Software Libre |
| Actualizaciones constantes | Actualizaciones constantes |
| Gran comunidad | Gran comunidad |
| Utilizado para todo tipo de análisis de datos | Principalmente usado para análisis estadísticos |
| Posee librerías o framework para desarrollo web, escritorio, móvil, etc | Posee paquetes para desarrollo web |
| Lenguaje de alto nivel | Lenguaje de bajo nivel |
| Multiplataforma | Multiplataforma |
| Fácil de aprender | Complejo de aprender sin experiencia |

Fuente: Elaboración propia

#### 2.2.8.3. Jupyter

Jupyter es un entorno de desarrollo de código libre, nació en 2014 bajo el proyecto IPython, hoy en día es uno de los IDE más utilizados por los científicos de datos. Su código se encuentra publicado enteramente en GitHub y es mantenido por su comunidad y desarrolladores constantemente [45].

La principal característica que posee Jupyter es su flexibilidad al ejecutar sentencias de códigos, ya que cada cuaderno o archivo está dividido en bloques que pueden variar su tamaño de una línea a las que uno quiera, ejecuta cada bloque de manera independiente, pero relacionando las variables, sin seguir un orden secuencial.

*Gráfico 7: Logotipo de Jupyter*


Fuente: Referencia [45]

Como se observa en la figura #8, Jupyter no se limita solo a código desarrollado en Python, sino que puede ser utilizado para escribir código en múltiples lenguajes, por ejemplo: R, PHP, C, Java, etc.

#### 2.2.8.4. RStudio

RStudio es un entorno de desarrollo integrado desarrollado exclusivamente para el lenguaje R, cuenta con una versión de código libre y otra comercial o privativa, posee la edición de escritorio o de navegador y ambas cuentan con las siguientes características: consola de comandos, resaltado de sintaxis, sección de gráficos, depuración y muchas otras herramientas que agilizan el desarrollo en R [46].

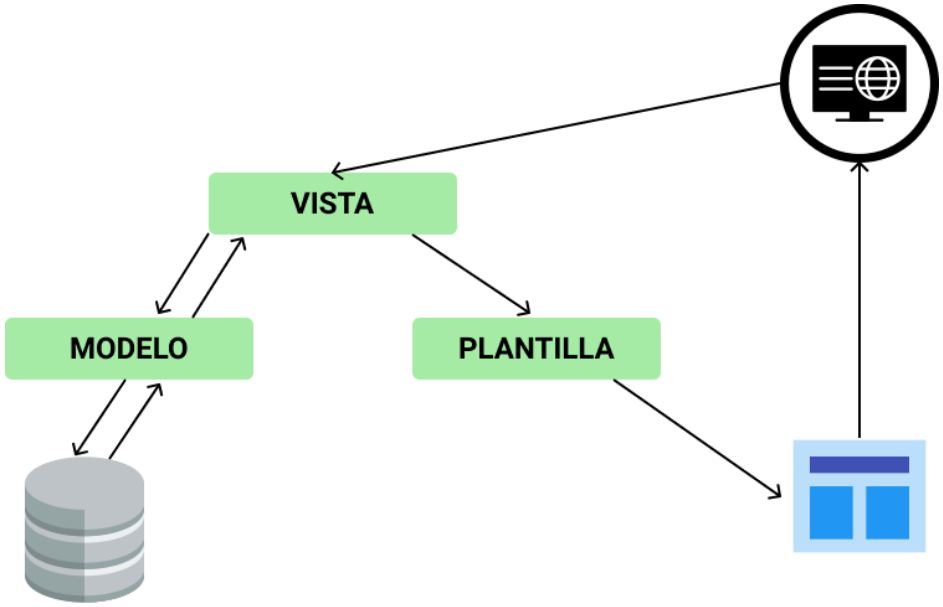
*Gráfico 8: Logotipo de RStudio*

Fuente: Referencia [46]

### 2.2.8.5. Django

Django es un framework de alto nivel desarrollado para Python, su principal característica es la velocidad con la que permite construir sitios web limpios. Es de código libre y maneja el patrón de diseño: modelo, vista, controlador; que es muy común en la programación orientada a objetos. Brinda módulos de seguridad que encriptan las sentencias y verifican la procedencia de paquetes para asegurar la confidencialidad de los datos de manera nativa [47].

*Gráfico 9: Arquitectura Django MVT*




Fuente: Elaboración propia

La arquitectura MVC, es utilizada por muchos programadores y framework's como Laravel, entre otros. Aunque Django utiliza una variación que es Modelo-Vista-Plantilla, gráfico 10. Este patrón de programación fue inicialmente introducido para el desarrollo de interfaces de usuario y luego fue implementada en la programación de aplicaciones,

en donde se introdujo los conceptos como se lo conoce actualmente. El modelo representa la parte lógica del negocio, la vista maneja lo que se le presenta al usuario y el controlador es el encargado de los cambios en las vistas solicitados por el usuario [48].

#### 2.2.8.6. Pandas

Pandas es una librería de Python creada en el 2008 por AQR Capital Managment y que en el 2009 pasó a ser de código libre y hasta el día de hoy es mantenida por usuarios de su comunidad alrededor del mundo que periódicamente la actualizan y añaden mejoras para potenciarla aún más. Su funcionamiento basado principalmente en convertir los datos ingresados en DataFrames, su agilidad y la gran cantidad de operaciones que se puede realizar con los mismos es una de las características principales que hacen de esta librería la mejor opción para el análisis de datos. Entre las características que más destacan sus autores sobre esta librería son los siguientes [49]:

- Manipulación de datos rápida y eficiente utilizando Dataframe.
- Variedad de herramientas para leer archivos de datos como csv, Excel, base de datos.
- Flexibilidad en el manejo de conjuntos de datos.
- Las columnas de datos pueden poseen una mutabilidad independiente.

## 2.3. Objetivos del prototipo

### 2.3.1. Objetivo general

Analizar los datos generados en la venta de insumos agrícolas de la empresa ABC utilizando técnicas de minería de datos para la mejora en la toma de decisiones y obtención de ventajas competitivas prediciendo el futuro de la empresa y su situación actual en torno al mercado.

### 2.3.2. Objetivos Específicos

- Generar gráficos estadísticos basado en datos históricos de la empresa utilizando modelos matemáticos.

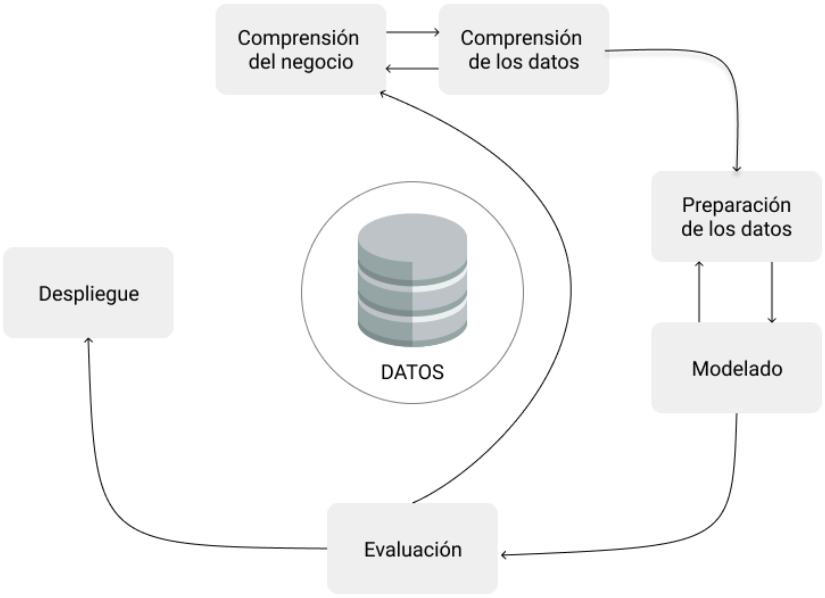
- Diseñar un sistema que integre todas las estadísticas analizadas de manera sencilla y flexible.
- Validar los prototipos de modelos predictivos diseñados utilizando las métricas estándar correspondientes.
- Utilizar una metodología de minería de datos para el aumento de la calidad en el análisis de la información.

## 2.4. Diseño del prototipo

CRISP-DM es definido comúnmente como un marco de trabajo que permite desarrollar trabajos de minería de datos de manera independiente de la tecnología que se utilice y el área a la que se aplique [50]. Consta de seis etapas que pueden ser ajustadas a las necesidades del proyecto, para construir un modelo que se adapte a las necesidades requeridas en un entorno real [51]. CRISP-DM es mayormente utilizado en trabajo orientados al ámbito empresarial que otras metodologías como KDD o SEMMA, por ser más completo y flexible.

La metodología mostrada en el gráfico 5, abarca durante las seis etapas todo un proceso de comprensión y análisis de los datos para lograr obtener conocimiento sobre la estructura y entorno que lo rodea.

*Gráfico 10: Fases de metodología CRISP-DM*




Fuente: Elaboración propia

### 2.4.1. Fase I. Comprensión del negocio

El objetivo de este trabajo es aplicar técnicas de minería de datos sobre la información obtenida en la venta de insumos agrícolas de la empresa ABC, con la finalidad de conocer su estado financiero y general de los procesos más relevantes para las ventas de manera actualizada y real. Actualmente obtener este tipo de reportes en la empresa ABC requiere de un tiempo prudencial de trabajo y no abarca todos los sectores que se requiere, además de requerir de un conjunto de datos muchas veces incompleto o que puede ser fácilmente manipulable.

Actualmente la empresa consta de un sistema transaccional que permite registrar las ventas, compras, inventario y transporte de productos, pagos y más datos contables a los usuarios que lo utilizan. El área de ventas consta con ejecutivos de ventas que recorren gran parte de la provincia y zonas del país, quienes tienen establecidos rutas y horarios que deben ser visitados según un cronograma. El área de adquisiciones se encarga del ingreso sistemático de las compras realizadas a los diferentes proveedores y de establecer los precios de los productos basados en una utilidad porcentual o datos históricos. El encargado de inventario, es responsable de revisar el stock de productos, analizar la demanda actual y solicitar la compra de nuevos productos, basado en varios reportes que le brinda el sistema.

Generalmente se realizan reuniones semanales o mensuales con los ejecutivos de ventas para analizar su rendimiento y discutir de nuevas estrategias y promociones que pueden ser puestas en el mercado, para ello, la administración solicita a los encargados de las áreas correspondientes generar reportes estadísticos que permitan visualizar dicho rendimiento.

Con respecto a los riesgos que podrían surgir en el desarrollo de este proyecto constan: la escasez de datos, la ampliación de tiempos en el desarrollo de las fases planificadas y las situaciones externas que puedan surgir durante el tiempo establecido.

El desarrollo del proyecto no significará ningún costo ya sea para la empresa o para algún miembro que intervenga en el mismo, lo que hace el desarrollo del proyecto tenga más posibilidades de ser factible de implementar.

El objetivo de la minería de datos en este proyecto es el de encontrar patrones o información que a simple vista o con los reportes estadísticos básicos no sea posible detectar rápidamente y de esa manera incrementar las ventas y optimizar los procesos internos de la empresa ABC.

Los criterios de éxito de la minería de datos a realizar son los siguientes:

- Identificar patrones o características en las áreas analizadas que permitan mejorar de manera eficiente los procesos actuales.
- Agilizar el sistema de reportes estadísticos que actualmente se realiza en la empresa.
- Presentar gráficos que permitan conocer el comportamiento de las ventas, compras, ventas, clientes y más áreas internas.

### 2.4.2. Fase II. Estudio y comprensión de los datos

El sistema transaccional que se utiliza en la empresa ABC, utiliza el gestor de base de datos PostgreSQL en su versión 9.4 con la interfaz PgAdmin 3, los datos utilizan la codificación ASCII.

Para realizar las pruebas de los modelos y técnicas que se utilizarán en el desarrollo del proyecto se cuenta con los respaldos de la base de datos de los últimos ocho años,

aunque para asegurar el anonimato de la empresa solo se utilizará los datos de los tres últimos años en el análisis del proyecto.

Tomando en cuenta lo anterior, los datos obtenidos serían los siguientes: 811 días, 33 meses o 3 años.

*Dataset de Ventas y Productos*

Para hacer los diferentes análisis de las ventas son necesarios los campos que se muestran en la tabla 1.

*Tabla 2: Dataset de ventas*

| Campo | Tipo de dato | Descripción |
|---|---|---|
| Fecha | Fecha | Fecha completa de las transacciones |
| Dia_texto | Texto | Dia de la semana en que se realizó la transacción |
| Vendedor | Texto | Persona que realizó la venta |
| Cliente | Texto | Persona que compró los productos |
| Ciudad | Texto | Ciudad del cliente |
| Productos | Texto | Productos listados en la venta |
| Marca | Texto | Marca del producto |
| Bodega | Texto | Bodega desde donde salió el producto |
| Total | Numérico | Monto total de la venta |

Fuente: Elaboración propia

*Dataset de Compras*

Para hacer los diferentes análisis de las compras son necesarios los campos de la tabla 2.

*Tabla 3: Dataset de compras*

| Campo | Tipo de dato | Descripción |
|---|---|---|
| Fecha | Fecha | Fecha en que se realizó la compra |
| Proveedor | Texto | Persona o entidad que proveyó los productos |
| Productos | Texto | Productos adquiridos por la empresa |
| Marca | Texto | Marca de los productos |
| Total | Numérico | Total en dólares de la compra realizada |

Fuente: Elaboración propia

### 2.4.3. Fase III. Análisis de los datos y selección de características

Para la preparación de los datos se han restaurado los datos de los respaldos en una base de datos PostgreSQL en su versión 9.4 debido a que es el gestor y la versión utilizada en el sistema transaccional original. Para manipular los datos y realizar consultas de prueba se ha optado por utilizar PgAdmin v3.

El principal problema que se presenta al extraer los datos, es el formato de codificación que la base de datos utiliza, por lo general los sistemas utilizan la codificación UTF-8 pero estos datos están codificados en ASCII; esto genera una incompatibilidad que limita el uso de diferentes tecnologías de extracción, transformación y carga, por lo que se procedió a realizar todo este proceso utilizando el lenguaje de programación Python, ya que este cuenta con múltiples librerías que permiten hacer el mismo proceso hace que sea la mejor herramienta para realizarlo.

Para el proceso de limpieza de los datos se procede a realizar las respectivas operaciones para que todos los datos sean legibles, sigan un estándar y permitan analizarlos sin problemas.

El proceso de integración de datos se lo realiza de forma integral utilizando Python para evitar que los datos tengan que pasar por un nuevo proceso de codificación y limpieza al ser cargados para ser utilizados ya sea por Python o R.

### 2.4.4. Fase IV. Modelado

En la tabla 3 se detallan las técnicas que se utilizan para el desarrollo del proyecto, junto con las fórmulas general que describen su funcionamiento para poder encontrar los valores requeridos.

*Tabla 4: Fórmulas de técnicas aplicadas*

| Tipo de técnica | Técnica | Modelo matemático | Descripción de la aplicación |
|---|---|---|---|
| **Estadísticos descriptivos** | Media | $\bar{x} = \frac{\Sigma_{i=1}^{n} x_i}{n}$ | Ventas, Compras |
| | Mediana | $Me = Li + a_i \left( \frac{\frac{n}{2} - F_{m-1}}{f_m} \right)$ | Ventas, Compras |
| | Cuartil | $Q_k = L_i + \left( \frac{\frac{k.N}{4} - F_{i-1}}{f_i} \right) a_i$ | Ventas, Compras |
| | Varianza | $S^2 = \sigma^2 = \frac{\Sigma_{i=1}^{n} (x_i . \bar{x})^2}{n}$ | Ventas, Compras |
| | Desviación estándar | $S = \sigma = \sqrt{\frac{\Sigma_{i=1}^{n} (x_i . \bar{x})^2}{n}}$ | Ventas, Compras |
| **Técnicas descriptivas** | Reglas de asociación | $X => Y$<br>$Soporte = \frac{frq(X,Y)}{N}$ | Patrones en las ventas |

| | | | |
|---|---|---|---|
| | | $Confianza = \frac{frq(X,Y)}{frq(X)}$<br>$Lift = \frac{Soporte}{Sop(X).Sop(Y)}$ | |
| **Técnicas predictivas** | Series temporales | Autorregresiva<br>$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$ | Ventas por año-mes |
| | | Media Móvil<br>$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t$ | Ventas por año-mes |
| | | ARIMA<br>$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t$ | Ventas por año-mes |
| | | Holt-Winters<br>$L_t = \alpha \frac{Y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1}$<br>$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$<br>$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma)S_{t-s}$<br>$F_{t-m} = (L_t - b_t m)S_{t-s+m}$ | Ventas por año-mes |

Fuente: Elaboración propia

La ejecución de estas técnicas se puede apreciar en el apartado 2.5 de este documento.

### 2.4.5. Fase V. Evaluación

Esta parte será desarrollada en el capítulo 3 de este documento, ahí se evaluarán los modelos e identificará el mejor para cada propósito.

### 2.4.6. Fase VI. Despliegue

Debido a que este proyecto solo se plantea como un prototipo no será implementado en ningún lugar, aunque se presentarán todas las bases para hacerlo.

## 2.5. Ejecución y/o ensamblaje del prototipo

Durante el desarrollo del proyecto se utilizó una serie de herramientas que permitieron obtener los resultados esperados, a continuación, se detalla la función de cada uno de ellos:

### 2.5.1. Técnicas utilizadas

#### 2.5.1.1. Series temporales

Las series temporales son modelos estadísticos que permiten predecir valores futuros basados en información histórica, en Python existen varias librerías que facilitan el uso de estas técnicas, entre las que se utilizó en este proyecto se encuentran statsmodels y sklearn. Para las entrenar y probar los modelos se utilizó la información de la empresa ABC de los últimos 8 años, aunque por cuestiones de privacidad solo se mostrarán los gráficos de los últimos 3 años.

