IMPLEMENTACION DE UN DASHBOARD DE SERIES TEMPORALES UTILIZANDO EL MODELADO DE ARIMA PARA PREDECIR FUTURAS GANANCIAS O PERDIDAS EN UN AÑO.

AUTORES:
Ing. Jimmy Fernando Castillo Crespín
jfcastilloc_est@utmachala.edu.ec
Ing. Esteban Fabricio Gonzabay Jimenez
egonzabay4@utmachala.edu.ec

Resumen

Una de las particularidades más interesantes y por ende importantes de la inteligencia de negocios (BI), es que el conocimiento de ciencias exactas tales como matemáticas, estadísticas, etc., se ven reflejadas al momento de modelar o diseñar una gráfica que muestre resultados para una problemática(ventas) dada. Este documento se enfoca en buscar la explicación paso a paso de un modelo ARIMA, para poder predecir el estado de las ventas de una empresa, mostrando así la importancia que tiene el utilizar este método. Los modelos ARIMA se basa en varios subprocesos como lo pueden ser estimación y verificación, ya que, uno de los principios generales de una empresa de compra y venta es saber si las ganancias son buenas o malas, dependiendo de la media que esta tendrá en los años anteriores, entonces ARIMA con esos datos trabaja y consigue aquella ventaja competitiva que puede lograr una empresa. El documento también cuenta con conceptos que enriquecen el conocimiento del lector, como herramientas BI, y definiciones de modelos estadísticos basándose en series temporales, funciones de auto relación y modelado ARIMA. Así mismo, artículos en donde se mostrarán otro tipo de análisis basándose en la misma temática de inteligencia de negocios.

Palabras claves: ARIMA, Series Temporales, Autocorrelación, CRISP-DM, Tablero de control.

I. INTRODUCCIÓN

En el presente trabajo se conceptualizará cada una de las herramientas que permitieron la creación de un análisis de inteligencia de negocios en una empresa comercial, poniendo en claro que la creación de un modelo BI es de mucha importancia no solo en la empresa que se tomara como ejemplo, sino en muchos otros casos de estudio.

En los últimos años el mercado del Bussiness Intelligence (BI) ha ido creciendo a medida que las empresas crecen en cantidad de datos almacenados. BI utiliza herramientas y técnicas ETL (Extraer, Transformar y Cargar) mismas que se explicara con más precisión a lo largo del documento.

Vivimos en una sociedad donde un análisis de negocios es muy importante basándose en que este, dará una explicación bastante detallada de la información más importante proveniente de un lago de datos al momento de presentar los resultados. Por lo general una empresa no cuenta con personal o clientes que sepan el estado de la información por lo tanto esto es una gran falla, la inteligencia de negocios puede brindar auxilio inmediato en el hecho de que los grandes lagos de datos (también llamados datos brutos) se conviertan en información válida y así ser comprendida por el personal para pasar a llamarse “conocimiento”. Esto es de gran ayuda para que los jefes, autoridades o gerentes que necesitan soluciones a diferentes problemáticas, sean estas a mediano o largo plazo.

Dicho esto, el documento pretende explicar puntualmente como una base de datos puede contestar varias preguntas que una empresa común y corriente puede realizarse, no tanto internamente sino para que esta información sea buena también para el cliente que invierte en la empresa.

El trabajo también contiene información interesante acerca de proyectos similares, en donde se aplican modelos de inteligencia de negocios, seguidamente una breve exposición de los detalles que resultaron al crear un diseño BI, y de igual manera se presentaran conclusiones y recomendaciones que puedan servir de apoyo para las personas que lean este documento.

II.TRABAJOS RELACIONADOS

Existen muchos estudiantes, profesores y profesionales que se dedican estrictamente en cambiar el paradigma de ajustes de datos en una organización, microempresa o macroempresa. Dando a entender que la creación de un diseño BI es de mucha importancia para la misma.

Se pretende resumir artículos científicos de casos en donde se utilizaron técnicas y herramientas para conseguir con eficiencia un perfecto diseño BI:

2.1. DASHBOARD PARA EL SOPORTE DE DECISIONES EN EMPRESAS DEL SECTOR MINERO

En la ciudad de Machala se pretendió realizar un análisis de inteligencia de negocios diseñando un sistema de soporte de decisiones (DSS) en un sector minero, mediante el uso de metodologías como lo son Hefesto y Kimball. (Mazon Olivo, Rivas Asanza, Gallegos Macas, & Pinta, 2017)

Hefesto. – Hefesto es una metodología que comprende varias etapas de investigación, entre ellas pueden ser:

(1) Análisis de requerimientos:

a) Identificar preguntas
b) Identificar indicadores y perspectivas
c) Modelo conceptual

(2) Análisis de los OLTP

a) Conformar indicadores
b) Establecer correspondencias
c) Nivel de granularidad
d) Modelo conceptual ampliado

(3) Modelo lógico del DW

a) Tipo de modelo lógico del DW
b) Tablas de dimensiones
c) Tablas de hechos
d) Uniones

(4) integración de datos

a) Carga inicial
b) Actualización

Esta metodología cuenta con las siguientes características:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos de los usuarios, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra a los usuarios finales en cada etapa para que tome decisiones respecto al comportamiento y funciones del DW

- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las estructuras físicas que contengan el DW y su respectiva distribución.
- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- Se aplica tanto para Data Warehouse como para Dara Mart. (Dario, 2010)

Kimball. - Kimball se refiere a un análisis dimensional, basándose en el ciclo de vida del proyecto o también llamado KLC (Kimball Life Cycle), desde el lanzamiento hasta el despliegue.

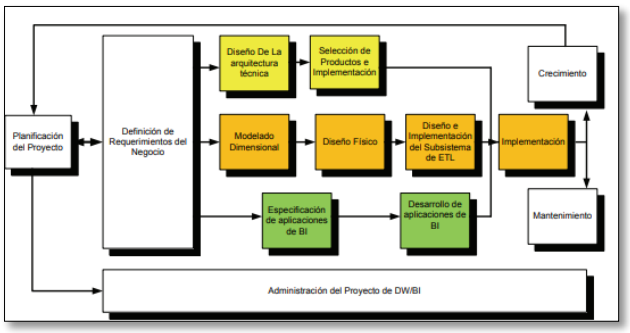


Fig. 1: Tareas de la metodología de Kimball, denominada Bussiness Dimensional Lifecycle (Rivadera, 2010)

Esto análisis incluye:

(1) Definir el alcance (entender los requerimientos del problema).
(2) Identificar las tareas.
(3) Programar las tareas.
(4) Planificar el uso de los recursos.
(5) Asignar la carga de trabajo a los recursos.
(6) Elaboración de un documento final que representa un plan del proyecto.

Además, cuentan con subfases tales como:

- Monitoreo del estado de los procesos y actividades
- Rastreo de problemas
- Desarrollo de un plan de comunicación comprensiva que direccione la empresa y las áreas de TI

2.2. SISTEMA PARA APOYAR LA TOMA DE DECISIONES EN LA DIRECCION GENERAL INFRAESTRUCTURA Y SERVICIOS.

Actualmente en las empresas se necesita tomar decisiones en respuesta a las exigencias de las actividades y funciones que sus trabajadores realizan. Grandes volúmenes de datos muchas veces provocan falta de concentración y homogeneización de estos, por lo que resulta difícil obtener una visión global del comportamiento del negocio. (Fernandez Henriquez, Prieto del Rio, & Rodriguez Freire, 2019)

El sistema que aplican los autores de este proyecto es la utilización de software y plataformas de código libre entre ellas son:

- **Visual Paradigm 6.4**, como herramienta de modelado soporta el ciclo de vida completo del desarrollo de software, es multiplataforma además permite realizar ingeniería inversa de bases de datos desde sistemas gestores de base de datos existentes a diagramas de Entidad-Relación.

  Es una herramienta CASE: ingeniería software asistida por computación. La misma propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación.
- **Suite Pentaho**, es una plataforma BI orientada a la solución y centrada en procesos que incluye los componentes requeridos para implementar soluciones basadas en procesos como minería de datos, ETL, generación de informes.
- **PostgreSQL**, es un sistema de código abierto de administración de base de datos del tipo relacional, aunque también es posible ejecutar consultas que sean no relaciones. En este sistema, las consultas relacionales se basan en SQL, mientras que las no relacionales hacen uso de JSON.

En conjunto con los programas anteriormente mencionados, este trabajo tiene como finalidad tener una visión de los subsistemas que citan en el trabajo, tales como: subsistema de integración, subsistema de almacenamiento y subsistema de visualización.