Los gráficos mostrados son realizados utilizando la librearía Matplotlib, la cual permite una gran flexibilidad tanto en el ajuste de las variables como el diseño de las figuras.

Cada uno de los modelos de series temporales realizados se presentará junto con todo su proceso de preparación y corrección de los datos, hasta llegar al resultado esperado.

#### 2.5.1.2. Preparación de los datos

Para elaborar de manera correcta una serie temporal es ideal escoger bien el periodo de los datos que van a ser analizados; para este caso luego de varias pruebas se escogió agrupar los datos de las ventas por meses. Luego es necesario extraer esos datos y juntarlos, para ello se hizo uso de la librería pandas y psycopg2 para la conexión con la base de datos, como se observa en el siguiente código de la tabla 5.

*Tabla 5: Extracción de datos*

**Función:** Extracción de datos

**Lenguaje:** Python

**Código:**

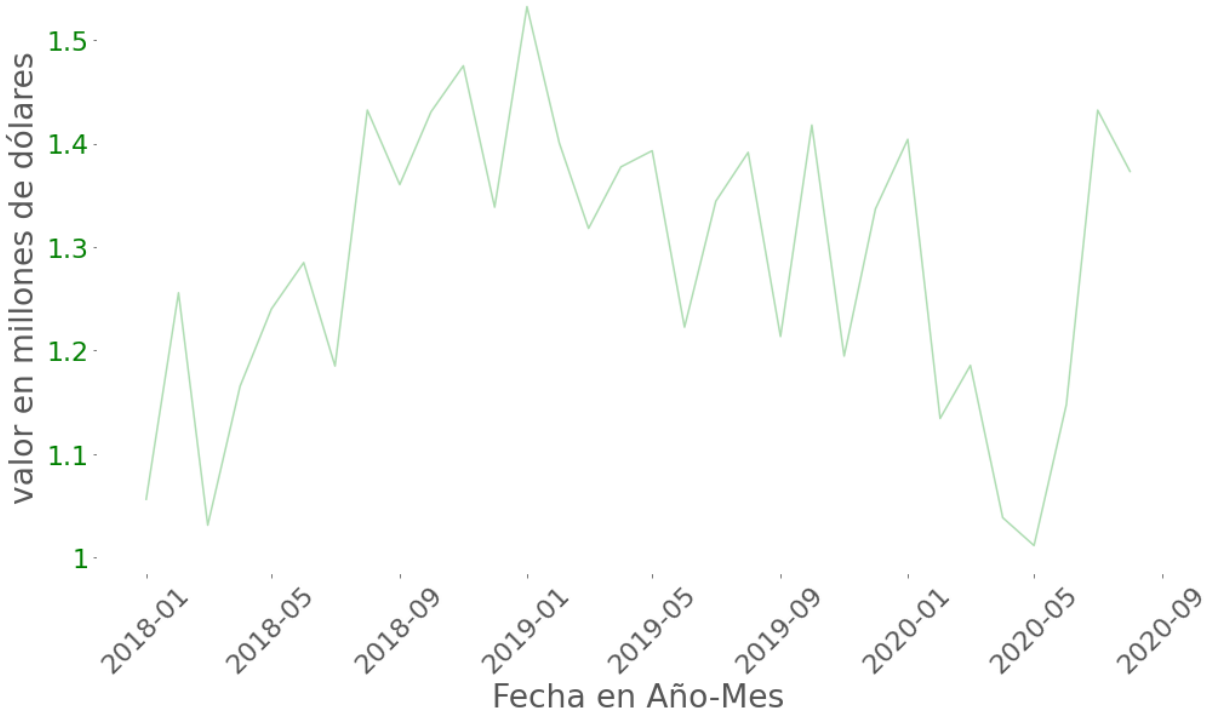
```
import pandas.io.sql as sqlio
import psycopg2

# Conexión con las diferentes bases de datos
conn20 = psycopg2.connect("dbname=bd20 user=postgres password=postgres")
conn19 = psycopg2.connect("dbname=bd19 user=postgres password=postgres")
conn18 = psycopg2.connect("dbname=bd18 user=postgres password=postgres")
psycopg2.extensions.register_type(psycopg2.extensions.BYTES)

# Extracción de datos
data20 = sqlio.read_sql_query(sql, conn20)
data19 = sqlio.read_sql_query(sql, conn19)
data18 = sqlio.read_sql_query(sql, conn18)

data = pd.concat([data18, data19, data20])
```

**Fuente:** Elaboración propia

Para facilitar el manejo y desarrollo de los gráficos, es necesario configurar las fechas como índices de cada fila, de esa manera se hace referencia a la fecha al que pertenece dicho valor.

Procedemos a graficar las ventas organizadas por mes, con el fin de encontrar algún patrón o tendencia que se repita y permita obtener los valores para próximamente poder utilizarlos al aplicar los modelos de predicción, (Ver gráfico 11).

*Gráfico 11: Total de ventas (2018 – 2020)*




**Fuente:** elaboración propia

Si observamos la figura tal no se puede encontrar un patrón en los datos de forma clara, la mayoría de los métodos de predicción para arrojar resultados precisos, requieren que la serie sea estacionaria, lo que significa que la media, varianza y covarianza no varíen con el tiempo, por ello, recurriremos a las pruebas: Dickey Fuller (ADF) y Kwiatkowski-Phillips-Schmidt-Shin (KPSS). Estos métodos permiten comprobar la estacionariedad de una serie.

Dickey Fuller plantea dos hipótesis:

- Hipótesis nula: La serie de datos tiene una raíz unitaria (No es estacionaria)
- Hipótesis alternativa: La serie de datos no tiene una raíz unitaria (Es estacionaria)

En donde si el valor del *p-value* es menor que el nivel de significancia 0.05 entonces la serie es estacionaria, caso contrario la serie no posee estacionariedad y no se puede rechazar la hipótesis nula. Como podemos ver en el gráfico 12, el valor de p-value es mayor que el nivel de significancia, lo que significa que la serie no es estacionaria.

*Gráfico 12: Resultados prueba Dickey-Fuller*

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.076693
p-value                          0.253989
#Lags Used                       3.000000
Number of Observations Used    100.000000
Critical Value (1%)             -3.497501
Critical Value (5%)             -2.890906
Critical Value (10%)            -2.582435
dtype: float64
```

**Fuente:** Elaboración propia

Ahora procedemos a aplicar el método KPSS para determinar las propiedades de la serie, en este caso KPSS plantea las hipótesis al contrario de ADF, por lo que al plantearlas quedaría de la siguiente manera:

- Hipótesis nula: La serie de datos no tiene una raíz unitaria
- Hipótesis alternativa: La serie de datos tiene una raíz unitaria

Como se puede ver en el gráfico 13 el valor de p es mayor al nivel de significancia, por lo tanto, la hipótesis nula no puede ser rechazada, esto significa que la serie es estacionaria.

*Gráfico 13: Resultados prueba KPSS*

```
Results of KPSS Test:
Test Statistic           0.405577
p-value                  0.074751
Lags Used                5.000000
Critical Value (10%)     0.347000
Critical Value (5%)      0.463000
Critical Value (2.5%)    0.574000
Critical Value (1%)      0.739000
dtype: float64
```

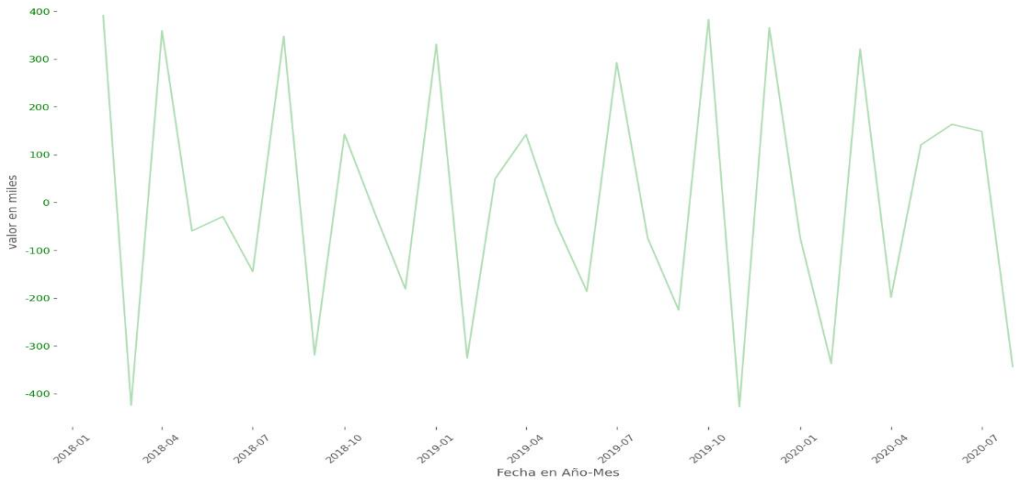
**Fuente:** Elaboración propia

Con los resultados de ADF y KPSS se plantea una situación, y para resolverlo, hay que revisar lo que establecen los modelos en estos casos. Si KPSS indica estacionariedad y ADF no, significa que la serie posee tendencia estacionaria, por lo tanto, hay que eliminar la tendencia, para que toda la serie se vuelva estacionaria.

Para resolver el problema de la tendencia, se procede a diferenciar la serie cuantas veces sea necesario, comprobando en cada diferenciación los valores resultantes.

Luego de aplicar una diferenciación a la serie se puede observar el resultado en el gráfico 14, también se evidencia que el valor de p en la prueba ADF es mucho menor al nivel de significancia, por lo que la serie es estacionaria, y al visualizar la prueba KPSS se verifica dicha afirmación, debido a que p es mayor al nivel de significancia, dando como resultado una serie estrictamente estacionaria. Con esto obtenemos que el número de diferenciaciones necesarias para que la serie se vuelva estacionaria es de 1.

*Gráfico 14: Serie diferenciada una vez*




| Results of Dickey-Fuller Test: | |
|---|---|
| Test Statistic | -5.530579 |
| p-value | 0.000002 |
| #Lags Used | 9.000000 |
| Number of Observations Used | 92.000000 |
| Critical Value (1%) | -3.503515 |
| Critical Value (5%) | -2.893508 |
| Critical Value (10%) | -2.583824 |
| dtype: float64 | |

| Results of KPSS Test: | |
|---|---|
| Test Statistic | 0.154609 |
| p-value | 0.100000 |
| Lags Used | 22.000000 |
| Critical Value (10%) | 0.347000 |
| Critical Value (5%) | 0.463000 |
| Critical Value (2.5%) | 0.574000 |
| Critical Value (1%) | 0.739000 |
| dtype: float64 | |

**Fuente:** Elaboración propia

### 2.5.1.2.1. ARIMA

El modelo de media móvil integrada autorregresiva (ARIMA) es la combinación de dos métodos predictivos: el modelo autorregresivo (AR) y el modelo de media móvil (MA). Esta técnica es muy utilizada para encontrar valores en series con componentes estacionales; en Python este modelo puede ser aplicado utilizando la librería statsmodel, en donde se puede configurar varios parámetros según necesiten los datos, además cuenta con modelo mejorado denominado SARIMAX, el cual incluye el componente estacional como parámetro. Este modelo también permite graficar cada

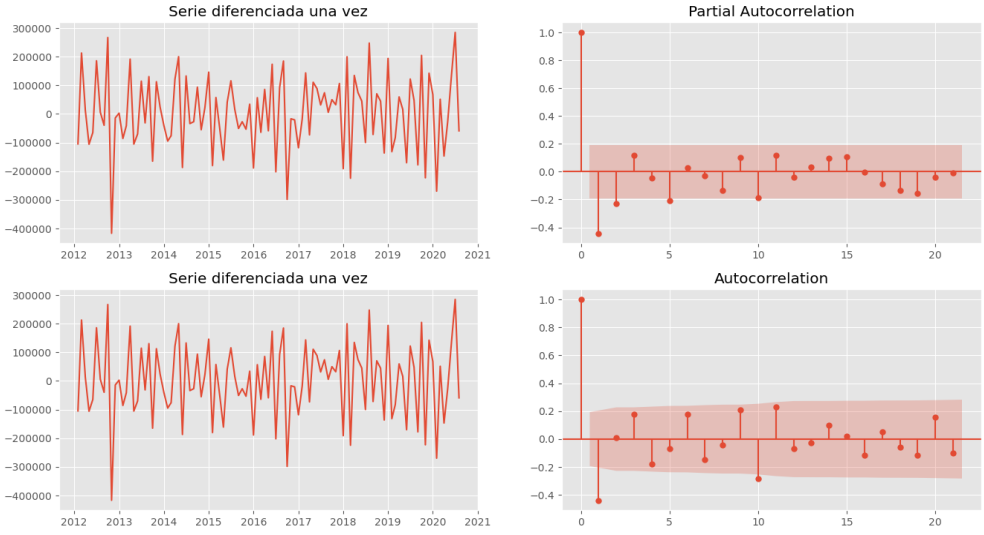
una de las técnicas que componen ARIMA de manera independiente, encerando los demás parámetros.

Al modelo, se lo expresa comúnmente de la siguiente forma: ARIMA (p, d, q). Donde *p* es el orden de retraso obtenido del número de observaciones, *d* es el número de diferenciaciones necesarias para el modelo, *q* es el orden de media móvil.

Una de las formas más común y práctica para conocer los valores de *p* y *q* es obtener los gráficos de autocorrelación parcial y autocorrelación respectivamente, para determinar el orden es necesario observar cuantos valores sobrepasan el espacio significativo denominado *Alpha*. Cabe recalcar que este análisis dado que nuestra serie no es estacionaria se le debe aplicar a la serie diferenciada.

En el gráfico 15 se puede observar que al hacer una diferenciación en la serie el valor de p sería dos, y el valor de q igualmente sería dos.

*Gráfico 15: Autocorrelación y Autocorrelación parcial*




**Fuente:** Elaboración propia

Entonces el modelo tentativo luego del análisis según las pruebas previamente realizadas sería ARIMA (2,1,2). Con el planteamiento ya definido, ahora se proceden a realizar las pruebas y verificar las métricas correspondientes.

En el código que se detalla a continuación se establece el modelo ARIMA siguiendo los órdenes determinados, utilizando la librería de statsmodels, detallado en la tabla 6.

*Tabla 6: Código para obtener valores estadísticos del modelo*

| **Función:** Resumen de valores estadísticos del modelo |
|---|
| **Lenguaje:** Python |
| **Código:** |

```
from statsmodels.tsa.arima.model import ARIMA

mod = sm.tsa.arima.ARIMA(df_month['total'], order=(2, 1, 2))
res = mod.fit()
print(res.summary())
```

**Fuente:** Elaboración propia

*Gráfico 16: Resultados estadísticos del modelo ARIMA*

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  total   No. Observations:                  104
Model:                 ARIMA(2, 1, 2)   Log Likelihood               -1345.492
Date:                Tue, 01 Dec 2020   AIC                           2700.984
Time:                        20:26:14   BIC                           2714.158
Sample:                  01-01-2012   HQIC                          2706.320
                       - 08-01-2020
Covariance Type:                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9337      0.091    -10.312      0.000      -1.111      -0.756
ar.L2         -0.8222      0.070    -11.669      0.000      -0.960      -0.684
ma.L1          0.7241      0.106      6.862      0.000       0.517       0.931
ma.L2          0.7684      0.093      8.288      0.000       0.587       0.950
sigma2      1.304e+10   2.03e-12   6.43e+21      0.000     1.3e+10     1.3e+10
===================================================================================
Ljung-Box (L1) (Q):                   5.13   Jarque-Bera (JB):                 0.73
Prob(Q):                              0.02   Prob(JB):                         0.69
Heteroskedasticity (H):               1.10   Skew:                            -0.11
Prob(H) (two-sided):                  0.78   Kurtosis:                         2.65
===================================================================================
```

**Fuente:** Elaboración propia

En el gráfico 16, se puede observar las métricas que arroja al entrenar el modelo con los parámetros ingresados, y como resultado vemos que el modelo parece ser muy eficiente, dado que el valor de p está muy por debajo del nivel de significancia 0,05 y vemos que el valor de AIC disminuyó considerablemente comparado con modelos realizados previamente. Si visualizamos la gráfica 17, al aplicar el modelo SARIMAX (2, 1, 2) x (2, 1, [1], 48), los valores de p varían sobrepasando el nivel de significancia, pero el AIC disminuye considerablemente, y es un factor muy importante para inclinarse más por este modelo como el final en la utilización de ARIMA, (Ver gráfico 17).

*Gráfico 17: Resultados estadísticos del modelo SARIMAX*

SARIMAX Results

| | | | |
|---|---|---|---|
| Dep. Variable: | total | No. Observations: | 104 |
| Model: | SARIMAX(2, 1, 2)x(2, 1, [1], 48) | Log Likelihood | -726.869 |
| Date: | Tue, 01 Dec 2020 | AIC | 1469.739 |
| Time: | 22:26:27 | BIC | 1485.798 |
| Sample: | 01-01-2012 - 08-01-2020 | HQIC | 1475.949 |
| Covariance Type: | opg | | |

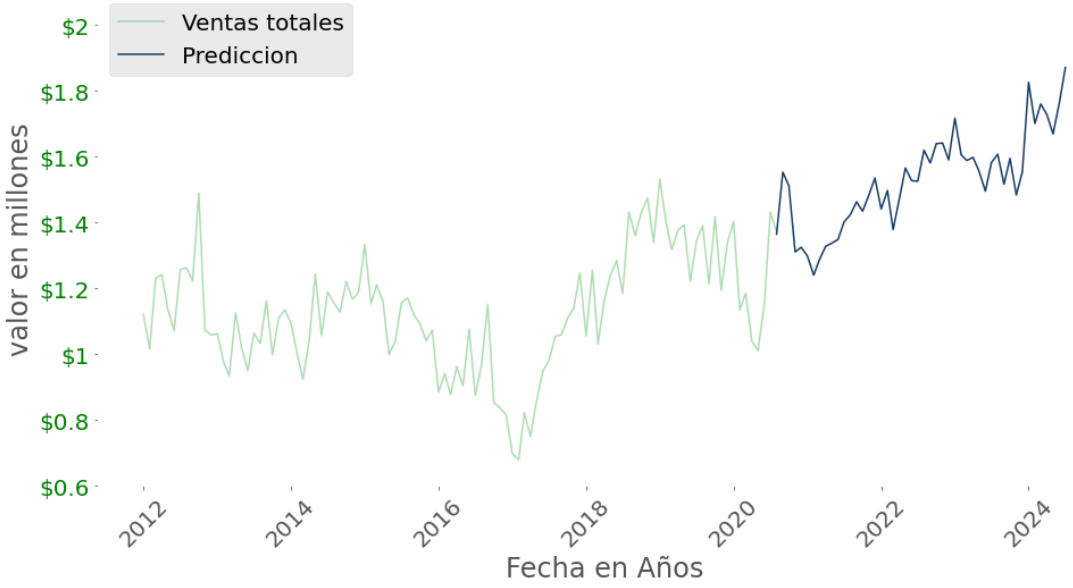
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| ar.L1 | -0.4394 | 0.266 | -1.650 | 0.099 | -0.961 | 0.082 |
| ar.L2 | -0.5259 | 0.322 | -1.635 | 0.102 | -1.156 | 0.104 |
| ma.L1 | 0.0847 | 0.263 | 0.322 | 0.747 | -0.431 | 0.600 |
| ma.L2 | 0.6692 | 0.189 | 3.536 | 0.000 | 0.298 | 1.040 |
| ar.S.L48 | -0.5756 | 4955.586 | -0.000 | 1.000 | -9713.346 | 9712.195 |
| ar.S.L96 | -0.2684 | 2250.912 | -0.000 | 1.000 | -4411.975 | 4411.438 |
| ma.S.L48 | -0.0006 | 5340.646 | -1.05e-07 | 1.000 | -1.05e+04 | 1.05e+04 |
| sigma2 | 2.161e+10 | 0.006 | 3.5e+12 | 0.000 | 2.16e+10 | 2.16e+10 |

| | | | |
|---|---|---|---|
| Ljung-Box (L1) (Q): | 2.23 | Jarque-Bera (JB): | 0.65 |
| Prob(Q): | 0.13 | Prob(JB): | 0.72 |
| Heteroskedasticity (H): | 1.25 | Skew: | -0.10 |
| Prob(H) (two-sided): | 0.64 | Kurtosis: | 3.50 |

**Fuente:** Elaboración propia

Una vez aplicado el modelo sobre los datos la solución que brinda es la que se observa en el gráfico 18.

*Gráfico 18: Predicción de ventas con ARIMA en Python*

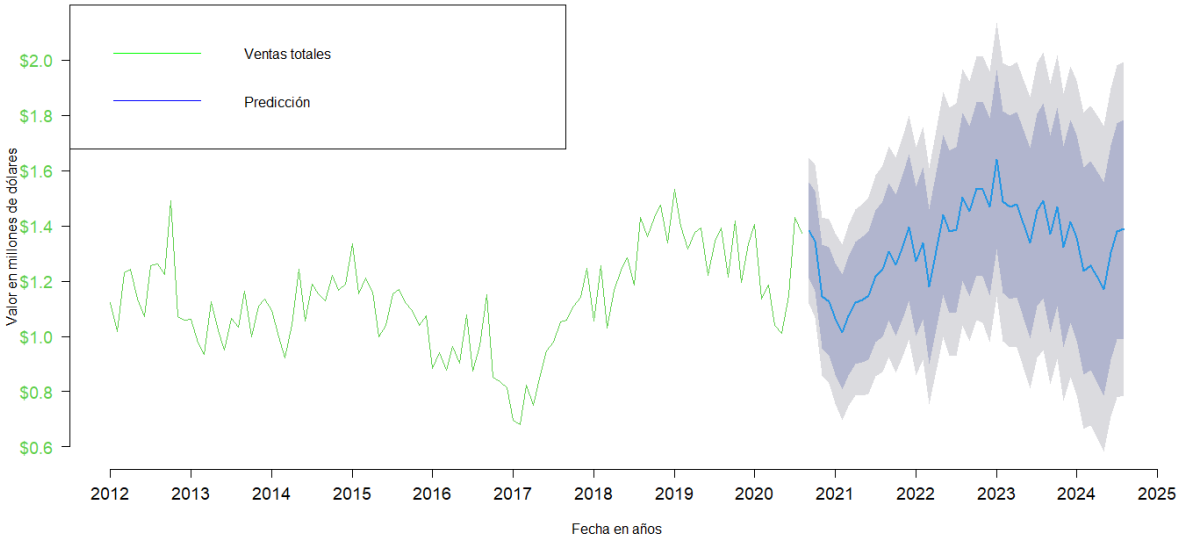



**Fuente:** Elaboración propia

Para realizar una comparación de los modelos y poder verificar que las predicciones son factibles en base a los datos analizados, se procede a realizar la misma predicción utilizando el modelo ARIMA, pero ahora desarrollado en el lenguaje R. Para ello haremos uso del IDE RStudio el cual es una herramienta muy potente y en parte parecido a Jupyter por las capacidades que le dan a su respectivo lenguaje.

Primero procedemos a cargar los datos que han sido extraídos utilizando las librerías de Python y transformados en archivo csv, luego debido a la gran cantidad de paquetes instalados en R, la forma de graficar predicciones de modelos estadísticos es mucho más sencillo, en este caso es simplemente establecer los órdenes de *p*, *d* y *q* y el periodo de estacionalidad que se deberá seguir como patrón, en este caso los valores serían ARIMA (0,1,1) y 48 periodos, (Ver gráfico 19).

*Gráfico 19: Predicción de ventas con ARIMA en R*



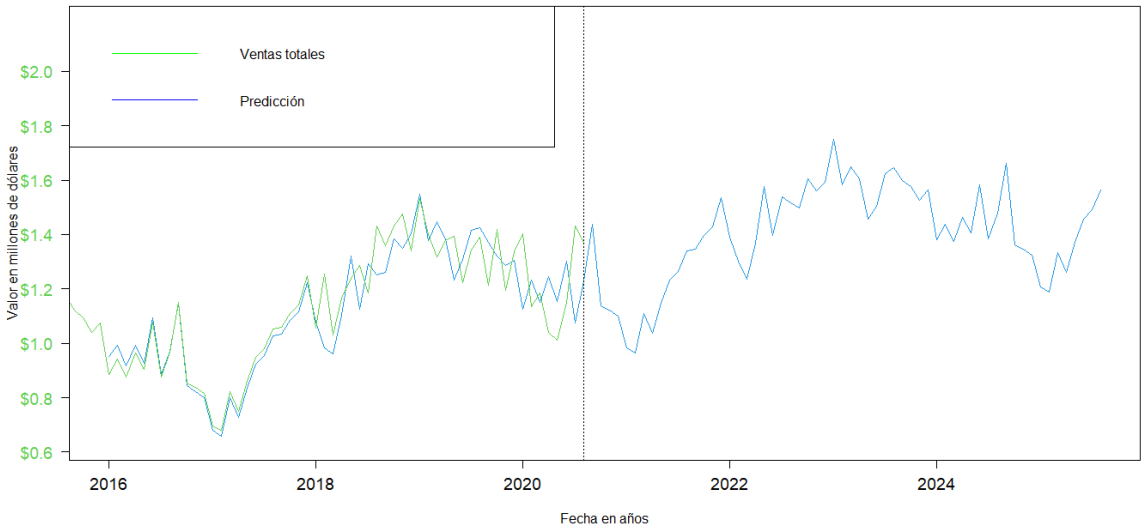

**Fuente:** Elaboración propia

A simple vista parece ser que las predicciones de las series desarrolladas con Python y R, no tienen mucha relación, pero esto se procederá a evaluar en el capítulo 3.

### 2.5.1.2.2. Holt-Winters

Utilizando el lenguaje de programación R, se aplicó un modelo Holt-Winters a los datos de las ventas mensuales de la empresa, este método es una modificación de la suavización exponencial, el cual es ideal para realizar predicciones no solo a corto plazo sino a mediano y largo plazo. Con los valores de la serie obtenidos previamente como el período y las herramientas de estadística de R se elaboró una predicción a 5 años, con una frecuencia de iteración de la serie de 48 meses, y con un nivel de confianza del 95%, (Ver gráfico 20).

*Gráfico 20: Serie temporal Holt-Winters en R*



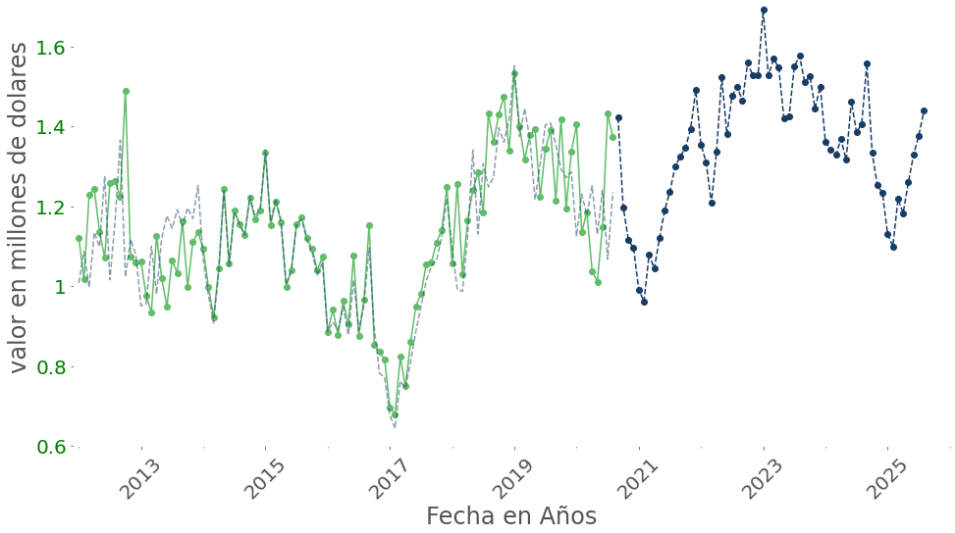

**Fuente:** Elaboración propia

En la gráfica 20, se puede observar con color azul luego de línea vertical segmentada, los valores que el modelo predijo. Con base en los periodos anteriores se ve en la imagen que el modelo puede ser factible, pero eso se comprobará en el capítulo 3 de este documento.

Ahora se procede a realizar una serie temporal aplicando el modelo Holt-Winters en Python, para ello se hará uso del método SmoothingExponential de la librería de statsmodel, como parámetros se establecerán los mismo usados anteriormente, por ello la cantidad de periodos estacionales se fijará en 48 y se utilizará una tendencia multiplicativa y estacionalidad aditiva.

Para entrenar el modelo se le establece como parámetro 100 repeticiones, se calcula el error multiplicativo y se procede a predecir los siguientes 5 años, (Ver gráfico 21).

*Gráfico 21: Serie temporal Holt-Winters en Python*




ExponentialSmoothing Model Results

| | | | |
|---|---|---|---|
| Dep. Variable: | total | No. Observations: | 104 |
| Model: | ExponentialSmoothing | SSE | 1277021406238.209 |
| Optimized: | True | AIC | 2520.041 |
| Trend: | Multiplicative | BIC | 2657.549 |
| Seasonal: | Additive | AICC | 2641.265 |
| Seasonal Periods: | 48 | Date: | Sun, 06 Dec 2020 |
| Box-Cox: | False | Time: | 10:29:07 |
| Box-Cox Coeff.: | None | | |

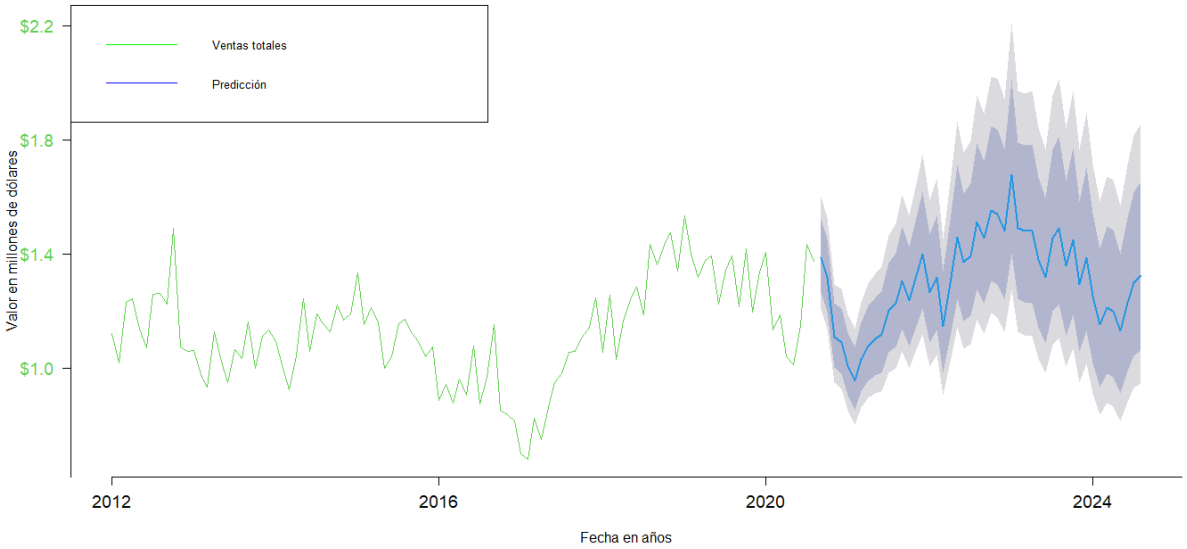
**Fuente:** Elaboración propia

Comparando los valores de AIC entre los dos modelos hechos en Python, parece ser que la predicción realizada con ARIMA está más ajustada al conjunto de datos de la empresa ABC, en el capítulo 3 se escogerá cuál de los modelos planteados termina siendo el mejor para la serie.

### 2.5.1.2.3. Suavizado Exponencial

Para encontrar el modelo más preciso para este conjunto de datos se hace uso de varias combinaciones de modelos, en este caso se implementa la técnica STL que aplica un suavizado exponencial al que se le puede aplicar otro modelo ya sea arima, ets, naive o drift, pero para efecto de este grafico se aplicara únicamente el modelo sin otro método.

*Gráfico 22: Predicción en R utilizando STLF*



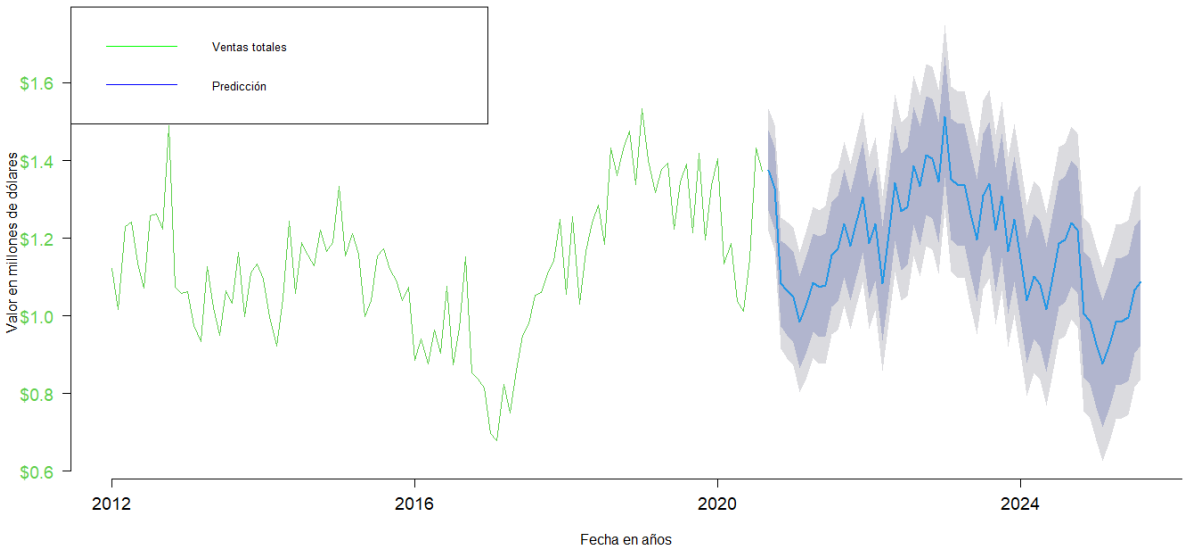

**Fuente:** Elaboración propia

Como se puede observar en el gráfico 22 el modelo arroja una predicción bastante coherente con los datos anteriores, además predice altas y bajas en las ventas, las mismas que deben ser analizadas una vez puesto en producción, si llegase a ser escogido el modelo.