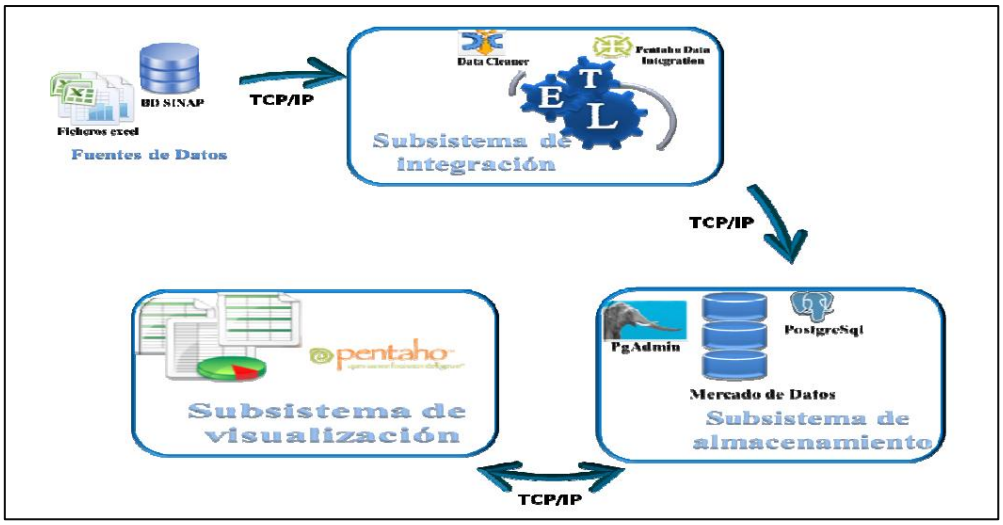


Fig. 2: Arquitectura de los subsistemas del mercado de datos (Fernandez Henriquez, Prieto del Rio, & Rodriguez Freire, 2019)

III.METODOLOGIA

Para la realización de este trabajo se opto por utilizar series temporales en la presentación de los datos estadísticos que muestra las data sets encontradas, esto nos ayudara a validar la información que podremos rescatar en cada uno de los diagramas.

3.1. SERIES TEMPORALES

Una serie temporal o de tiempo es una secuencia de observación, medidos en determinados momentos del tiempo ordenados cronológicamente y, espaciados entre si de manera uniforme, así los datos usualmente son dependientes entre sí. El principal objetivo de una serie de tiempo X_t, donde t = 1, 2, …, n es su análisis para hacer un pronóstico. (Villavicencio, 2011)

Las series temporales se pueden utilizar en varios campos de la ciencia, tales como:

- Marketing: Proyecciones del empleo y desempleo, evolución del índice de precios de la leche, beneficios netos mensuales de cierta entidad bancaria.
- Demografía: Numero de habitantes por año, tasa de mortalidad infantil por año.
- Medioambiente: Evolución horaria de niveles de oxido de azufre y de niveles de óxido de nitrógeno en una ciudad durante una serie de años, lluvia recogida diariamente en una localidad, temperatura media mensual, medición diaria del contenido en residuos tóxicos en un rio.

3.1.1. COMPONENTES DE LAS SERIES TEMPORALES

Componente tendencia. - Se puede definir como un cambio a largo plazo que se produce en la relación al nivel medio, o el cambio a largo plazo de la media. La tendencia se identifica con un movimiento suave de la serie a largo plazo.

Componente estacional. – Muchas series temporales presentan cierta periodicidad o, dicho de otro modo, variación de cierto periodo (semestral, mensual, etc.). Por ejemplo, las ventas al detalle en una ciudad donde las mismas aumentan por el pasar de los meses en un rango determinado de tiempo. Estos efectos son fáciles de entender y se pueden medir explícitamente o incluso se pueden eliminar de la serie de datos, a este proceso se le llama desestacionalización de la serie.

Componente aleatoria. - Esta componente no responde a ningún patrón de comportamiento, sino que es el resultado de factores fortuitos o aleatorios que inciden de forma aislada en una serie de tiempo.

Cabe destacar que estos tres componentes, los dos primeros son componentes determinísticos y el tercero es aleatorio.

3.1.2. AUTOCORRELACION

En ocasiones una serie de tiempo acontece, que los valores que toma una variable en el tiempo no son independientes entre sí, sino que un valor determinado depende de los valores anteriores, existen dos formas de medir esta dependencia de las variables: Función de autocorrelación y función de autocorrelación parcial.

3.1.2.1. Función de Autocorrelación Simple

La función de autocorrelación simple (FAS) es una herramienta de análisis estadístico que nos permite encontrar el nivel de autocorrelación de los datos y en que retardos, k, se produce. También se dice que es una función matemática que nos ayuda en saber que dependencia tienen los datos de un periodo determinado con los mismos de hace k periodos anteriores.

La importancia de las FAS radica más en su representación que en su fórmula matemática dado que son los resultados los que representamos y a partir de los cuales sacaremos nuestras conclusiones.

La utilización de la FAS consiste en medir la inercia o tendencia de una serie temporal, es decir, ver qué grado de dependencia muestran los datos del ahora con los datos de hace k periodos anteriores.

Dado que la metodología de trabajo son las series temporales, establecemos el análisis sobre una única variable en distintos momentos del tiempo. Un ejemplo típico seria el

precio de cotización de un activo financiero entre 1990 y 2020. Aunque los precios cambien, la variable de estudio será la misma: precio de cotización.

¿Cómo se representa una FAS?

Según la tipología de los datos, ira cambiando dado que no todos los datos son iguales, ni tienen el mismo nivel de correlación con el pasado. (Rodó, 2020)

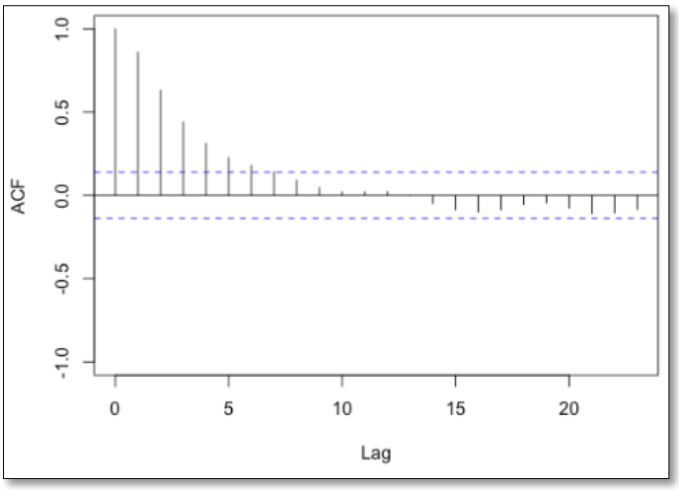


Fig. 3: Ejemplo de una FAS con datos reservados

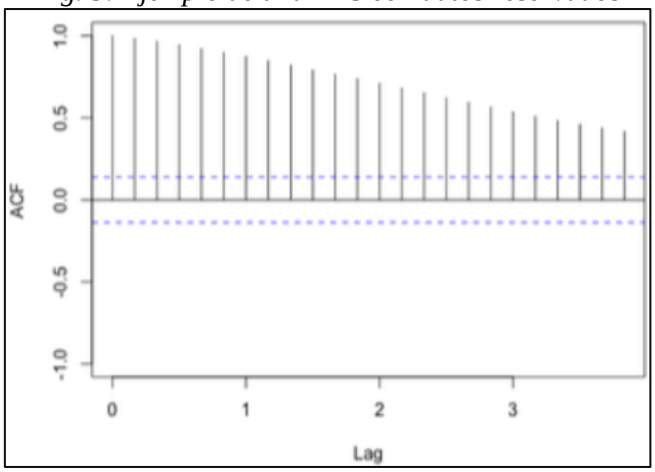


Fig. 4: Ejemplo de una FAS con datos reservados

Se denomina LAG, al retardo de las líneas discontinuas, las mismas que representan bandas de desconfianza al 95% de defecto.

3.1.2.2. Función de autocorrelación parcial (PACF)

La función de autocorrelación parcial es una medida de la correlación entre observaciones de una serie de tiempo que se encuentran separadas por k unidades de tiempo ($X_t - X_{t-k}$), después de ajustarse para la presencia de los demás términos de desfase más corto ($X_{t-1}, X_2, \dots, X_{t-k-1}$).

3.1.3. CONTRASTE DICKEY FULLER

El contraste de Dickey Fuller permite tener un análisis mas abierto a la información que tenemos presente, para de tal manera clasificarlas en "tendencias estocásticas" o "no tendencia estocástica" dentro de las series temporales.