### 2.5.1.2.4. Suavizado Exponencial y Autorregresión

En el gráfico 23 se observa la aplicación de la técnica de suavizado exponencial acompañado del método autorregresivo sobre los datos, esto da como consecuencia una predicción más acertada en forma visual y también predice bajas y altas sobre las ventas.

*Gráfico 23: Predicción en R utilizando STLM y Autorregresión*

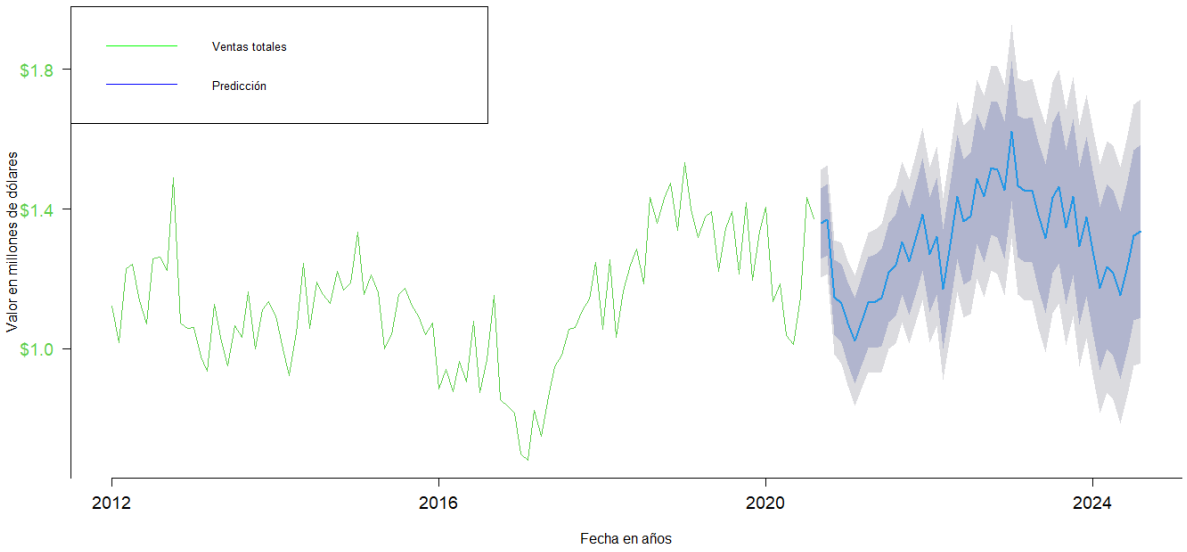



**Fuente:** Elaboración propia

### 2.5.1.2.5. Suavizado Exponencial y ARIMA

Esta técnica es utilizada para combinar suavización exponencial con modelos estacionales, por ello se hizo uso del método en R denominado s*tlm* que permite utilizar ambos modelos, con el fin de encontrar un modelo aún más preciso.

*Gráfico 24: Predicción en R utilizando STLM y ARIMA*




**Fuente:** Elaboración propia

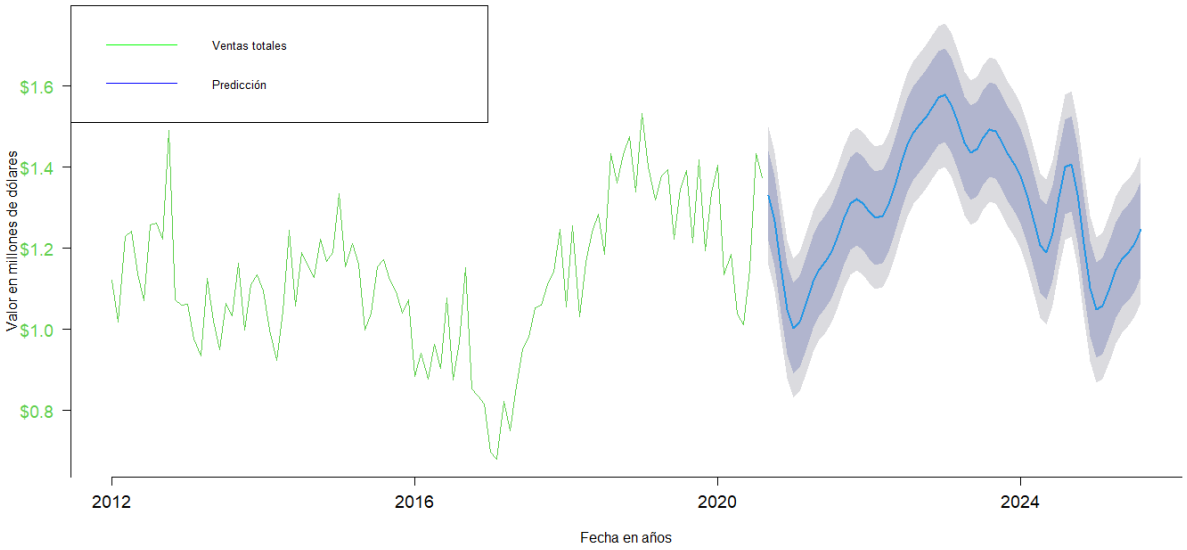
El resultado de la aplicación del método se puede observar en el gráfico 24 el cual es muy parecido al del gráfico 22, pero se puede observar que el índice de error en cuanto

a las predicciones se disminuye considerablemente y es más ajustado al valor de la predicción, esto será analizado en el capítulo 3 donde se hará la selección del modelo ideal a aplicar finalmente.

#### 2.5.1.2.6. TBATS

El modelo TBATS es parte de los modelos de suavizado exponencial, el cual aplica una combinación con múltiples técnicas que son: Transformación Box-Cox, errores ARMA, componentes de tendencia y estacionalidad. Se utilizó esta técnica por la cantidad de atributos que puede recibir y la forma en que combina los métodos que se selecciona.

*Gráfico 25: Predicción en R utilizando TBATS*




**Fuente:** Elaboración propia

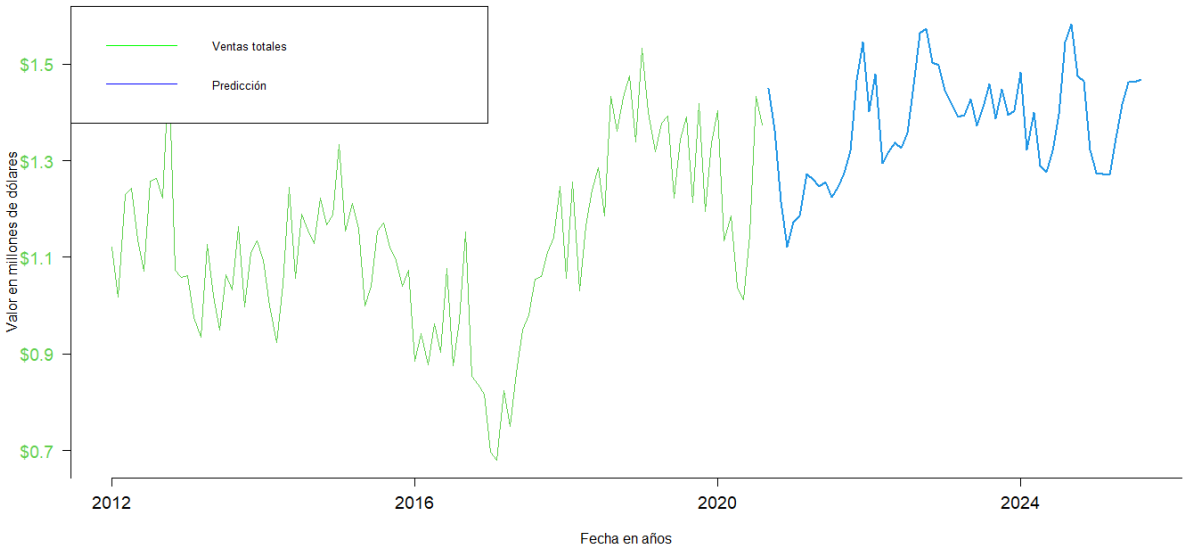
La aplicación del modelo se puede observar en el gráfico 25, presenta un gráfico en términos generales parecido al modelo realizado con ARIMA, pero los picos parecen estar ajustados para no verse muy irregular, lo cual al parecer puede no ajustarse a los datos anteriores.

#### 2.5.1.2.7. NNETAR

NNETAR es un método exclusivo de R que utiliza modelos de redes neuronales para encontrar predicciones de series temporales, muy útil debido a su forma de encontrar

nuevos valores en base a técnicas de aprendizaje automático que calculan pesos automáticamente.

*Gráfico 26: Predicción en R utilizando NNETAR*




**Fuente:** Elaboración propia

Como se puede observar en el gráfico 26, el cual a simple vista parece no ir tan acorde con los datos anteriores, debido a que las ventas a lo largo del tiempo han presentado altas y bajas pero no de forma tan brusca sino más bien de forma periódica. Pero eso será analizado con más detalle en el capítulo 3.

#### 2.5.1.3. Reglas de asociación

Las reglas de asociación son utilizadas en la minería de datos para descubrir patrones de comportamientos de los elementos analizados, en este caso se aplicará esta técnica para descubrir conocer que productos son los que con más frecuencia se compran juntos y brindar un ayuda en la creación de promociones y nuevas estrategias de ventas.

Debido a que el lenguaje R contiene métodos que facilitan el manejo de elementos conjuntos, el análisis principalmente se lo realizará en este lenguaje.

Primero cargamos los datos extraídos utilizando Python y transformados en csv, para realizar un análisis correcto es necesario tener los siguientes datos de los productos a

analizar: el número de factura y el nombre o código de los productos que fueron comprados por los clientes.

Para asegurar el anonimato y seguridad de los datos de la empresa ABC se optó por renombrar los productos según un orden aleatorio. Partimos de un conjunto de datos conformado por más de 73 mil transacciones y 772 ítems, para conocer mejor los datos y las características de los productos a analizar, buscamos la frecuencia relativa y absoluta de las transacciones de cada producto, en la tabla 7 se ven los 10 productos que más apariciones tienen.

*Tabla 7: Tabla de frecuencias de productos*

| Producto | Frecuencia relativa | Frecuencia absoluta |
|---|---|---|
| PRODUCTO-44 | 0.12228489 | 9019 |
| PRODUCTO-11 | 0.08192098 | 6042 |
| PRODUCTO-4 | 0.07992787 | 5895 |
| PRODUCTO-5 | 0.06257288 | 4615 |
| PRODUCTO-19 | 0.06159666 | 4543 |
| PRODUCTO-38 | 0.05816634 | 4290 |
| PRODUCTO-2 | 0.04574667 | 3374 |
| PRODUCTO-83 | 0.04472978 | 3299 |
| PRODUCTO-26 | 0.04470266 | 3297 |
| PRODUCTO-13 | 0.04406541 | 3250 |

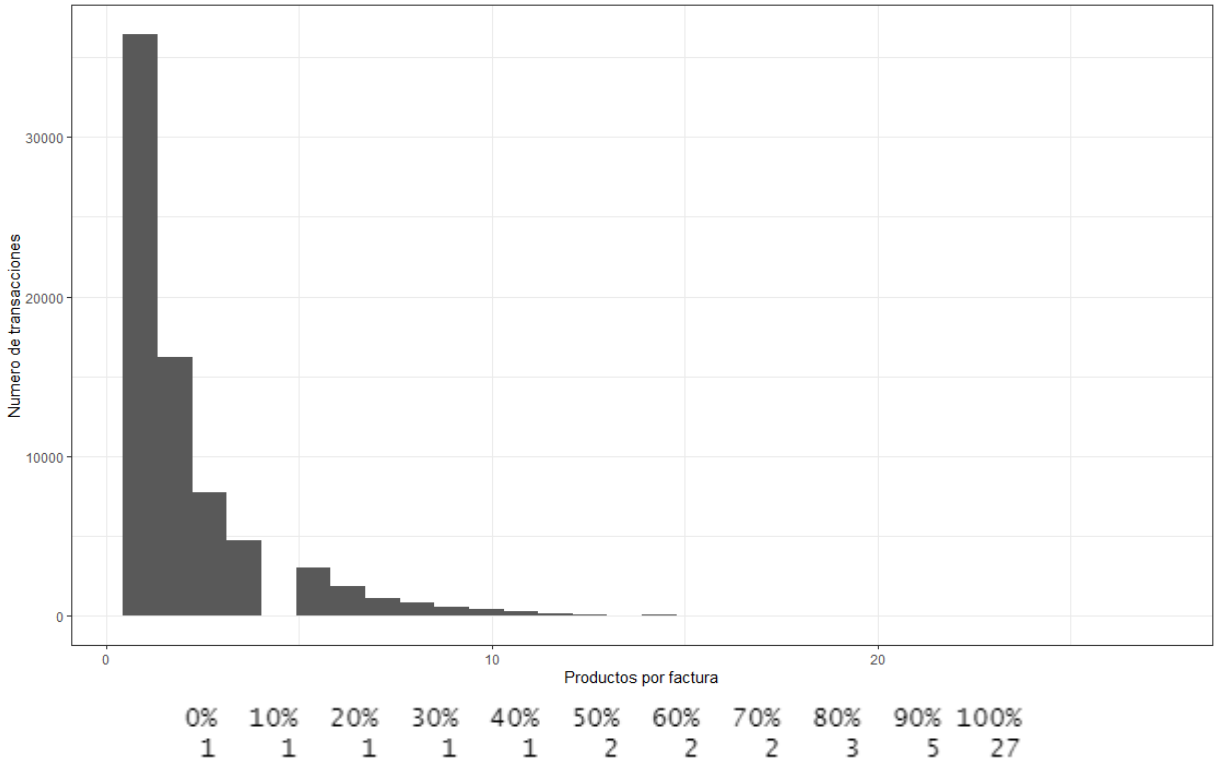
**Fuente:** Elaboración propia

Antes de continuar con el análisis es necesario conocer dos conceptos fundamentales que se debe tener en cuenta para el posterior análisis: el número de transacciones que contienen el ítem o conjunto de ítems llamado soporte, y la confianza que es la probabilidad que existe en la que una transacción que contiene *X*, también contenga el ítem *Y*.

Debido a que queremos conocer el comportamiento de las ventas basándonos en lo que llevan los clientes, se debe tomar en cuenta solo aquellas facturas que tengan mínimo 2 productos, para ello se ha clasificado por cuartiles la cantidad de productos que lleva cada cliente por factura lo que dio como resultado lo que se observa en la gráfica 22 en donde se aprecia en el histograma que un gran porcentaje de transacciones cuentan con un solo producto, y revisando los cuartiles vemos que ese

porcentaje equivale al 50% del total de transacciones, además, se observa que las facturas con más de 10 ítems muy pocas comparadas con el resto. Si nos basáramos en lo que nos revela los cuartiles, se podría trabajar esta regla de asociación con los límites 2 y 5 para el descubrimiento de patrones y con eso se abarcaría el 50% de los datos sumado al 40% que solo tiene una transacción se estaría trabajando con el 90% de los datos disponibles; aunque para efectos de prueba en este proyecto se dejará como límites 2 y 20 ítems por factura, (Ver gráfico 27).

*Gráfico 27: Distribución de productos por cantidad*




**Fuente:** Elaboración propia

Una vez conocido los conceptos, procedemos a establecer los parámetros para la selección de datos; para el soporte establecemos todos aquellos productos que aparezcan al menos en 500 facturas, en la tabla 8 se muestra el código para dividir los conjuntos de datos.

*Tabla 8: Reglas de asociación - Algoritmo Apriori*

| **Función:** Establecer conjunto de datos en algoritmo Apriori |
|---|

**Lenguaje:** R

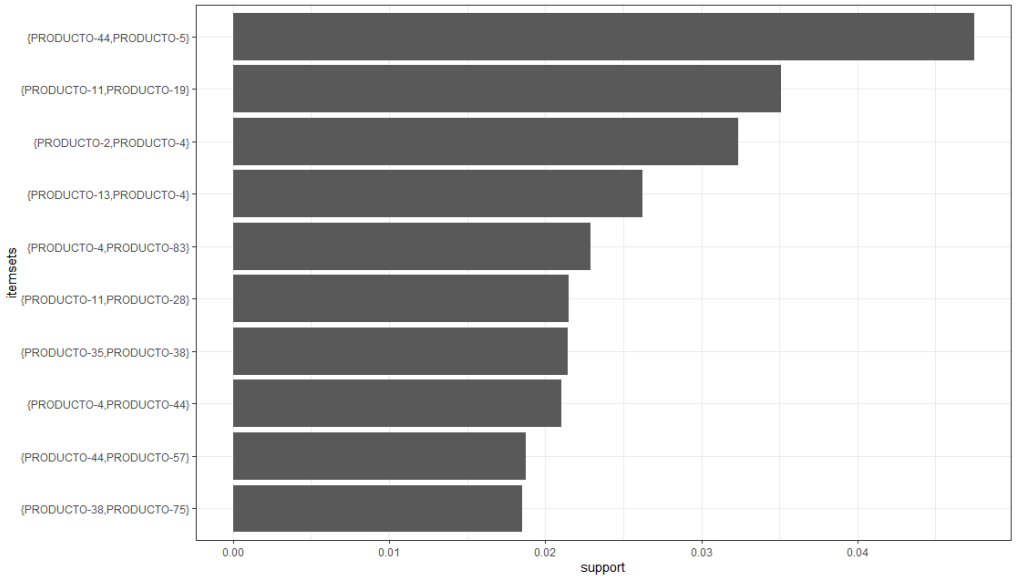
**Código:**

```
soporte <- 500 / dim(transacciones)[1]

itemsets <- apriori(data = transacciones,

                    parameter = list(support = soporte,

                                     minlen = 2,

                                     maxlen = 20,

                                     target = "frequent itemset"))
```

**Fuente:** Elaboración propia

Al ordenar los valores obtenidos previamente, da como resultado un listado de los 10 conjuntos de datos que más se repiten en las facturas hechas por la empresa, que es lo que se observa en la gráfica 28. Hasta ahora ya hemos extraído mucha información sobre los datos y la asociación que existe entre productos, pero es ahora cuando hay que empezar a descubrir patrones y crear reglas en base a ello.

*Gráfico 28: Conjunto de ítems por frecuencia*




**Fuente:** Elaboración propia

Ahora se procede a establecer las reglas con un porcentaje de confidencialidad, para efectos de prueba se puede empezar con un 70% y revisar si es adecuado o no para

el tipo de análisis que se quiera realizar, el código en R de esa técnica se muestra en la tabla 9.

*Tabla 9: Establecer reglas en algoritmo Apriori*

| **Función:** Establecer reglas en algoritmo a priori por soporte y confianza |
|---|
| **Lenguaje:** R |
| **Código:** |

```
soporte <- 500 / dim(transacciones)[1]
reglas <- apriori(data = transacciones,
                  parameter = list(support = soporte,
                                   confidence = 0.70,
                                   target = "rules"))
```

**Fuente:** Elaboración propia

Al inspeccionar los datos que nos arroja aplicando la regla anterior, vemos que debido a la estructura que maneja la empresa al establecer promociones, afecta en el análisis que se está realizando, por lo que es necesario evaluar con criterio las asociaciones que el algoritmo analiza, debido a que existen productos con el mismo nombre y con la etiqueta '*promo'* en medio del nombre, esto produce una confusión al algoritmo ya que es imposible para él determinar si se trata del mismo producto o no. Teniendo en cuanta lo antes mencionado procedemos a consultar los conjuntos de ítems. Como se puede observar en la tabla 10 el algoritmo logra encontrar las relaciones porcentuales existentes entre los ítems, en este caso se muestran los 5 productos con más Lift, este término hace referencia al aumento de probabilidad que el cliente compre el producto *y* en base a los productos *x*.

*Tabla 10: Resultados de aplicación de algoritmo Apriori*

| Productos adquiridos (lhs) | Posible producto a adquirir (rhs) | Soporte | Confianza | Lift |
|---|---|---|---|---|
| ▪ PRODUCTO-116 | PRODUCTO-92 | 0.010738401 | 0.7180417 | 28.109580 |
| ▪ PRODUCTO-37<br>▪ PRODUCTO-4 | PRODUCTO-13 | 0.008514792 | 0.7001115 | 15.888007 |
| ▪ PRODUCTO-105<br>▪ PRODUCTO-35 | PRODUCTO-38 | 0.009748624 | 0.8621103 | 14.821465 |
| ▪ PRODUCTO-35<br>▪ PRODUCTO-49 | PRODUCTO-38 | 0.006928438 | 0.8176000 | 14.056240 |

| | | | | |
|---|---|---|---|---|
| ▪ PRODUCTO-105<br>▪ PRODUCTO-75 | PRODUCTO-38 | 0.008446999 | 0.8111979 | 13.946175 |

**Fuente:** Elaboración propia

Con esto concluye la ejecución de la técnica de reglas de asociación, y se puede obtener una conclusión muy buena en cuanto a la estructura del negocio y de los datos que almacena. En base a los datos que arroja el modelo, es cuestión de analizar factores estadísticos sobre las ventas y productos de mayor o menor rotación o de aquellos ítems a los que se quiera dar un impulso y revisar su relación con los demás productos para comprobar cómo se le podría dar un impulso aún mayor.

### 2.5.2. Estadísticas

Toda empresa en la actualidad debería contar con reportes estadísticos sobre sus datos que permitan tomar decisiones de manera rápida y controlar los sectores más críticos, ya sea con ayuda de herramientas ofimáticas o con sistemas especializados, ya que este análisis permite conocer el estado de la entidad en base a la realidad, ya que aunque las cuentas contables reflejen los datos reales de la empresa, no es muy práctico revisar los datos manualmente, establecer diferencias sobre las compras y ventas y determinar la utilidad de la empresa. En este caso puntual, para la empresa ABC, se desarrolló un análisis estadístico sobre los campos más importantes de la organización: compras, ventas, clientes e inventario. Aunque las compras y ventas están relacionados con clientes e inventario de productos, es importante separar cada entidad como un factor de análisis diferente para evaluar las áreas puntuales de cada uno y determinar posteriormente la cohesión de toda la información.

Es de recalcar que las gráficas presentadas en el desarrollo de esta sección, corresponden al total general de los datos analizados, en este caso de los últimos 3 años correspondiente al período del 2018 hasta el mes de agosto del 2020, salvo excepciones que serán detallados en los gráficos en que se analice solo ciertos periodos puntuales.

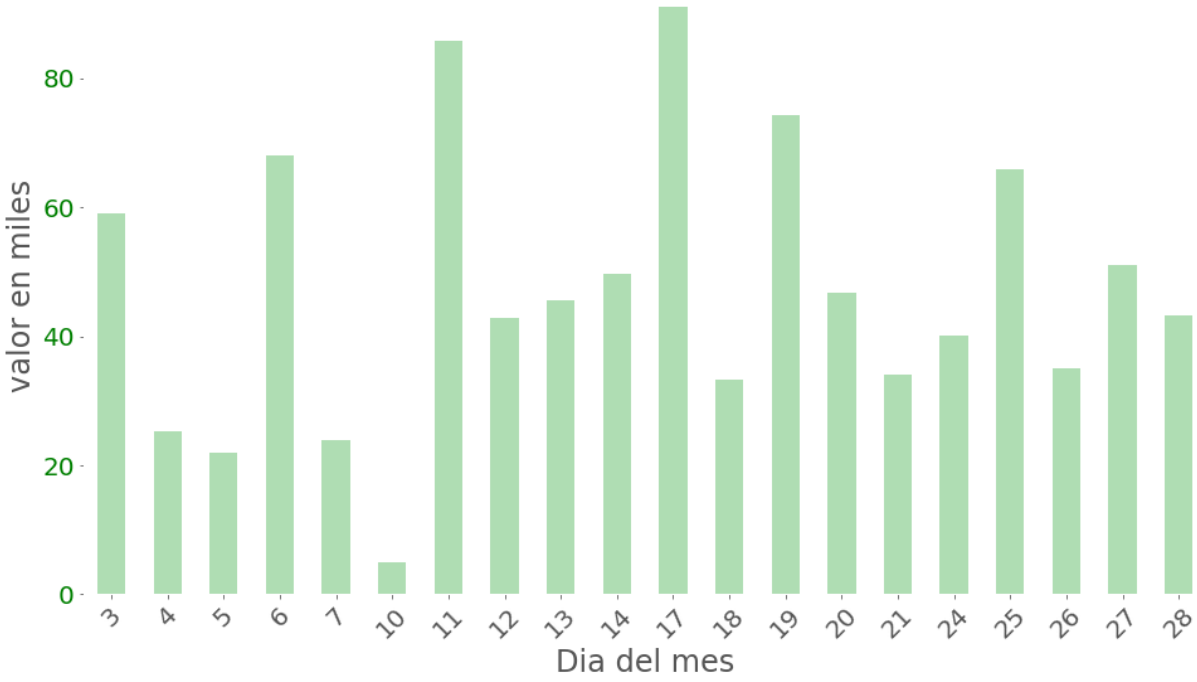
#### 2.5.2.1. Ventas

Las ventas es sin duda el sector más importante a analizar en una empresa, debido a que la misma se mantiene o genera una ganancia en base a si las ventas son mayores a todos los gastos generados durante el proceso.

*Ventas diarias*

En el gráfico 29 se presentan las ventas ordenadas por día del mes de agosto, analizando este gráfico se puede determinar los días en que más se vende y establecer en el caso que en ese periodo existiera algún día feriado si varía o no el comportamiento de las ventas.

*Gráfico 29: Estadística - Ventas diarias del mes de Agosto 2020*

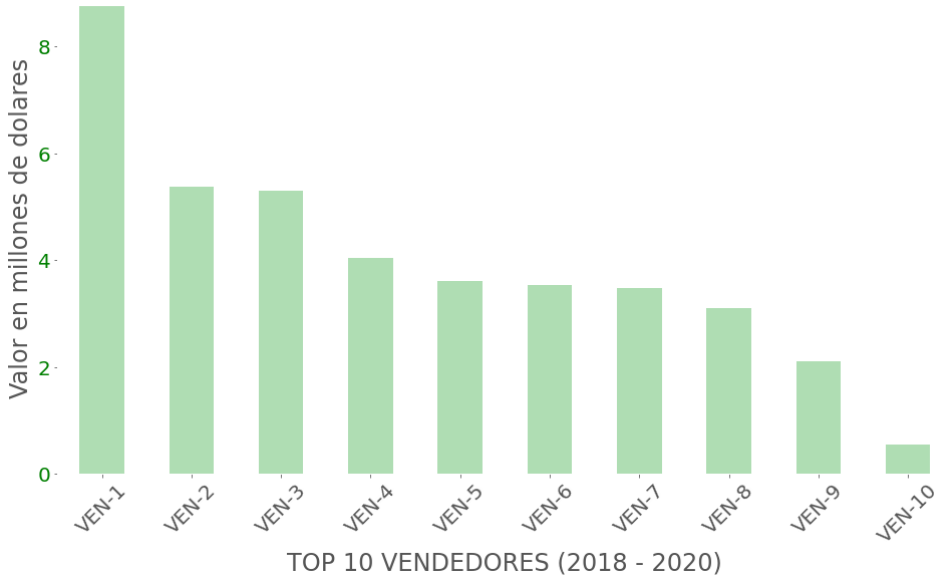



**Fuente:** Elaboración propia

*Ventas por vendedor*

El reporte de top de vendedores permite visualizar principalmente quienes son los ejecutivos de ventas que más ingresos generan a la empresa, además es útil para comprobar su rendimiento individual, (Ver gráfico 30).

*Gráfico 30: Estadística - Top vendedores (2018 – 2020)*

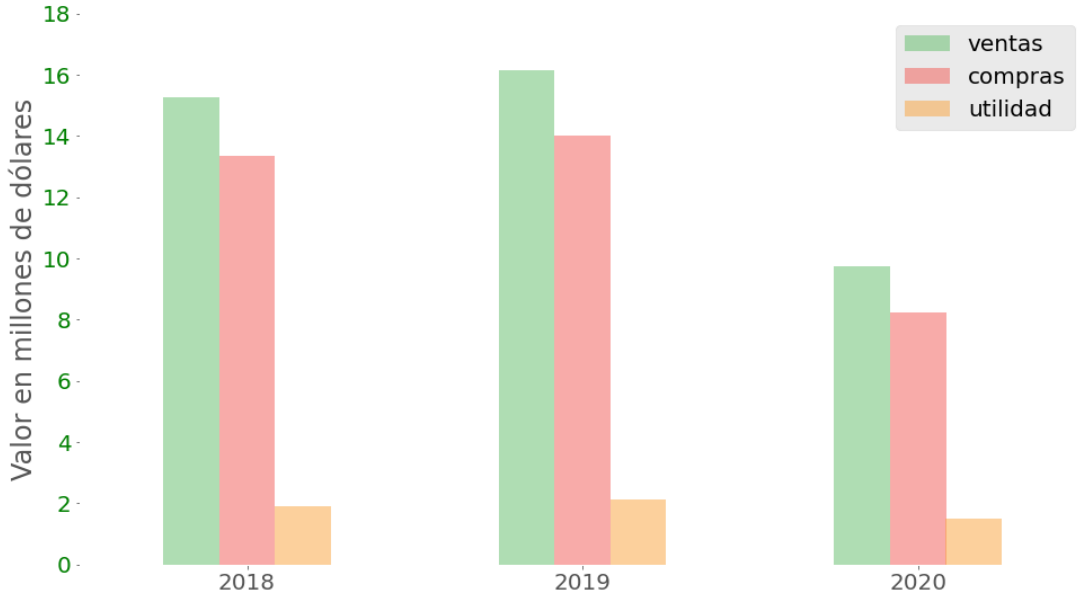



**Fuente:** Elaboración propia

### 2.5.2.2. Compras

La estadística de las ventas contra las compras es de gran ayuda para los encargados de adquisiciones de productos para comprobar si se está teniendo una concordancia entre lo que se compra versus lo que se gana, (Ver gráfico 31).

*Gráfico 31: Estadística - Ventas vs Compras vs Utilidad por año*

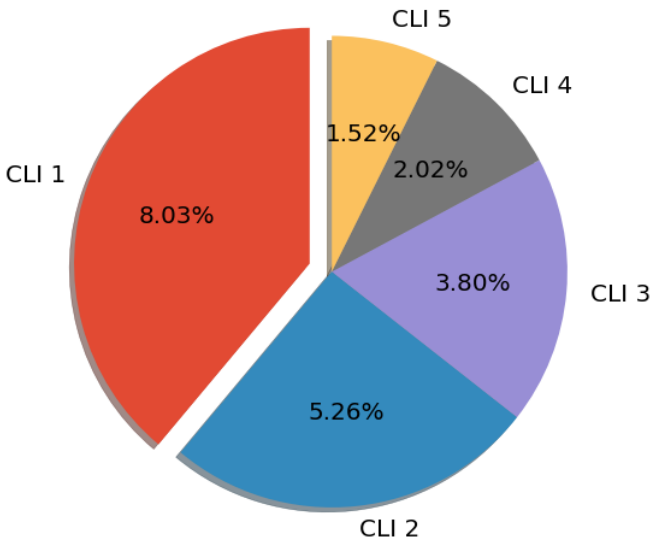



**Fuente:** Elaboración propia

### 2.5.2.3. Clientes

El reporte de los mejores clientes sirve para determinar quiénes pueden tener un trato especial frente al resto de clientes, pueden acceder a descuentos especiales, o simplemente para tener en cuenta quiénes son esos clientes de un trato diferente, (Ver gráfico 32).

*Gráfico 32: Estadística - Top clientes (2018 - 2020)*

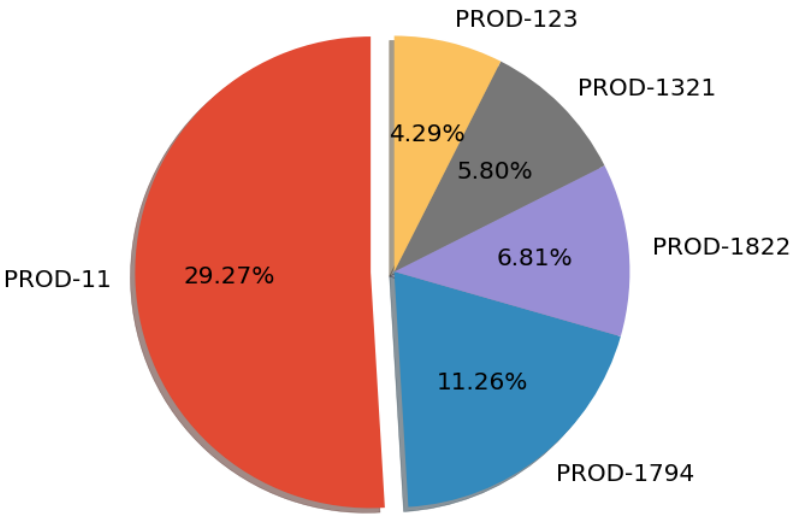



**Fuente:** Elaboración propia

### 2.5.2.4. Productos

El gráfico 33 ayuda a visualizar los 5 productos más vendidos en cantidad, aunque eso no significa que sea el que más dinero hace que ingrese a la empresa, y ese es precisamente el gráfico que ya se encuentra dentro del Dashboard.

*Gráfico 33: Estadística - Top productos por cantidad (2018 – 2020)*



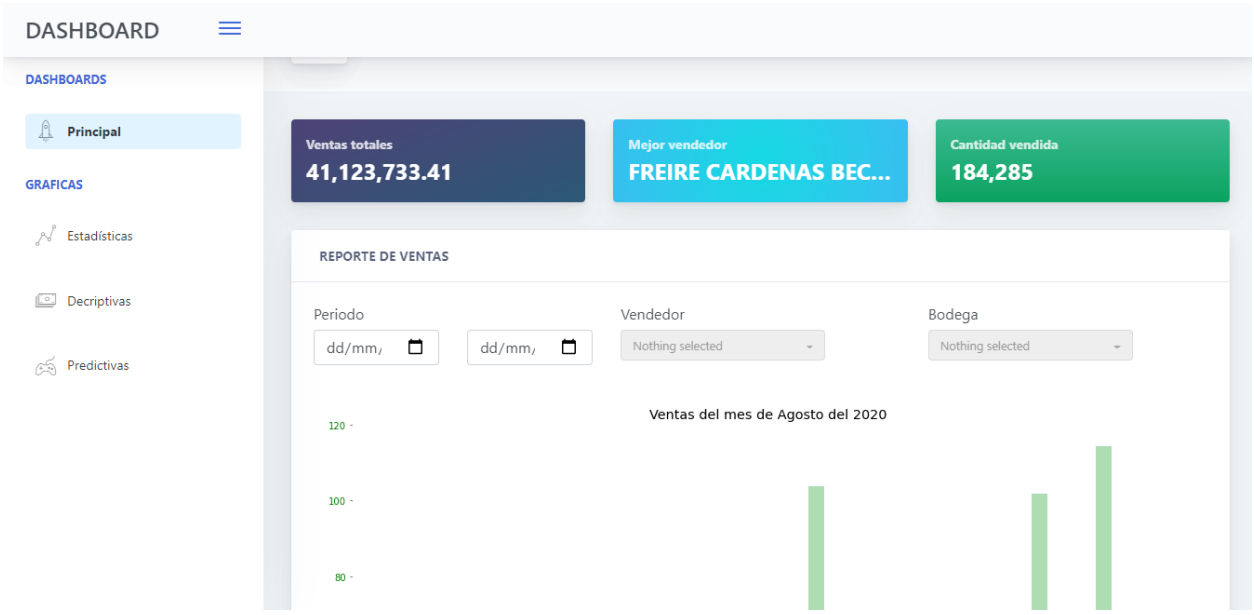

**Fuente:** Elaboración propia

### *2.5.3. Dashboard*

Para poder presentar todos los gráficos realizados durante este proyecto, se realizó un Dashboard desarrollado en Python con el framework Django, aunque para la presentación de las estadísticas se utilizó la librería de Matplotlib por la flexibilidad al graficar series temporales y sus predicciones.