Matemáticamente hablando se presentan los siguientes puntos, teniendo en cuenta un modelo autorregresivo (AR):

Hipotesis nula $\boldsymbol{H0}$

Hipotesis alternativa $\boldsymbol{H1}$

- $\boldsymbol{H1}$: $\varphi = \mathbf{1} = \boldsymbol{Tendencia\ estocastica\ en\ las\ series\ temporales}$
- $\boldsymbol{H0}$: $\varphi < \mathbf{1} = \boldsymbol{No\ tendencia\ estocastica\ en\ las\ series\ temporales}$

El contraste Dickey Fuller es comúnmente aplicado en econometría para comprobar la presencia de tendencia sobre las series temporales. La particularidad del contraste Dickey Fuller es que es la herramienta más fácil de usar comparado con otros contrastes más complejos que también prueban la presencia de tendencia en los datos.

3.1.4. MODELO ARIMA

El modelo arima o también llamado de predicción, se utiliza para la realización de análisis predictivos en un modelo a corto plazo.

El modelo ARIMA se puede clasificar en 4 etapas:

1. Identificación. Utilizando los datos y/o cualquier tipo de información disponible sobre como ha sido generada la serie, se intentará sugerir una subclase de modelos ARIMA (p, d, q) que merezca la pena ser investigada. El objetivo es determinar los ordenes p, d, q que parecen apropiados para reproducir las características de la serie bajo estudio y si se incluye o no la constante φ, En esta etapa es posible identificar mas de un modelo candidato a haber podido generar la serie.
2. Estimación. Usando de forma eficiente los datos se realiza inferencia sobre los parámetros condicionada a que el modelo investigado sea apropiado.

   Dado un determinado proceso propuesto, se trata de cuantificar los parámetros de este.
3. Validación. Se realizan contrastes de diagnostico para comprobar si el modelo se ajusta a los datos, o, si no es así, revelar las posibles discrepancias de modelo propuesto para poder mejorarlo.

4. Predicción. Obtener pronósticos en términos probabilísticos de los valores futuros de la variable. En esta etapa se tratará también de evaluar la capacidad predictiva del modelo.

Esta metodología se basa fundamentalmente, en dos principios:

- Selección de un modelo en forma interactiva. En cada etapa se plantea la posibilidad de rehacer las etapas previas.
- Principio de parametrización escueta, también denominado parsimonia. Se trata de proponer un modelo capaz de representar la serie con el mínimo de parámetros posibles y únicamente acudir a una ampliación del mismo en caso de que sea estrictamente necesario para describir el comportamiento de la serie.

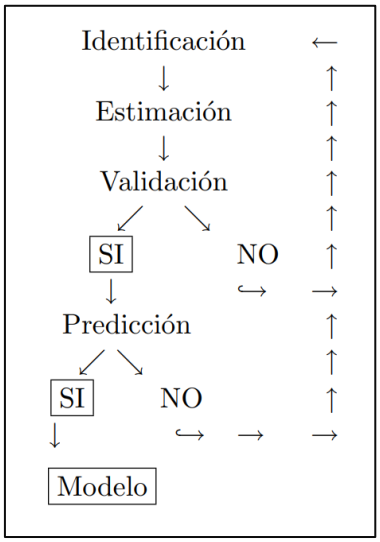


Fig. 5: Etapas del modelo ARIMA

IV.Resultados

Los resultados que se presentan a continuación en esta investigación fueron obtenidos siguiendo la metodología CRIPS-DM, en cual indica varias fases a seguir:

1. **Primera fase. -** la primera fase comprende al entendimiento del negocio, “Pagar es Fácil” es un eje de negocios digitales que permite a pequeños y medianos empresarios exhibir y vender productos o servicios, así como transaccionar con tarjetas de crédito y billetera virtual, transferir saldos a cuentas bancarias y pagar servicios básicos. Pagar es Fácil ofrece múltiples servicios, pero la que se utilizó en esta investigación fue el servicio de Marketplace. Pagar es Fácil aloja muchos comercios los cuales día a día suben sus productos a la plataforma e igualmente se transacciona con las compras realizadas por las personas.

Regularmente la plataforma recibe ingresos económicos por comisiones de cada venta, siempre y cuando el vendedor entregue correctamente el producto. En base a lo mencionado anteriormente, Pagar es Fácil requiere conocer una predicción de ganancias de las ventas que se podrían generar en futuros años, conocer cuáles son sus compradores más recurrentes, las categorías y productos más solicitados, y sus vendedores más confiables.

2. **Segunda fase**, el entendimiento de los datos, EL data-set fue obtenido de las diferentes bases de datos proporcionadas por la plataforma, las cuales son las ventas conformadas por los clientes, vendedores, productos y categorías, a continuación, se detalla cada una ellas:

- **Cliente:** Nombres y apellidos de los clientes.
- **Nombre_producto:** El nombre del producto.
- **Name_comercio:** El nombre del vendedor o mejor conocido como el comercio.
- **Precio_producto:** Precio del producto proporcionado por el vendedor.
- **Precio_producto_venta:** Precio del producto sumado con la comisión de Pagar es Fácil.
- **Ganancias:** Es la resta de precio_producto_venta con precio_producto obteniendo así las ganancias por producto por venta.
- **Category:** Las categorías de los productos.
- **Fecha:** Fecha de venta.
- **Venta_concretada:** Variable utilizada para conocer si un vendedor entrego el producto al cliente.

3. **Tercera fase**, los datos pasaron por un preprocesamiento, las cuales se indican a continuación, lo primero que se hizo con el data-set fue corregir aquellos campos con caracteres que están mal codificados.

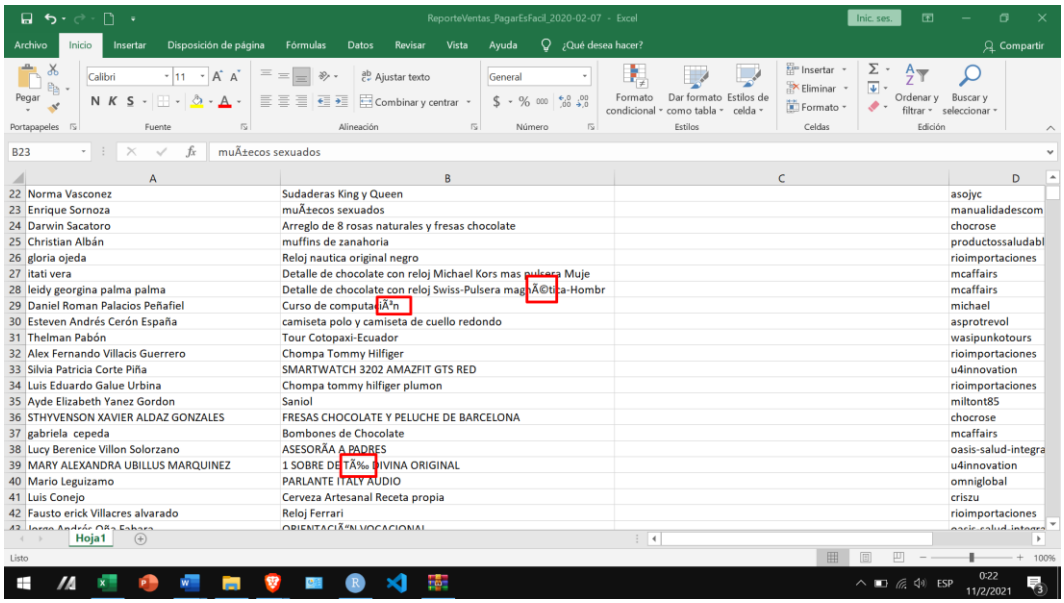


Fig. 6: Estado inicial de la data en Excel

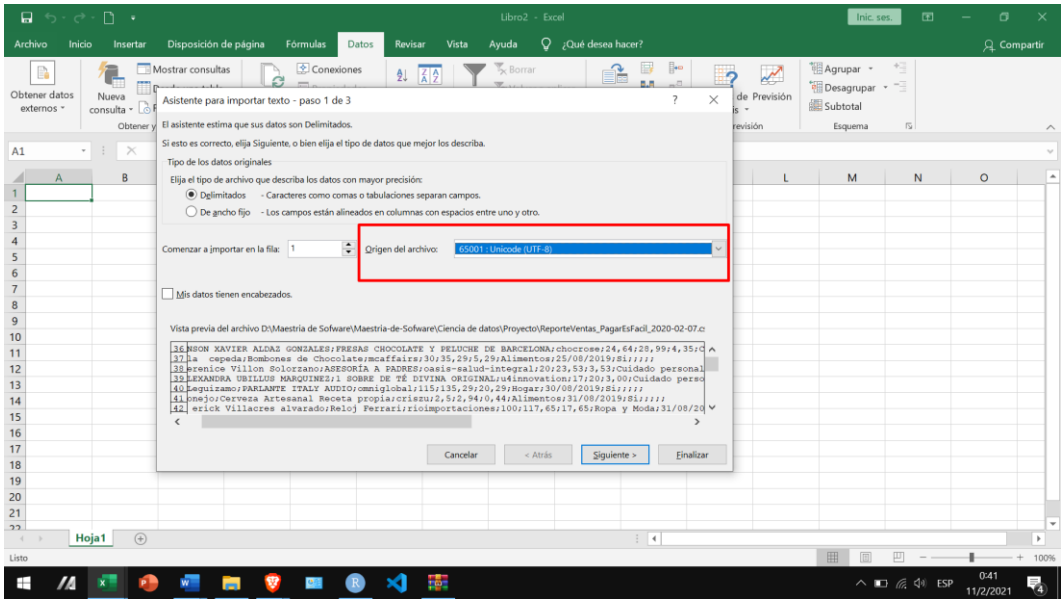


Fig. 7: Proceso de corrección de datos en Excel

Se cambiaron las comas que poseen los campos numéricos de precio_producto, precio_producto_venta y ganancias por el punto, para que posteriormente la herramienta R nos los reconozca como variable numérica y no una cadena de texto.