Como se puede ver en el gráfico 34 se dividió en tres secciones la presentación del análisis, el primer apartado es solamente para estadística básica, donde se muestran valores como la media y varianza de los datos, además de figuras estadísticas que pueden ser manipulables según los parámetros que se escojan; el siguiente apartado es de minería descriptiva y de visualización en donde se muestran algunas gráficas de comprensión de los datos que permiten visualizar los mejores meses de la empresa entre otros gráficos, y finalmente el apartado de predicción es donde se encuentran las series temporales que pueden ser modificables según los períodos quesean seleccionados.

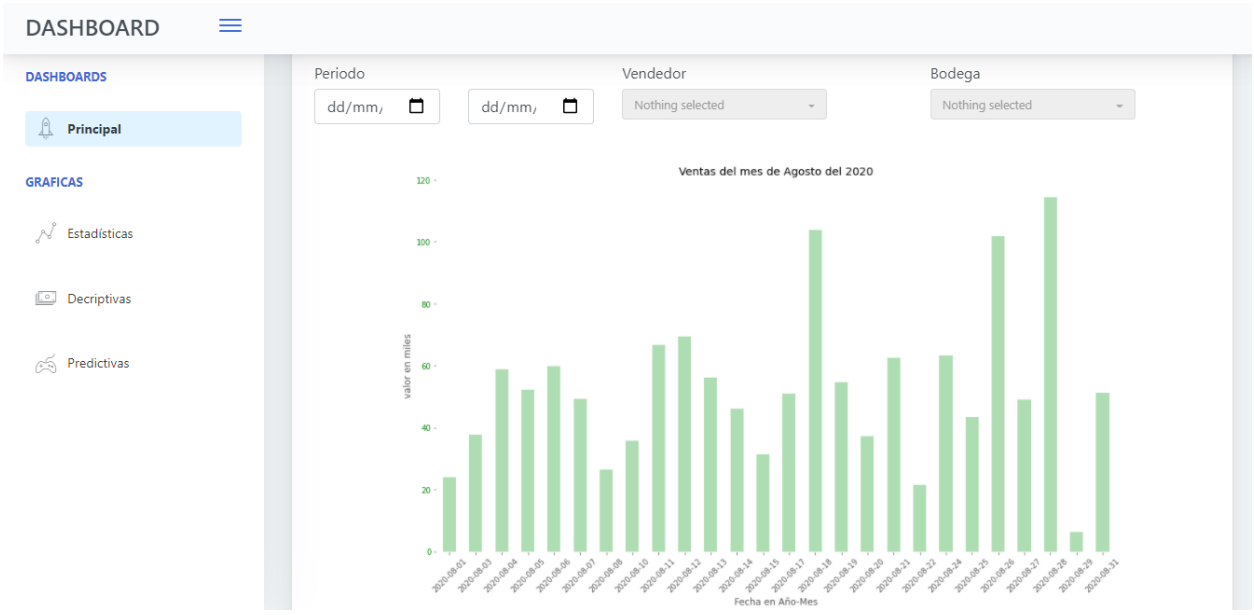
*Gráfico 34: Dashboard - Diseño*




**Fuente:** Elaboración propia

Para efectos prácticos en el gráfico 35 se presenta una demostración de los campos seleccionables en uno de los gráficos, estos parámetros son ajustables según la necesidad del gráfico.

*Gráfico 35: Dashboard - carga y visualización de figuras*




**Fuente:** Elaboración propia

# 3. CAPÍTULO III. EVALUACIÓN DEL PROTOTIPO

## 3.1. Plan de evaluación

### 3.1.1. Evaluación de los modelos predictivos

Para evaluar la precisión de los modelos de las series temporales realizados se deben realizar medidas con fórmulas en base a los resultados obtenidos, por ello en la tabla 11, se detallan las métricas que se utilizarán para comprobar cuál de los modelos se adapta mejor a los datos de la empresa ABC.

*Tabla 11: Métricas de evaluación de modelos predictivos*

| Métricas | Fórmula | Descripción |
|---|---|---|
| Error absoluto medio | $MAE = \frac{\sum_{i=0}^{n} \lvert y_i - \hat{y}_i \rvert}{n}$ | Valor promedio de los valores absolutos de la desviación. |
| Desviación mediana absoluta | $MedAE(y, \hat{y}) = median(\lvert y_1 - \hat{y}_1 \rvert, \lvert y_2 - \hat{y}_2 \rvert, \dots, \lvert y_n - \hat{y}_n \rvert)$ | Valor mediano de las diferencias absolutas. |
| Error cuadrático medio | $MSE = \frac{\sum_{i=0}^{n} (y_i - \hat{y}_i)^2}{n}$ | Promedio del cuadrado del error pronosticado. |
| Error absoluto medio porcentual | $MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert$ | Medida del error porcentual absoluto medio. |

**Fuente:** Elaboración propia

## 3.2. Resultados de la evaluación

### 3.2.1. Evaluación de los modelos predictivos

En la tabla 12 se muestra los valores resultantes luego de ejecutar las fórmulas de cada métrica utilizando el conjunto de datos reales vs los obtenidos por el modelo.

*Tabla 12: Resultados de aplicación de métricas*

| Modelo | MAE | RMSE | MPE | MAPE |
|---|---|---|---|---|
| ARIMA (2,1,2)(2,1,1)[48] PYTHON | 105565 | 150902 | 0.0129 | 0.0971 |
| ARIMA (0,1,1)(2,1,1)[48] R | 56546 | 97414 | 0.8846 | 0.0520 |
| Holt-Winters PYTHON | 73327 | 110810 | -0.008 | 0.0628 |
| Holt-Winters R | 78741 | 87541 | -0.3547 | 0.0851 |
| Suavizado Exponencial R | 60949 | 78603 | -0.0557 | 0.0551 |
| Suavizado Exponencial y Autorregresión R | 55811 | 72250 | -0.2986 | 0.0506 |
| Suavizado Exponencial y ARIMA R | 59045 | 76546 | -0.0393 | 0.0536 |
| TBATS R | 69884 | 86789 | -0.4882 | 0.0629 |
| NNETAR R | 27977 | 37925 | -0.29763 | 0.0252 |

| RESULTADOS<br>Mejor modelo según métrica | NNETAR | NNETAR | HOLT-WINTERS PYTHON | NNETAR |
|---|---|---|---|---|

**Fuente:** Elaboración propia

En base a los resultados obtenidos en la tabla 12 podemos deducir que el mejor modelo de predicción aplicado a los datos de la empresa ABC es el modelo NNETAR por una diferencia considerable, aunque cabe mencionar que el modelo que más se ajusta a los datos de manera visual pareciera ser el modelo de Holt-Winters.

### 3.2.2. Comprobación de valores reales vs predichos

*Tabla 13: Comparación de predicciones con valores reales*

| Modelo | VR 09-20 | VP 09-20 | VR 10-20 | VP 10-20 | VR 11-20 | VP 11-20 | Porcentaje de error |
|---|---|---|---|---|---|---|---|
| ARIMA (2,1,2)(2,1,1)[48] PYTHON | 1392512 | 1553339 | 1388124 | 1511981 | 1202065 | 1310992 | 3.29% |
| Holt-Winters R | 1392512 | 1932344 | 1388124 | 1885176 | 1202065 | 2069344 | 2.40% |
| Holt-Winters PYTHON | 1392512 | 1421707 | 1388124 | 1197792 | 1202065 | 1115964 | 2.07% |
| NNETAR R | 1392512 | 1562993 | 1388124 | 1367261 | 1202065 | 1297609 | 2.05% |
| TBATS R | 1392512 | 1329516 | 1388124 | 1265339 | 1202065 | 1152185 | 1.97% |
| Suavizado Exponencial y Autorregresión R | 1392512 | 1375161 | 1388124 | 1329506 | 1202065 | 1083974 | 1.62% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Suavizado Exponencial R | 1392512 | 1387994 | 1388124 | 1317398 | 1202065 | 1108852 | 1.41% |
| ARIMA (0,1,1)(2,1,1)[48] R | 1392512 | 1384134 | 1388124 | 1345379 | 1202065 | 1144712 | 0.91% |
| Suavizado Exponencial y ARIMA R | 1392512 | 1358440 | 1388124 | 1369483 | 1202065 | 1147200 | 0.90% |

Como podemos observar en la tabla 13, el modelo que mejor se ajusta a los datos comparando las ventas de los meses de septiembre, octubre y noviembre con las predicciones dadas por cada técnica es el de suavizado exponencial combinado con ARIMA.

## CONCLUSIONES

- Se realizó un análisis profundo de los datos de la empresa ABC, en la que se descubrieron patrones de comportamiento en las ventas que permitieron realizar gráficas predictivas con el uso de modelos estadísticos como ARIMA y Holt-Winters, además, se encontró relaciones entre los productos que compran los clientes utilizando reglas de asociación.
- Aplicando conceptos matemáticos como la media, varianza, frecuencia y porcentajes se construyó una serie de gráficos estadísticos que permiten conocer el estado actual de la empresa; con respecto a las ventas, permiten evaluar el desempeño tanto de los vendedores como de las estrategias que se ponen en marcha; en las compras a proveedores, ayuda a determinar si es factible o no adquirir más mercadería, y en los demás aspectos como los productos o clientes, los gráficos informativos ayudan a visualizar la información de forma clara y tomar mejores decisiones.
- Para poder visualizar tanto los gráficos estadísticos, descriptivos y predictivos se desarrolló un sistema integrador en el Python utilizando el framework Django, que permite manipular la información de manera sencilla y ajustar parámetros relevantes como el período de los datos.
- Los modelos predictivos desarrollados fueron evaluados para escoger el que mejor se adaptaba a los datos de la empresa, garantizando de esa manera que las predicciones que realice serán acertadas y tendrán el menor error posible, en donde se determinó que el mejor modelo es el Holt-Winters desarrollado en Python.
- Se utilizó la metodología CRISP-DM para el desarrollo del proyecto, esto ayudó de gran manera ya que posee fases que esclarecen los procesos que se deben realizar en cada momento y existe mucha documentación sobre cada etapa.

## RECOMENDACIONES

- Aunque se realizaron varias técnicas de minería sobre los datos, existen muchos más métodos que se pueden aplicar, no solo de minería sino ciencias mucho más avanzadas como machine learning o Deep learning que pueden ser ejecutadas para potenciar las ventas, reducir costes, mejorar los procesos y en general, aumentar la productividad de la empresa.
- Los gráficos son de gran ayuda para comprender la situación de una entidad en el mercado, pero si no se interpretan de la manera correcta puede significar una pérdida para la misma, por lo tanto, en necesario que las personas que se encarguen de tomar las decisiones reciban asesoría sobre la intención real de los gráficos, y entiendan que se está evaluando en cada situación.
- El sistema integrador desarrollado, aunque cumple con la función de presentar los gráficos, no posee módulos de seguridad que garanticen la privacidad de los datos, por lo que no es recomendable que sea puesto en producción sin antes implementar al menos alguna función de seguridad que garantice el acceso solo de las personas autorizadas.
- Probar más modelos predictivos diferentes a los planteados para comparar y verificar el modelo actual, con el fin de encontrar un mejor modelo que se ajuste a los datos de la empresa y plantee nuevas formas de analizar los datos.
- Escoger una metodología desde el principio, planificar y seguir estrictamente el cronograma establecido es esencial para aumentar el éxito de un proyecto de minería de datos.

## BIBLIOGRAFÍA


[1] H. Li, Y. J. Wu, y Y. Chen, «Time is money: Dynamic-model-based time series data-mining for correlation analysis of commodity sales», p. 27.

[2] J. Kim y J. Y. Lee, «Server-Edge dualized closed-loop data analytics system for cyber-physical system application», *Robot. Comput.-Integr. Manuf.*, vol. 67, p. 102040, feb. 2021, doi: 10.1016/j.rcim.2020.102040.

[3] B. Mazon-Olivo, W. Rivas-Asanza, J. Novillo-Vicuña, y C. Flores-Cabrera, «Análisis de producción avícola mediante técnicas de inteligencia de negocios y minería de datos», *Alternativas*, vol. 19, n.º 2, Art. n.º 2, ago. 2018, doi: 10.23878/alternativas.v19i2.203.

[4] B. Mazon-Olivo, M. Pinta, y F. Refrovan, «Desarrollo de competencias en Minería de Datos, una experiencia didáctica», en *Sistematización de experiencias educativas innovadoras*, 1.ª ed., Universidad Técnica de Machala, 2020, pp. 383-406.

[5] S. J. Qin y L. H. Chiang, «Advances and opportunities in machine learning for process data analytics», *Comput. Chem. Eng.*, vol. 126, pp. 465-473, jul. 2019, doi: 10.1016/j.compchemeng.2019.04.003.

[6] L. N. Sanchez-Pinto, Y. Luo, y M. M. Churpek, «Big Data and Data Science in Critical Care», *Chest*, vol. 154, n.º 5, pp. 1239-1248, nov. 2018, doi: 10.1016/j.chest.2018.04.037.

[7] J. R. Saura, «Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics», *J. Innov. Knowl.*, p. S2444569X20300329, ago. 2020, doi: 10.1016/j.jik.2020.08.001.

[8] B. Mazon-Olivo, A. Pan, y R. Tinoco-Egas, «Inteligencia de negocios en el sector agropecuario», en *Análisis de Datos Agropecuarios*, First Ed., Eds. Machala-Ecuador: Universidad Técnica de Machala, 2018, pp. 246–278.

[9] E. Szymańska, «Modern data science for analytical chemical data – A comprehensive review», *Anal. Chim. Acta*, vol. 1028, pp. 1-10, oct. 2018, doi: 10.1016/j.aca.2018.05.038.

[10] I. Ramírez-Morales, B. Mazon-Olivo, y A. Pan, «Ciencia de datos en el sector agropecuario», en *Análisis de Datos Agropecuarios*, 1.ª ed., Eds. Machala-Ecuador: Universidad Técnica de Machala, 2018, pp. 12–44.

[11] A. M. Jimenez-Carvelo, «Data mining/machine learning methods in foodomics», *Curr. Opin. Food Sci.*, vol. 37, pp. 76-82, feb. 2021, doi: 10.1016/j.cofs.2020.09.008.

[12] W. Rivas-Asanza, B. Mazon-Olivo, y E. Mejía-Peñafiel, «Generalidades de las redes neuronales artificiales», en *Redes neuronales artificiales aplicadas al reconocimiento de patrones*, 1.ª ed., Eds. Universidad Técnica de Machala, 2018.

[13] C. Gutierrez-Osorio y C. Pedraza, «Modern data sources and techniques for analysis and forecast of road accidents: A review», *J. Traffic Transp. Eng. Engl. Ed.*, vol. 7, n.º 4, pp. 432-446, ago. 2020, doi: 10.1016/j.jtte.2020.05.002.

[14] P. Sunhare, R. R. Chowdhary, y M. K. Chattopadhyay, «Internet of things and data mining: An application oriented survey», *J. King Saud Univ. - Comput. Inf. Sci.*, p. S131915782030416X, jul. 2020, doi: 10.1016/j.jksuci.2020.07.002.

[15] S. M. Drayton-Brooks, P. A. Gray, N. P. Turner, y J. A. Newland, «The use of big data and data mining in nurse practitioner clinical education», *J. Prof. Nurs.*, p. S8755722320300788, mar. 2020, doi: 10.1016/j.profnurs.2020.03.012.

[16] B. MAZÓN-Olivo, M. JARAMILLO-Paredes, O. ROMERO-Hidalgo, A. Borja, M. AGUIRRE-Benalcazar, y M. CONTENTO-Segarra, «Tecnologías de Inteligencia de Negocios y Minería de datos para el análisis de la producción y comercialización de cacao», p. 15.

[17] Q. Zheng, Y. Li, y J. Cao, «Application of data mining technology in alarm analysis of communication network», *Comput. Commun.*, vol. 163, pp. 84-90, nov. 2020, doi: 10.1016/j.comcom.2020.08.012.

[18] H. Zhou, J. Zhang, Y. Zhou, X. Guo, y Y. Ma, «A feature selection algorithm of decision tree based on feature weight», *Expert Syst. Appl.*, vol. 164, p. 113842, feb. 2021, doi: 10.1016/j.eswa.2020.113842.

[19] G. Hu, S. Mohammadiun, A. A. Gharahbagh, J. Li, K. Hewage, y R. Sadiq, «Selection of oil spill response method in Arctic offshore waters: A fuzzy decision tree based framework», *Mar. Pollut. Bull.*, vol. 161, p. 111705, dic. 2020, doi: 10.1016/j.marpolbul.2020.111705.

[20] A. Beucher, A. B. Møller, y M. H. Greve, «Artificial neural networks and decision tree classification for predicting soil drainage classes in Denmark», *Geoderma*, vol. 352, pp. 351-359, oct. 2019, doi: 10.1016/j.geoderma.2017.11.004.

[21] M. Sabah, M. Talebkeikhah, F. Agin, F. Talebkeikhah, y E. Hasheminasab, «Application of decision tree, artificial neural networks, and adaptive neuro-fuzzy inference system on predicting lost circulation: A case study from Marun oil field», *J. Pet. Sci. Eng.*, vol. 177, pp. 236-249, jun. 2019, doi: 10.1016/j.petrol.2019.02.045.

[22] A. Augello, I. Infantino, G. Pilato, y F. Vella, «Sensing the Web for Induction of Association Rules and their Composition through Ensemble Techniques», *Procedia Comput. Sci.*, vol. 169, pp. 851-859, 2020, doi: 10.1016/j.procs.2020.02.152.

[23] J. Hong, R. Tamakloe, y D. Park, «Application of association rules mining algorithm for hazardous materials transportation crashes on expressway», *Accid. Anal. Prev.*, vol. 142, p. 105497, jul. 2020, doi: 10.1016/j.aap.2020.105497.

[24] X. Dong, F. Hao, L. Zhao, y T. Xu, «An efficient method for pruning redundant negative and positive association rules», *Neurocomputing*, vol. 393, pp. 245-258, jun. 2020, doi: 10.1016/j.neucom.2018.09.108.

[25] X. Guo, D. Z. W. Wang, J. Wu, H. Sun, y L. Zhou, «Mining commuting behavior of urban rail transit network by using association rules», *Phys. Stat. Mech. Its Appl.*, vol. 559, p. 125094, dic. 2020, doi: 10.1016/j.physa.2020.125094.

[26] M. Nasr, M. Hamdy, D. Hegazy, y K. Bahnasy, «An efficient algorithm for unique class association rule mining», *Expert Syst. Appl.*, vol. 164, p. 113978, feb. 2021, doi: 10.1016/j.eswa.2020.113978.

[27] L. Baroni *et al.*, «An analysis of malaria in the Brazilian Legal Amazon using divergent association rules», *J. Biomed. Inform.*, vol. 108, p. 103512, ago. 2020, doi: 10.1016/j.jbi.2020.103512.

[28] J. Maia *et al.*, «Evolving clustering algorithm based on mixture of typicalities for stream data mining», *Future Gener. Comput. Syst.*, vol. 106, pp. 672-684, may 2020, doi: 10.1016/j.future.2020.01.017.

[29] K. Fukui, Y. Okada, K. Satoh, y M. Numao, «Cluster sequence mining from event sequence data and its application to damage correlation analysis», *Knowl.-Based Syst.*, vol. 179, pp. 136-144, sep. 2019, doi: 10.1016/j.knosys.2019.05.012.

[30] M. J. Zaki y W. Meira, Jr, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2.ª ed. Cambridge University Press, 2020.

[31] H. Akoglu, «User's guide to correlation coefficients», *Turk. J. Emerg. Med.*, vol. 18, n.º 3, pp. 91-93, sep. 2018, doi: 10.1016/j.tjem.2018.08.001.

[32] C. Katris, «A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece», *Expert Syst. Appl.*, vol. 166, p. 114077, mar. 2021, doi: 10.1016/j.eswa.2020.114077.

[33] W. Jiang, X. Wu, Y. Gong, W. Yu, y X. Zhong, «Holt–Winters smoothing enhanced by fruit fly optimization algorithm to forecast monthly electricity consumption», *Energy*, vol. 193, p. 116779, feb. 2020, doi: 10.1016/j.energy.2019.116779.

[34] C. Liu, B. Sun, C. Zhang, y F. Li, «A hybrid prediction model for residential electricity consumption using holt-winters and extreme learning machine», *Appl. Energy*, vol. 275, p. 115383, oct. 2020, doi: 10.1016/j.apenergy.2020.115383.

[35] D. Barrow, N. Kourentzes, R. Sandberg, y J. Niklewski, «Automatic robust estimation for exponential smoothing: Perspectives from statistics and machine learning», *Expert Syst. Appl.*, vol. 160, p. 113637, dic. 2020, doi: 10.1016/j.eswa.2020.113637.

[36] M. Liu, J. W. Taylor, y W.-C. Choo, «Further empirical evidence on the forecasting of volatility with smooth transition exponential smoothing», *Econ. Model.*, vol. 93, pp. 651-659, dic. 2020, doi: 10.1016/j.econmod.2020.02.021.

[37] M. Gos, J. Krzyszczak, P. Baranowski, M. Murat, y I. Malinowska, «Combined TBATS and SVM model of minimum and maximum air temperatures applied to wheat yield prediction at different locations in Europe», *Agric. For. Meteorol.*, vol. 281, p. 107827, feb. 2020, doi: 10.1016/j.agrformet.2019.107827.

[38] M. Kalantari, «Forecasting COVID-19 Pandemic Using Optimal Singular Spectrum Analysis», *Chaos Solitons Fractals*, p. 110547, dic. 2020, doi: 10.1016/j.chaos.2020.110547.

[39] N. Wang, M. Reformat, W. Yao, Y. Zhao, y X. Chen, «Fuzzy Linear regression based on approximate Bayesian computation», *Appl. Soft Comput. J.*, vol. 97, oct. 2020, doi: 10.1016/j.asoc.2020.106763.

[40] P. Mccaffrey, «A selective introduction to Python and key concepts», en *An Introduction to Healthcare Informatics*, Elsevier, 2020, pp. 145-157.

[41] J. E. Tomlinson, J. H. Arnott, y J. J. Harou, «A water resource simulator in Python», *Environ. Model. Softw.*, vol. 126, p. 104635, abr. 2020, doi: 10.1016/j.envsoft.2020.104635.

[42] «Welcome to Python.org», *Python.org*. https://www.python.org/ (accedido oct. 29, 2020).

[43] «R: The R Project for Statistical Computing». https://www.r-project.org/ (accedido nov. 29, 2020).

[44] C. E. Galván-Tejada *et al.*, «Demographic and Comorbidities Data Description of Population in Mexico with SARS-CoV-2 Infected Patients(COVID19): An Online Tool Analysis», *Int. J. Environ. Res. Public. Health*, vol. 17, n.º 14, p. 5173, jul. 2020, doi: 10.3390/ijerph17145173.

[45] «Project Jupyter». https://www.jupyter.org (accedido nov. 29, 2020).

[46] «RStudio | Open source & professional software for data science teams». https://rstudio.com/ (accedido nov. 30, 2020).

[47] «The Web framework for perfectionists with deadlines | Django». https://www.djangoproject.com/ (accedido nov. 30, 2020).

[48] A. Sunardi y Suharjito, «MVC Architecture: A Comparative Study Between Laravel Framework and Slim Framework in Freelancer Project Monitoring System Web Based», *Procedia Comput. Sci.*, vol. 157, pp. 134-141, 2019, doi: 10.1016/j.procs.2019.08.150.

[49] «pandas - Python Data Analysis Library». https://pandas.pydata.org/ (accedido nov. 30, 2020).

[50] S. Huber, H. Wiemer, D. Schneider, y S. Ihlenfeldt, «DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model», *Procedia CIRP*, vol. 79, pp. 403-408, 2019, doi: 10.1016/j.procir.2019.02.106.

[51] A. Pinto, D. Ferreira, C. Neto, A. Abelha, y J. Machado, «Data Mining to Predict Early Stage Chronic Kidney Disease», *Procedia Comput. Sci.*, vol. 177, pp. 562-567, 2020, doi: 10.1016/j.procs.2020.10.079.

## ANEXOS

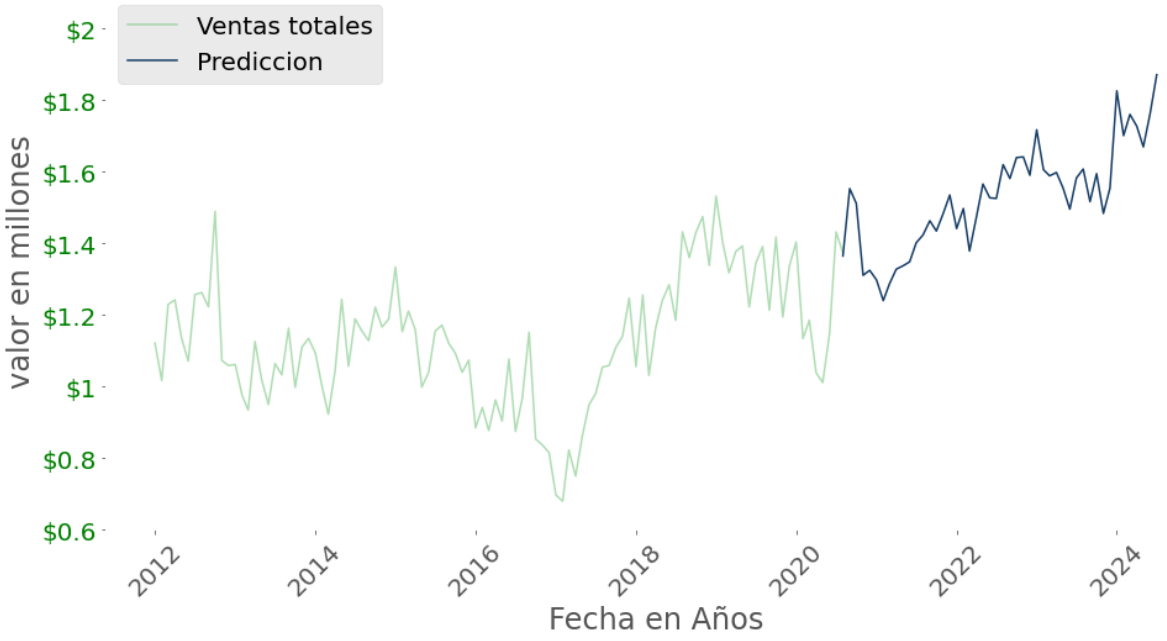
### Anexo 1. Código del modelo ARIMA en Python

**Función:** Modelo predictivo SARIMAX (2, 1, 2) x (2, 1, 1, 48)

**Lenguaje:** Python

**Código:**

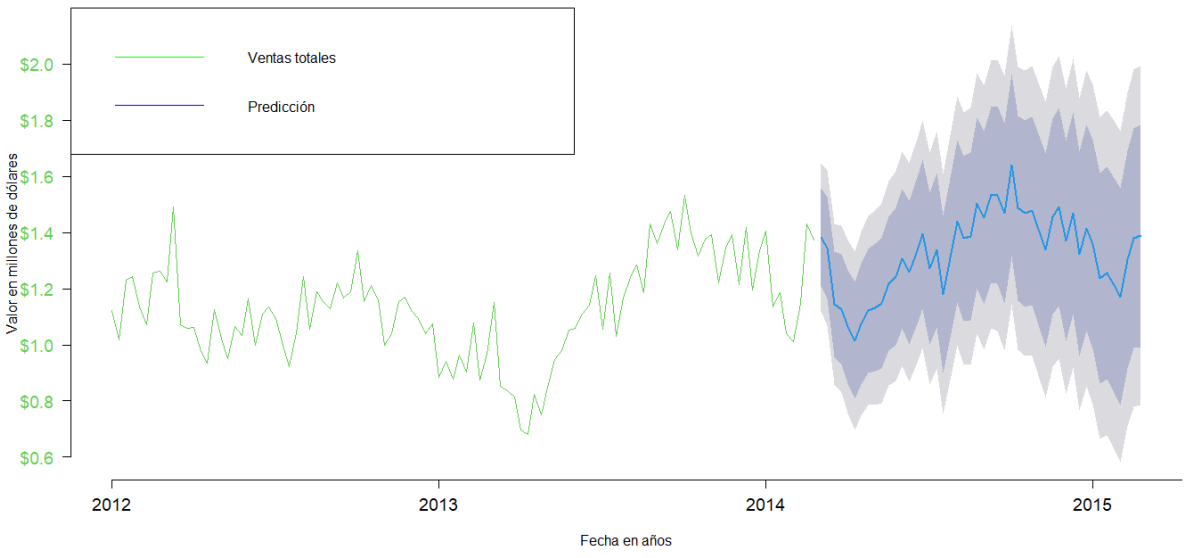
```
# Entrenar el modelo
import statsmodels.api as sm
model=sm.tsa.statespace.SARIMAX(df['total'],order=(2, 1,
2),seasonal_order=(2,1,1,48))
results=model.fit()
print(results.summary())
df['forecast']=results.predict(start=5,end=103,dynamic=False)
df[['total','forecast']].plot(figsize=(12,8))

# Prediccion de datos
from pandas.tseries.offsets import DateOffset
future_dates=[df.index[-1]+ DateOffset(months=x)for x in range(0,48)]
future_datest_df=pd.DataFrame(index=future_dates[1:],columns=df.columns)

future_datest_df.tail()

future_df=pd.concat([df,future_datest_df])

future_df['forecast'] = results.predict(start = 103, end = 184, dynamic=
True)

# Graficación de serie temporal
plts = future_df[['total', 'forecast']][72:].plot(figsize=(16, 8),
color=[colors['green_a'],colors['blue']])
plt.xticks(rotation=45)
scale_y = 1e6
ticks_y = ticker.FuncFormatter(lambda x, pos: '{0:g}'.format(x/scale_y))
plts.yaxis.set_major_formatter(ticks_y)
plts.yaxis.set_tick_params(which='major', labelcolor='green')
# plts.yaxis.set_label_position("right")
# plts.yaxis.tick_right()
plts.set_ylabel('valor en millones')
plts.set_xlabel('Fecha en Años')
plts.set_facecolor('#fff')
plt.title("Predicción de ventas en los próximos 5 años - SARIMAX(2, 1,
2)x(2, 1, [1], 48)")

plt.show()
```


Ventas totales
Prediccion
$2
$1.8
$1.6
$1.4
$1.2
$1
$0.8
$0.6
valor en millones
2012
2014
2016
2018
2020
2022
2024
Fecha en Años

## Anexo 2. Código del modelo ARIMA en R

| |
|---|
| **Función:** Modelo predictivo ARIMA (0, 1, 1) x (2, 1, 1, 48) |
| **Lenguaje:** R |
| **Código:** |

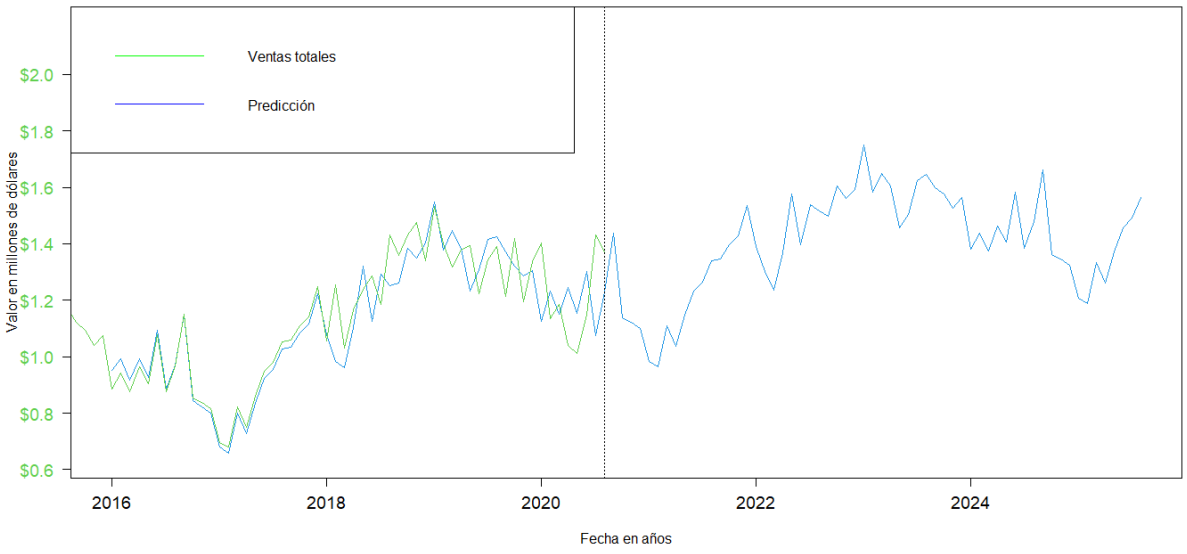
```
#Librería para encontrar datos
library(forecast)

fitARIMA <- arima(data$total, order=c(0,1,1),seasonal = list(order =
c(2,1,1), period=48),method="ML")

coeftest(fitARIMA)

confint(fitARIMA)

acf(fitARIMA$residuals)

auto.arima(data$total, trace=TRUE)

predict <- predict(fitARIMA, n.ahead = 104)
```



## Anexo 3. Código del modelo Holt-Winter en R

| **Función:** Modelo predictivo Holt-Winters |
|---|
| **Lenguaje:** R |
| **Código:** |

```
demand <- ts(data$total, start = c(2012, 1), frequency = 12)
plot(demand)

hw <- HoltWinters(demand)
plot(hw)

forecast <- predict(hw, n.ahead = 20, prediction.interval =
T, level = 0.95)
plot(hw, forecast)
```