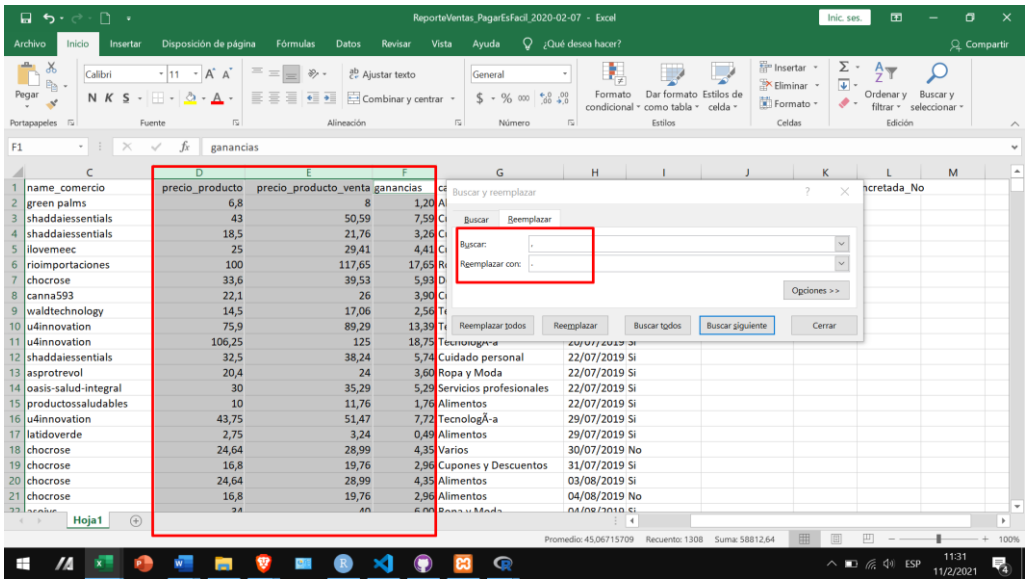


Fig. 8: Cambio del punto decimal al punto en valores numéricos en Excel

Se borrarón aquellos campos adicionales que se crearon al transformar el archivo .xlms a .csv, para esto se utilizó la librería Rattle.

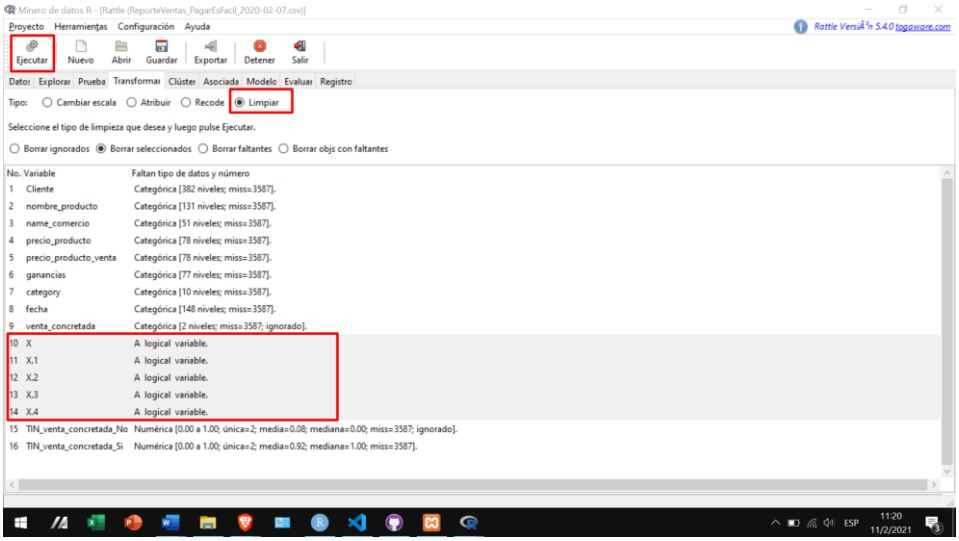


Fig. 9: Eliminación de campos adicionales de la data con Rattle.

Se utilizó la opción RECODE de la librería Rattle, para crear variables indicadoras a la columna venta_concretada que poseen registros con valores de SI/NO y posteriormente utilizarlas para un posible análisis.

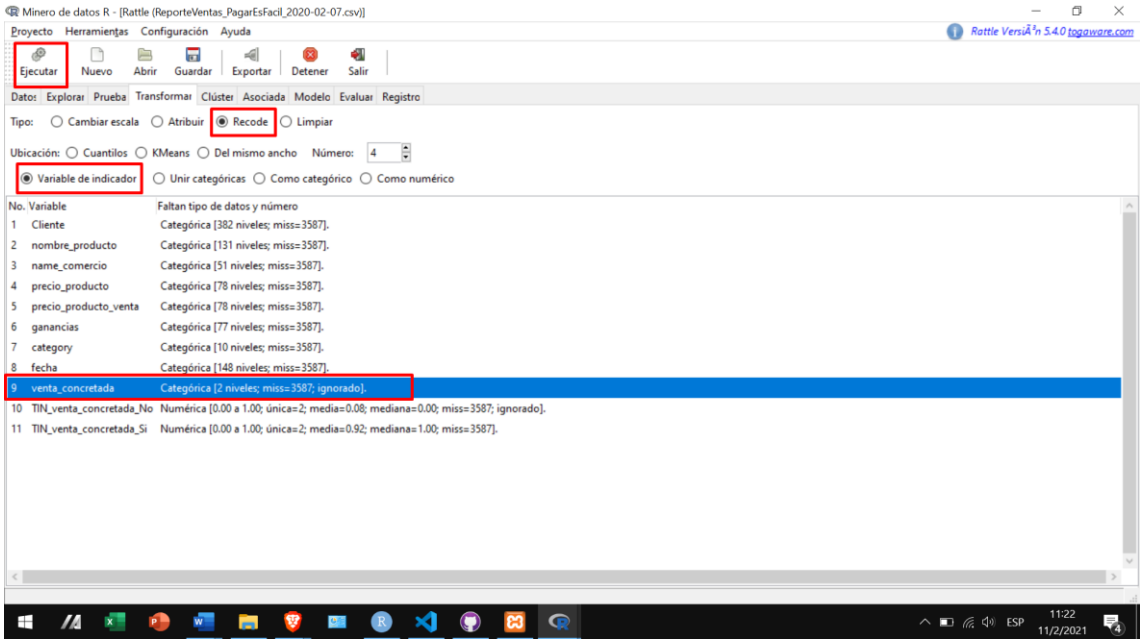


Fig. 10: Creación de variable indicadora venta_concretada con Rattle.

A lo que respecta con la fase de modelamiento, se utilizó la nube de palabras para determinar cuáles son los productos más comprados en la tienda y también se aplicó el modelo ARIMA en las series temporales para la predicción de ganancias de las futuras ventas en Pagar es Fácil.

4. Cuarta Fase

4.1. Nubes de palabras

En la siguiente imagen se detalla el resultado de esta técnica de minería de datos, en la cual se analizaron todos los productos vendidos desde el junio del 2019 hasta febrero del 2021, con una cantidad de frecuencia de por ejemplo 7 y un total de 89 palabras, dando como resultado que las palabras chocolate, aceite, fresas y computador fueron las más buscadas y por ende vendidas por la tienda.

Fig. 11: Empleo de la nube de palabras en R.

El código utilizado para la creación de la nube de palabras fue el siguiente:

```
productos <- VCorpus(VectorSource(datosc$nombre_producto)) #convierte en arreglo el texfo

productos <- tm_map(productos, content_transformer(tolower)) #convierte a miniculas
productos <- tm_map(productos, removePunctuation) #quita los signos de puntuacion

colores <- brewer.pal(8, "Dark2") #para dar colores a las palabras

output$wordClouds <- renderPlot({

  wordcloud(productos, scale=c(4,0.5),min.freq = input$freq,
            max.words=input$max, random.order=FALSE, rot.per=0.35, colors=colores)
```

Fig. 11.1: Código fuente de la creación de nubes de palabras en R

4.2. Series temporales

El primer paso para trabajar con las series temporales es convertir los datos en una serie temporal, se utilizó el siguiente código en donde se detalla el mes y año de inicio de la primera venta y el mes y año final en la que se quiere hacer

```
 # convertimos la data a series temporales
ventas<-ts(datosc$ventas_mensuales, start = c(2019,6), end=c(2025, 12), frequency = 12)
```

Fig. 12: Conversión de la data en serie temporales

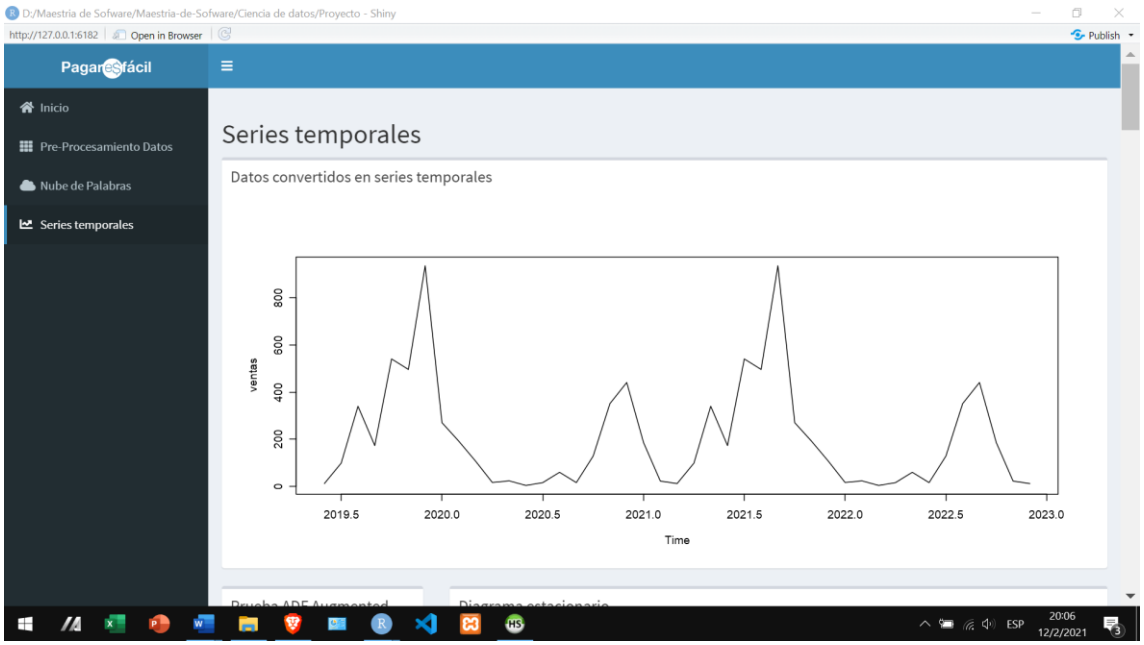


Fig. 12.1: Conversión de la data en serie temporales

Para que se pueda aplicar el modelo Arima, un requisito es que la serie temporal debe ser estacionaria, una serie es estacionaria cuando los valores están cercas de una misma media, entonces para comprobar si la serie temporal es estacionaria se aplicó el test de Dickey Fuller el cual indica que la del p-value debe ser menor a 0.05, dando como resultado 0.03 lo cual indica que nuestra serie es estacionaria.

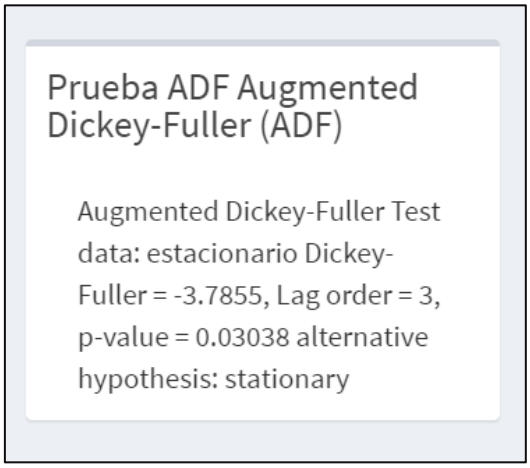


Fig. 12.2: Presentación del Dickey Fuller

El código utilizado fue el siguiente:

```
#CONVERTIR LA SERIE TEMPORAL EN ESTACIONARIA
# Prueba ADF Augmented Dickey-Fuller (ADF) t-test

#determina el numero de diferencias necesarias para la serie de tiempo, esto para hacerla estacionaria
ndiffs(ventas) # da resultado 1

estacionario= diff(ventas, differences = 1)
# estacionario <- ventas

respuesta <- adf.test(estacionario, alternative = "stationary")
```

Fig. 12.3: Codigo del Dickey fuller

Luego se obtuvieron los valores de las funciones de auto correlación (ACF) y la función de auto correlación parcial (PACF), estas dos funciones nos sirven para saber cuántas medias móviles y cuantos autoregresivos vamos a utilizar en nuestro modelo ARIMA entonces para eso se utiliza el siguiente comando:

```
#La función de auto correlación y autocorrelación parcial
#sirven para conocer cuantos medias moviles y cuantos auto regresivos utilizaremos en nuestra modelo ARIMA

# autocovarianza
acf <- autoplot(acf(ts(estacionario)))

# autcorrelacion parcial de la muestra
plotPACF <- autoplot(pacf(ts(estacionario)))
```

Fig. 13: Codigo para la función autocorrelacion

Analizando los resultados:

El ACF nos indica el número de medias móviles y el PACF nos indica el número de autorregresivos y en ambos nos aparecen los rezagos de la serie, del cual para el ACF el rezago es 3, y en PACF es 6, estos dos valores nos servirán posteriormente para la aplicación de ARIMA.

Fig. 13.1: Presentación de las funciones de autocorrelación

Hasta el momento tenemos de resultado:

ACF= 3

PACF= 6

Diferencias=1

Se procedió a descomponer la serie para analizar más a fondo la tendencia, estacionalidad, componente irregular y los datos observados.

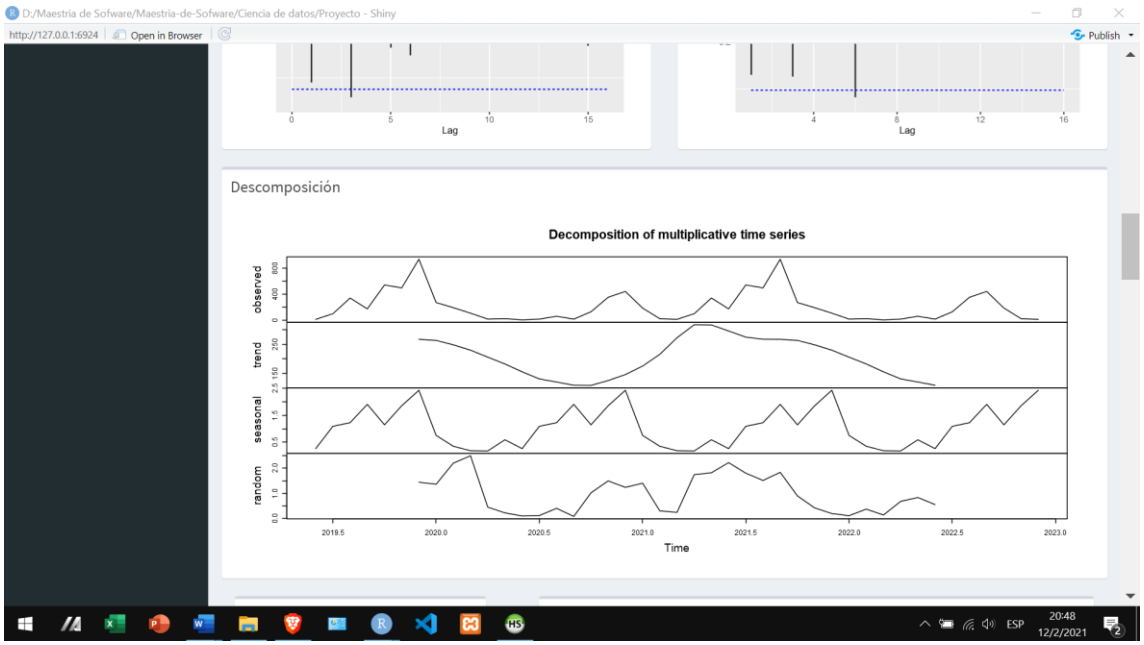


Fig. 14: Grafico de la descomposición de las series temporales

Para aplicar el modelo ARIMA, se utilizó los valores previamente obtenidos, los cuales son ACF, diferencias y PACF

```
modelo1 <- Arima(ventas, order=c(3,1,6)) #el primer parametro es el ACF, el 2do es el nro_diferencias y el 3er parametro es el PA
```

Fig. 15: Valores que se usaron para aplicar ARIMA en el código de las series temporales

Obteniendo el siguiente resultado, en el cual nos indica los coeficientes de autorregresivos, los coeficientes de la media móvil y los errores estándar.