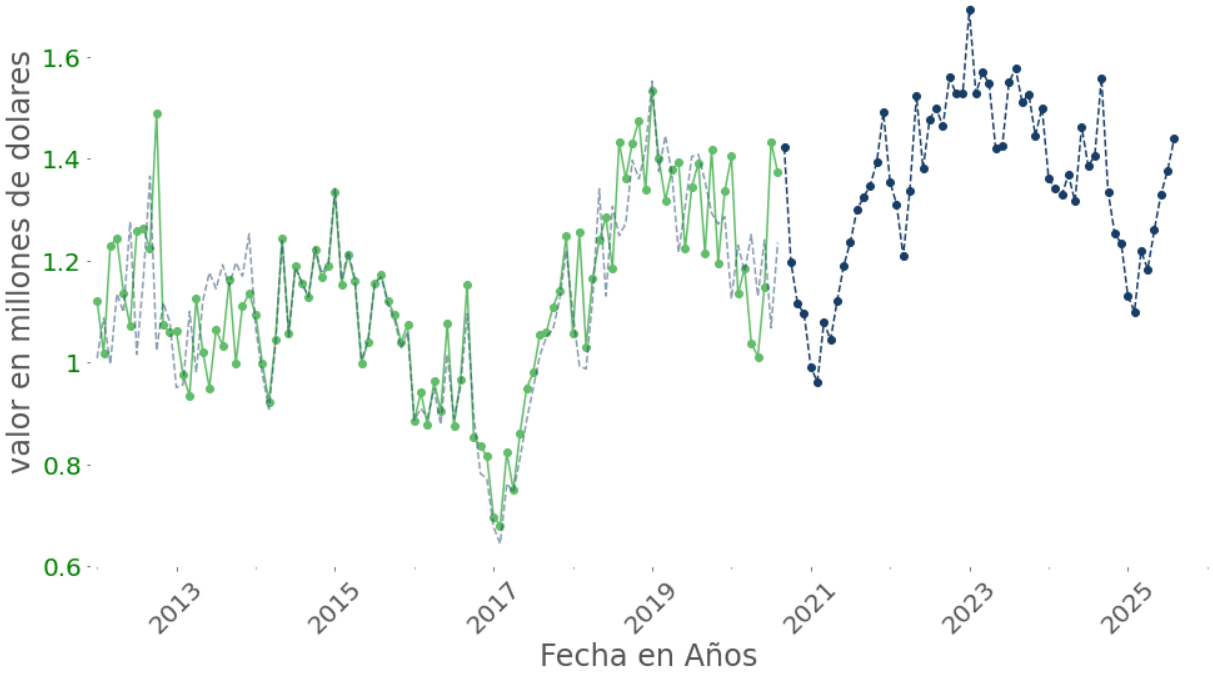


**Anexo 4. Código del modelo Holt-Winter en Python**

| **Función:** Modelo predictivo Holt-Winter |
| --- |
| **Lenguaje:** Python |
| **Código:** |

```
import matplotlib.ticker as ticker

aust = data

fit = ExponentialSmoothing(aust.total, seasonal_periods=48,
trend='mul', seasonal='add',
initialization_method="estimated").fit()
simulations = fit.simulate(40, repetitions=100, error='mul',
random_errors='bootstrap')

plts = aust.total.plot(figsize=(16,9), marker='o',
color=colors['green'],
               title="Forecasts and simulations from Holt-
Winters' multiplicative method" )
fit.fittedvalues.plot(ax=plts, style='--',
color=colors['blue_a'])
# simulations.plot(ax=ax, style='-', alpha=0.05,
color='grey', legend=False)
fit.forecast(60).plot(ax=plts, style='--', marker='o',
color=colors['blue'], legend=False)

plt.xticks(rotation=45)
scale_y = 1e6
ticks_y = ticker.FuncFormatter(lambda x, pos:
'{0:g}'.format(x/scale_y))
plts.yaxis.set_major_formatter(ticks_y)
plts.yaxis.set_tick_params(which='major', labelcolor='green'
#                            ,labelleft=False,
labelright=False
                          )
# plts.yaxis.set_label_position("right")
# plts.yaxis.tick_right()
plts.set_ylabel('valor en millones')
plts.set_xlabel('Fecha en Año-Mes')
plts.set_facecolor('#fff')
```

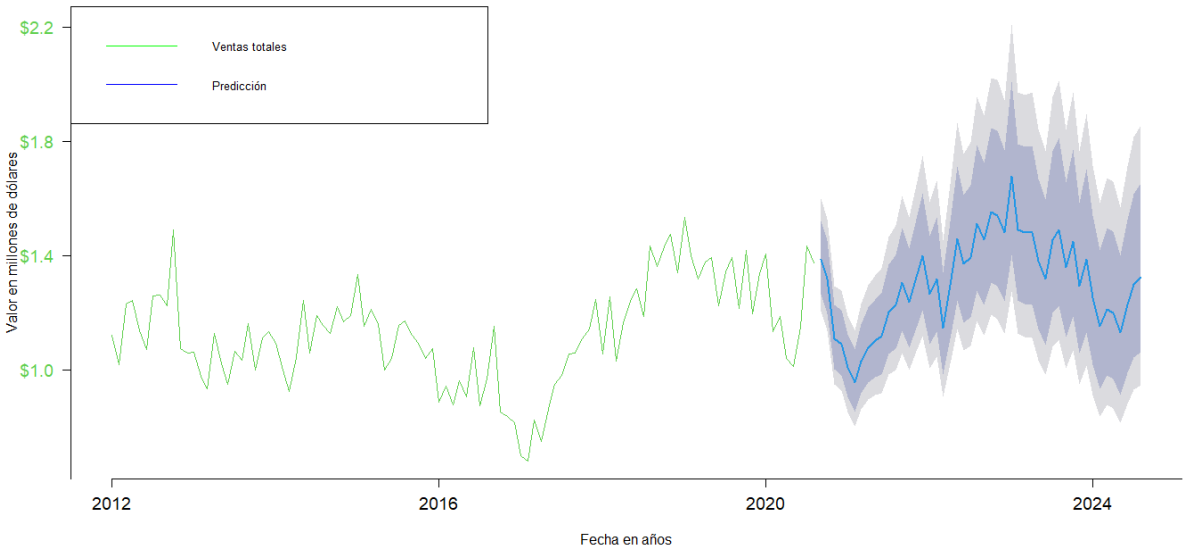
```
plt.title("Holt-Winters en las ventas mensuales")

plt.show()
print(fit.summary())
```



## Anexo 5. Código del modelo Suavizado Exponencial en R

| **Función:** Modelo predictivo Suavizado Exponencial |
|---|
| **Lenguaje:** R |
| **Código:** |

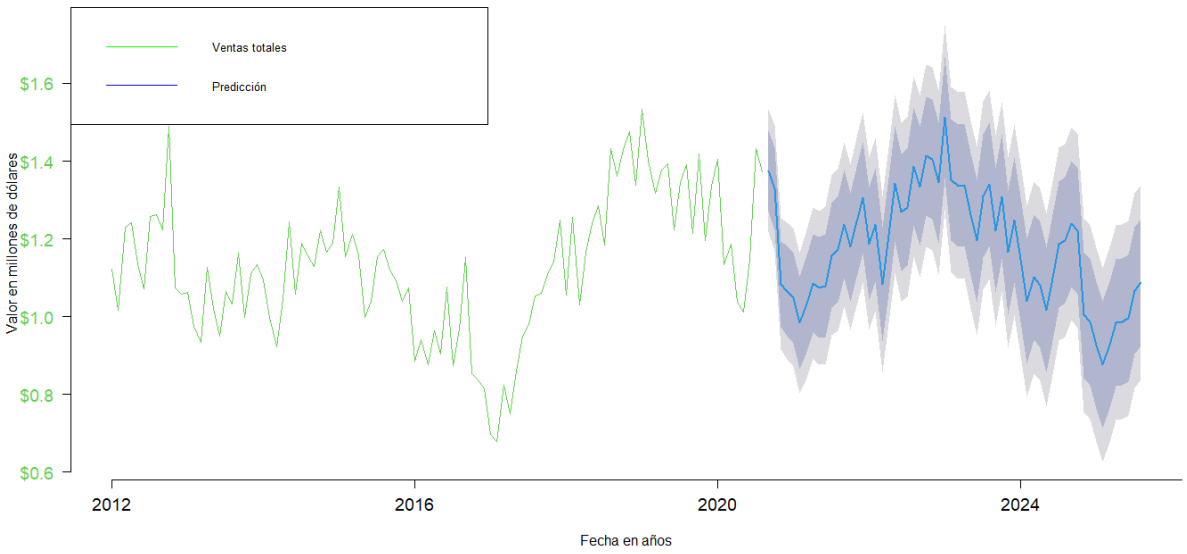
```
#STLF suavisacion exponencial
fitc2 <- stlf(serie1, lambda=0)
fc2=""
fc2<- forecast(fitc2, h=60)

xpos <- seq(600000, 2600000, by=400000)
plot(fc2, main="Predicción de ventas de los próximos 5 años
Suavizacion Exponencial STLF", axes = FALSE, xlab="Fecha en
años", ylab="Valor en millones de dólares", col=3)
axis(2, at=xpos, labels=sprintf("$%.1f", xpos/1000000),
col.axis=3, las=1, cex.axis=1.2)
axis(1,at=c(2012, 2013, 2014, 2015,
2016),labels=c("2012","2016","2020", "2024", "2028"),
cex.axis=1.2)
legend("topleft", legend=c("Ventas totales", "Predicción"),
       col=c("green", "blue"), lty=1:1, cex=.8)

accuracy(fc2)
```



**Anexo 6. Código del modelo Suavizado Exponencial con Autorregresión en R**

| |
|---|
| **Función:** Modelo predictivo Suavizado Exponencial con Autorregresión |
| **Lenguaje:** R |
| **Código:** |

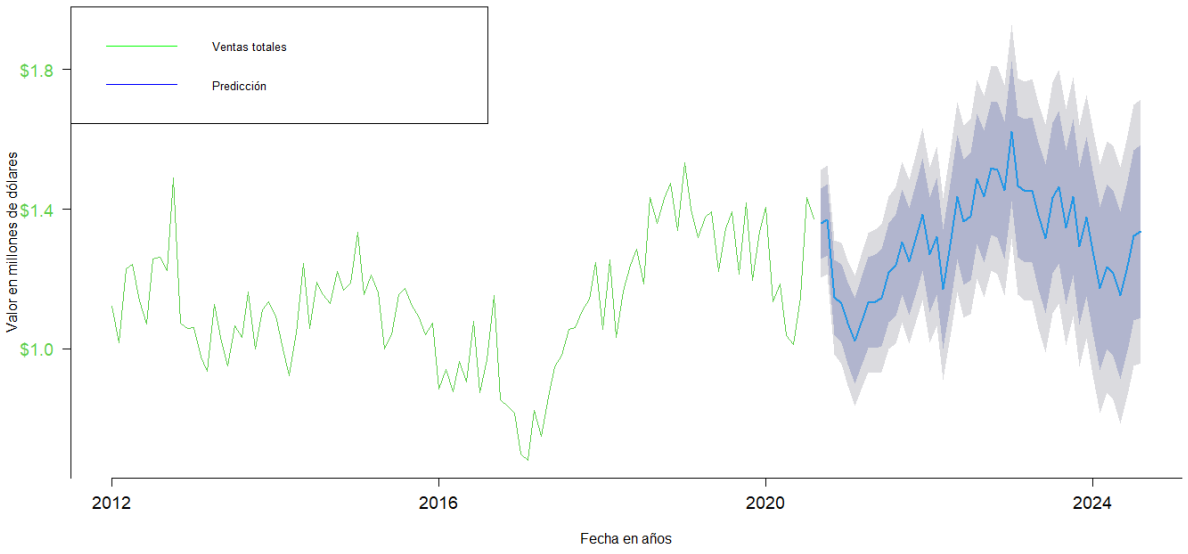
```
#modelo STLM
fitc1 <- stlm(serie1, modelfunction=ar)
fc1<- forecast(fitc1, h=60)

xpos <- seq(600000, 1800000, by=200000)
plot(fc1, main="Predicción de ventas de los próximos 5 años -
MODELO STL-Autorregresivo", axes = FALSE, xlab="Fecha en
años", ylab="Valor en millones de dólares", col=3)
axis(2, at=xpos, labels=sprintf("$%.1f", xpos/1000000),
col.axis=3, las=1, cex.axis=1.2)
axis(1,at=c(2012, 2013, 2014, 2015,
2016),labels=c("2012","2016","2020", "2024", "2024"),
cex.axis=1.2)
legend("topleft", legend=c("Ventas totales", "Predicción"),
       col=c("green", "blue"), lty=1:1, cex=.8)

accuracy(fc1)
```



## Anexo 7. Código del modelo Suavizado Exponencial y ARIMA en R

| **Función:** Modelo predictivo Suavizado Exponencial y ARIMA |
|---|
| **Lenguaje:** R |
| **Código:** |

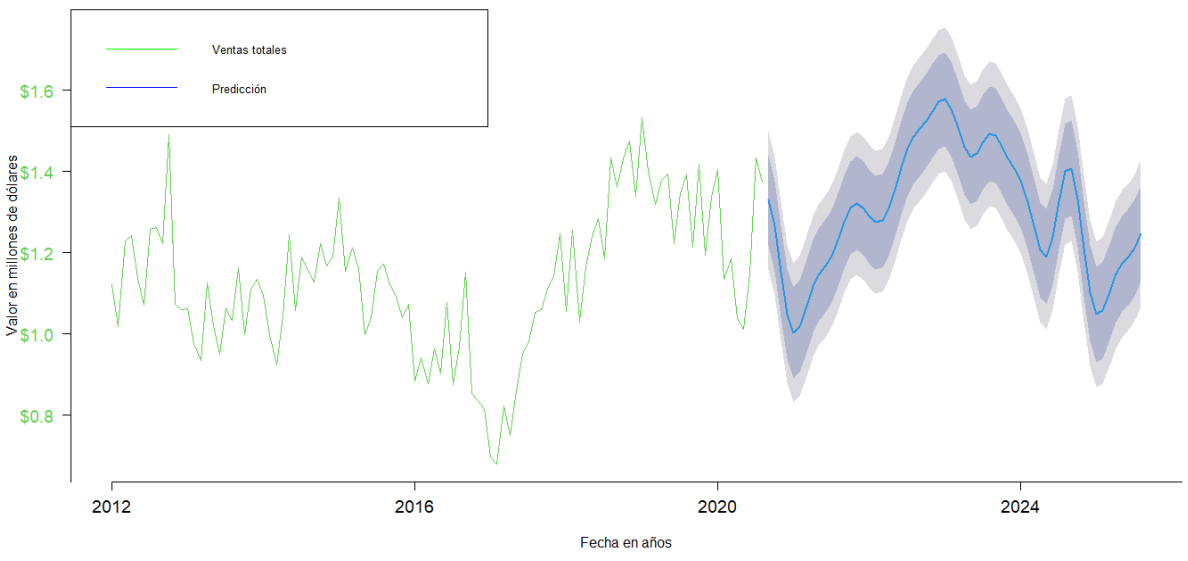
```
fit3 <- stlm(serie1, modelfunction=Arima, order=c(2,1,2))
fc3 <- forecast(fit3, 60)

xpos <- seq(600000, 2600000, by=400000)
plot(fc3, main="Predicción de ventas de los próximos 5 años -
STL y ARIMA(2,1,2)", axes = FALSE, xlab="Fecha en años",
ylab="Valor en millones de dólares", col=3)
axis(2, at=xpos, labels=sprintf("$%.1f", xpos/1000000),
col.axis=3, las=1, cex.axis=1.2)
axis(1,at=c(2012, 2013, 2014, 2015,
2016),labels=c("2012","2016","2020", "2024", "2028"),
cex.axis=1.2)
legend("topleft", legend=c("Ventas totales", "Predicción"),
       col=c("green", "blue"), lty=1:1, cex=.8)

accuracy(fc3)
```



## Anexo 8. Código del modelo TBATS en R

| **Función:** Modelo predictivo TBATS |
|---|
| **Lenguaje:** R |
| **Código:** |

```
fit6 <- tbats(serie1, biasadj=FALSE, use.arma.errors,
use.trend = TRUE)
fc6<- forecast(fit6, h=60)

xpos <- seq(600000, 2000000, by=200000)
plot(fc6, main="Predicción de ventas de los próximos 5 años -
TBATS", axes = FALSE, xlab="Fecha en años", ylab="Valor en
millones de dólares", col=3)
axis(2, at=xpos, labels=sprintf("$%.1f", xpos/1000000),
col.axis=3, las=1, cex.axis=1.2)
axis(1,at=c(2012, 2013, 2014, 2015,
2016),labels=c("2012","2016","2020", "2024", "2028"),
cex.axis=1.2)
legend("topleft", legend=c("Ventas totales", "Predicción"),
       col=c("green", "blue"), lty=1:1, cex=.8)

accuracy(fc6)
```



## Anexo 9. Código del modelo NNETAR en R

| **Función:** Modelo predictivo NNETAR |
|---|
| **Lenguaje:** R |
| **Código:** |

```
fit5 <- nnetar(serie1, maxit=100)
fc5<- forecast(fit5, h=60)

xpos <- seq(500000, 1800000, by=200000)
plot(fc5, main="Predicción de ventas de los próximos 5 años -
NNETAR", axes = FALSE, xlab="Fecha en años", ylab="Valor en
millones de dólares", col=3)
axis(2, at=xpos, labels=sprintf("$%.1f", xpos/1000000),
col.axis=3, las=1, cex.axis=1.2)
axis(1,at=c(2012, 2013, 2014, 2015,
2016),labels=c("2012","2016","2020", "2024", "2028"),
cex.axis=1.2)
legend("topleft", legend=c("Ventas totales", "Predicción"),
       col=c("green", "blue"), lty=1:1, cex=.8)

accuracy(fc5)
```