Resultados del modelo ARIMA

Series: ventas ARIMA(3,1,6) Coefficients:
ar1 ar2 ar3 ma1 ma2 ma3 ma4 ma5
ma6 -0.0678 0.0314 -0.7662 -0.3405
-0.0217 0.3683 -0.3484 -0.0278 -0.6299
s.e. 0.1772 0.1923 0.1391 0.1806 0.2367
0.1991 0.1606 0.2041 0.1747 sigma^2
estimated as 36671: log
likelihood=-278.22 AIC=576.44
AICc=583.54 BIC=593.82

Fig. 16: Coeficientes del modelo ARIMA

Se puede realizar un diagnóstico para saber si el modelo es bueno, se utiliza el siguiente comando.

```
tsdiag(modelo1)
```

Fig. 16.1: Codigo para revisar el diagnostico del modelo

La siguiente gráfica muestra un diagnóstico del modelo que utilizamos, en la primera gráfico se muestran los errores estandarizados y estos deben parecerse mucho al ruido blanco, pero para eso se hará uso de los valores p (p-values) que entrega la tercera gráfica, el estadístico de Ljung Box, los cuales nos indica si hay o no ruidos blanco, y como se aprecia en la gráfica #3 los valores son mayores a 0.5 lo que indica según el test de Ljung que si es mayor a 0.5 entonces hay ruido blanco y nuestro modelo se ajusta bien.

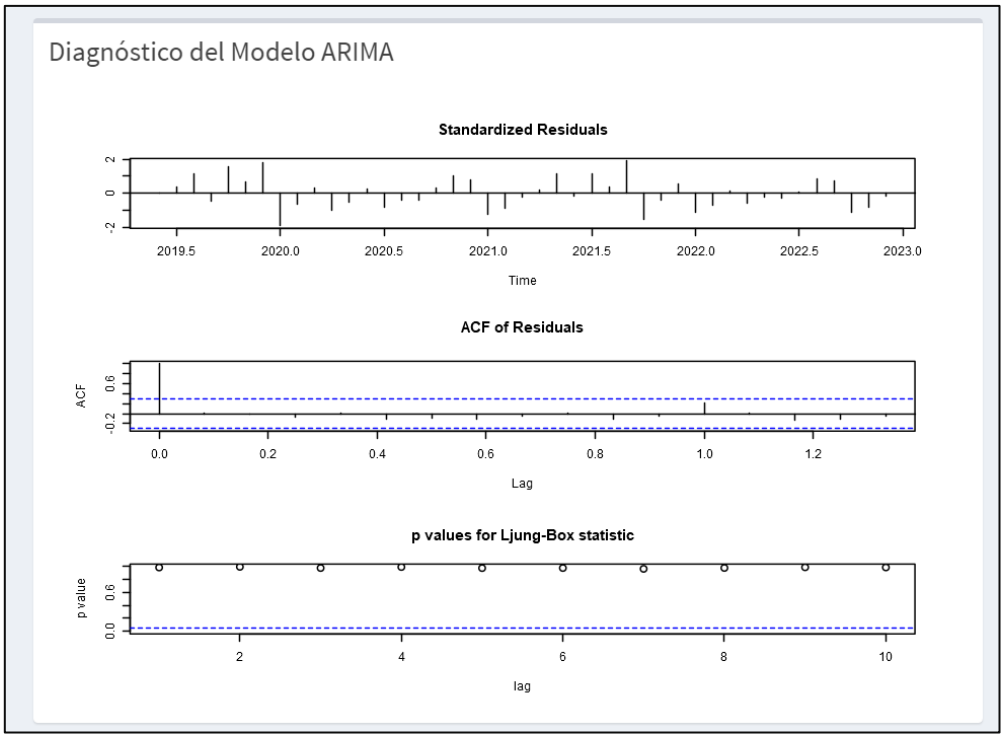


Fig. 16.2: gráficos para revisar los valores del modelo ARIMA

Para confirmar lo anteriormente dicho, se ejecutó el test de Ljung dando como resultado un p-value de 0.98 siendo este mayor a 0.5.

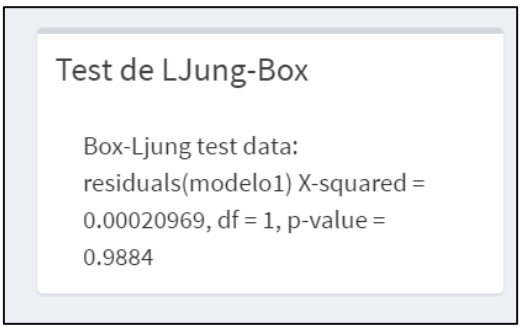


Fig. 17: Test de Ljung para revisar los valores de la fig. 16.2

5. Quinta Fase

5.1. Evaluación

Para dar una predicción o pronóstico de la cantidad de ganancias de ventas dentro de los próximos 12 meses de Pagar es Fácil hay que analizar los limites inferiores y superiores con el 80% y 95% de confianza.

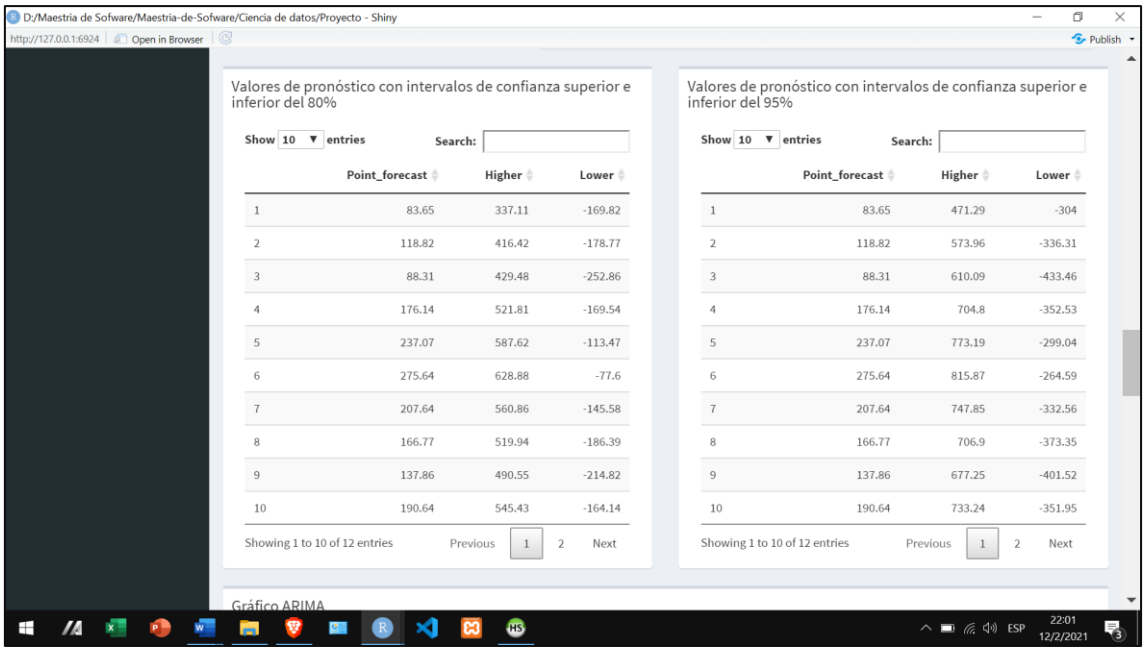


Fig. 18: Tabla de resultados con posibles pronósticos según ARIMA

La interpretación del resultado indica, por ejemplo, para el mes de enero del 2023, Pagar es Fácil tendrá un ingreso de $83.65 tanto a nivel de 80% y 95% de confianza, siendo su valor más alto $337.11 y su valor más bajo $-169.82 para el nivel de confianza de 80% y para el 95% su valor más alto es de $471.29 y su nivel más bajo $-304.

La siguiente gráfica muestra el resultado de la tendencia de las ventas a futuro.

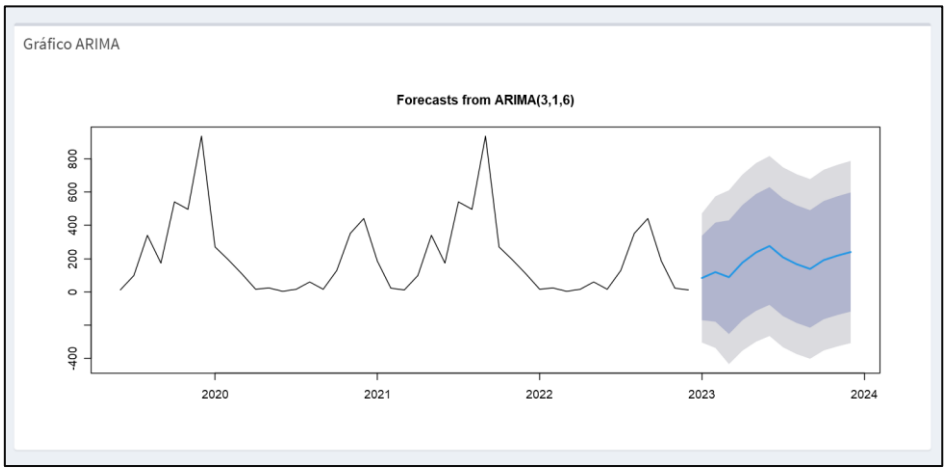
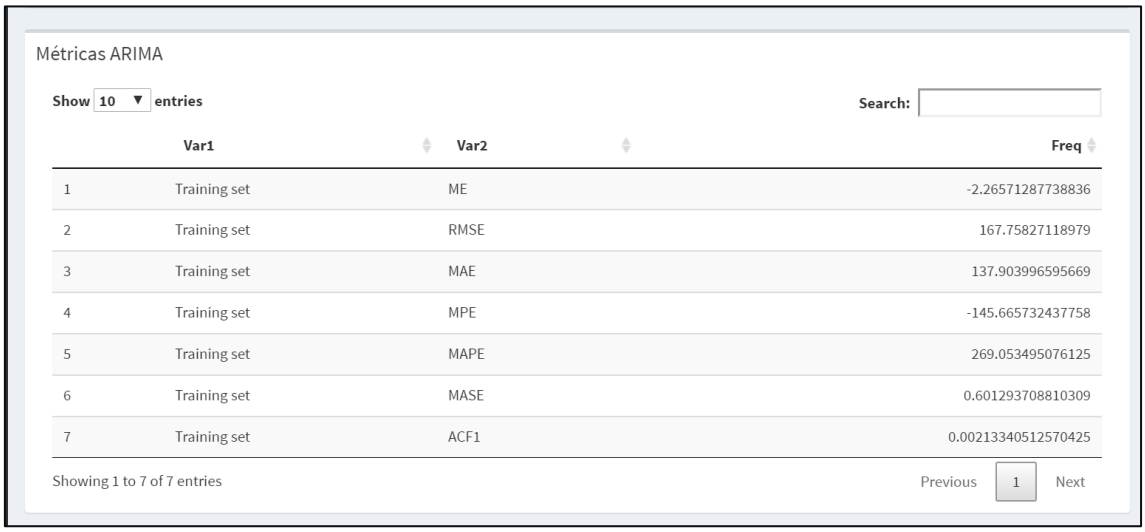


Fig. 19: Grafica con tendencia a futuro

La siguiente tabla muestra las cantidades de indicadores que miden errores de pronóstico.



Métricas ARIMA

Show 10 entries Search:

	Var1	Var2	Freq
1	Training set	ME	-2.26571287738836
2	Training set	RMSE	167.75827118979
3	Training set	MAE	137.903996595669
4	Training set	MPE	-145.665732437758
5	Training set	MAPE	269.053495076125
6	Training set	MASE	0.601293708810309
7	Training set	ACF1	0.00213340512570425

Showing 1 to 7 of 7 entries Previous 1 Next

Fig. 20: Indicadores de errores de pronostico

Como paso final de la metodología CRISP-DM, se muestra a continuación el despliegue del dashboard en producción.

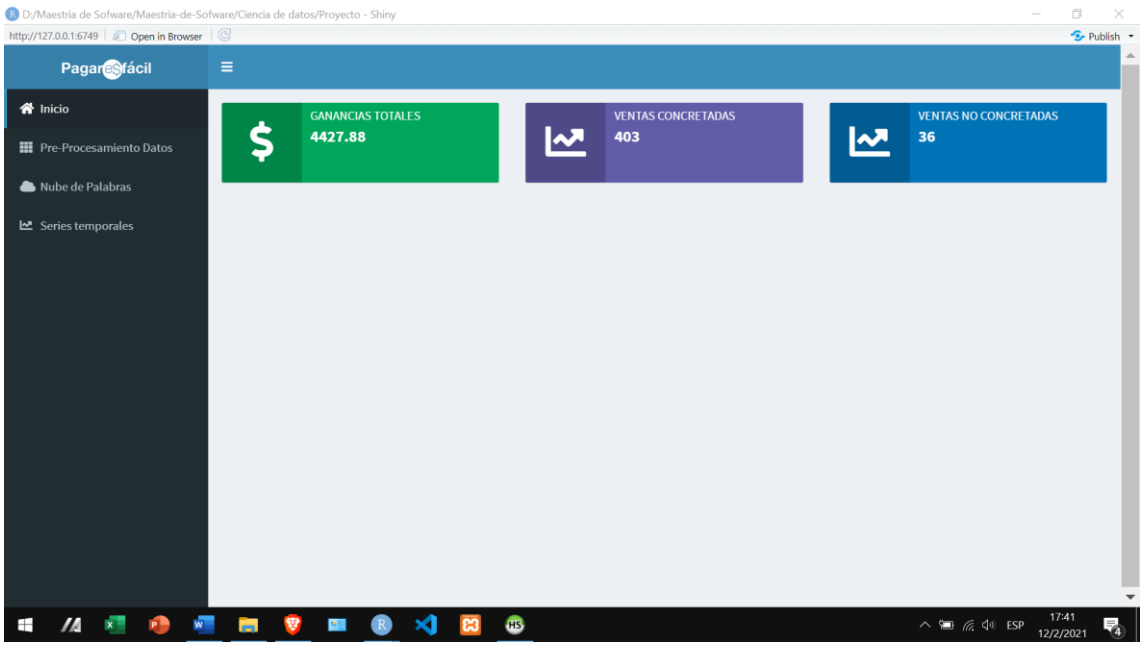


Fig. 21: Menú Inicio del Dashboard

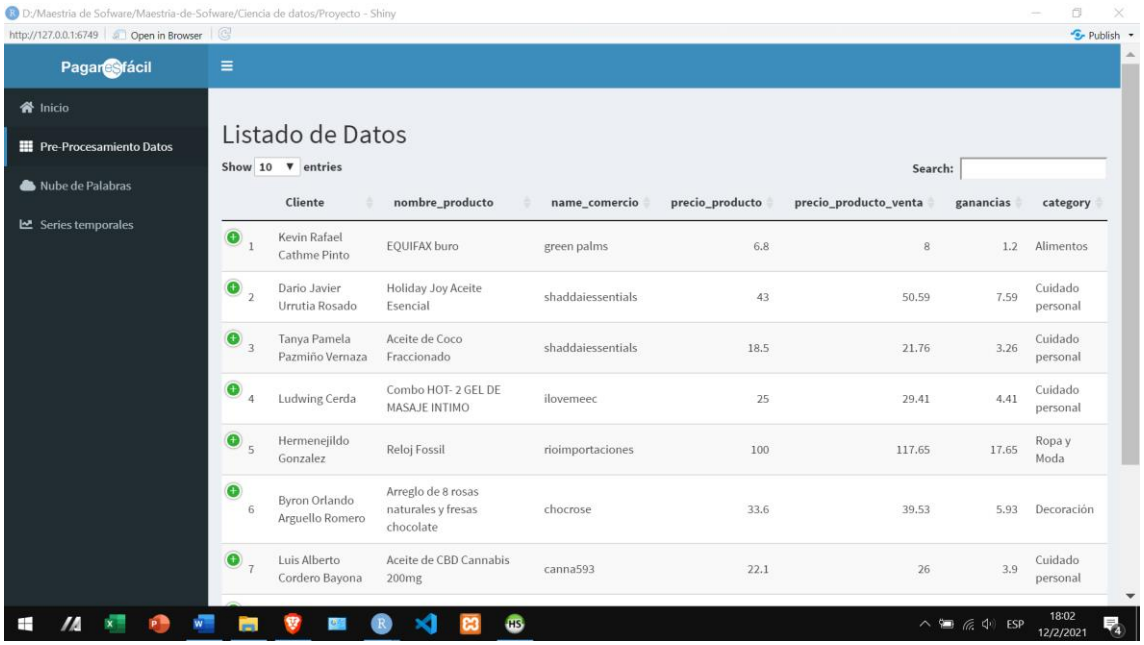


Fig. 22: Menú de preprocesamiento de datos.

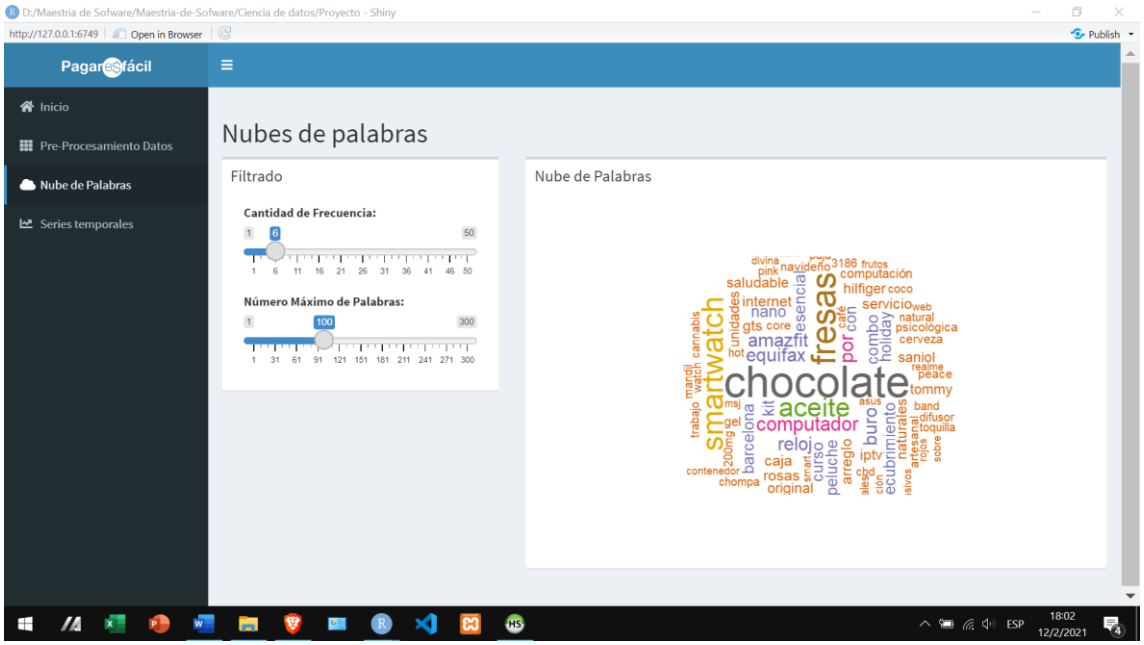


Fig. 23: Menú de nubes de palabras

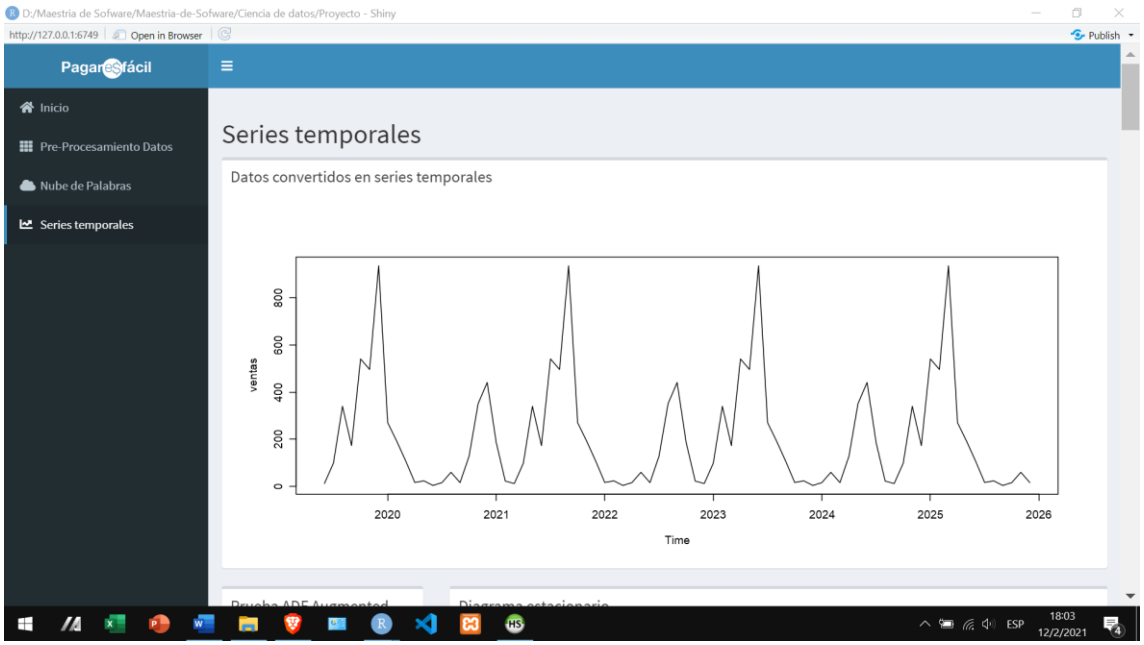


Fig. 24: Menú de series temporales

6. Conclusiones

Esta investigación se realizó utilizando la metodología CRISP DM la cual nos permitió lograr cumplir con los objetivos planteados. En cuestión de la simulación y pronóstico de las ganancias futuras para el Marketplace de Pagar es Fácil se realizó utilizando las series de tiempo de las cuales se han evaluado los modelos más comunes en este ámbito. Mediante el análisis de los resultados se puede concluir que, para realizar un pronóstico aproximado, se debe elegir el modelo más acertado para el evento que se quiere estudiar. El análisis estadístico permitió tomar una decisión del modelo escogido, el cual cumple con los parámetros requeridos de normalidad, varianza constante y aleatoriedad. En relación con el caso de estudio presentado, el pronóstico de ganancias mensuales más altos con un intervalo de confianza del 80% entre los 12 meses del próximo año oscila entre $337.11 a $628.88 y su valor más bajo oscila entre $ -77.6 a $-252.86, ahora con respecto al intervalo de confianza del 95%, el valor más alto oscila entre $471.29 a $815.87 y el valor más bajo oscila entre $ -264.59 a $ -433.46. La simulación a partir del modelo ARIMA (3,1,6) demuestra que los pronósticos, desde un registro histórico, se adapta muy bien a los niveles máximos y mínimos, y se mantienen en el rango. Se concluye, entonces, que estos modelos no permiten simular el comportamiento exacto en el tiempo, pero es una buena herramienta con la cual se obtiene una aproximación de posibles eventos máximos y mínimos. Por último, mediante los modelos de predicción estadística y simulación es posible tener aproximaciones de los comportamientos de las ganancias para periodos cortos de tiempo, y que estas metodologías constituyen una herramienta que, una vez optimizada, permitirá obtener una aproximación de las ganancias de ventas.

REFERENCIAS BIBLIOGRAFICAS

Dario, I. B. (2010). *Hefesto (DATA WAREHOUSING: Investigacion y sistematizacion de conceptos).* Cordoba, Argentina: Free Software Foundation.

Fernandez Henriquez, S., Prieto del Rio, D. R., & Rodriguez Freire, M. (2019). *Sistema para apoyar la toma de decisiones en la Direccion General Infraestructura y Servicios.* Cancun, Mexico: Innovation in Engineering, Technology and Education for Competitiveness and Prosperity.

Mazon Olivo, B., Rivas Asanza, W., Gallegos Macas, H., & Pinta, M. (2017). *Dashboard para el soporte de decisiones en empresas del sector minero.* Machala, El Oro: ResearchGate.

Rivadera, G. R. (2010). *Rivadera: La Metodologia de Kimballl para el diseño de almacenes.* Cancun, Mexico: Cuaderno de la facultad n. 5.

Rodó, P. (2020). *Funcion de autorrelacion simple.* Barcelona, España: economipedia.com.

Villavicencio, J. (2011). *Introduccion a series de tiempo.* Puerto Rico: Introduccion a series de tiempo.