UNIVERSIDAD DE SONORA
DIVISIÓN DE INGENIERÍA

POSGRADO EN INGENIERÍA INDUSTRIAL
MAESTRÍA EN INGENIERÍA EN SISTEMAS Y TECNOLOGÍA

ANÁLISIS DEL COMPORTAMIENTO DEL PRECIO DE LA UVA DE MESA SONORENSE EN ESTADOS UNIDOS MEDIANTE ESTRATEGIAS DE MINERÍA DE DATOS

T E S I S

PRESENTADA POR

VICETE SOLIS SANDOVAL

Desarrollada para cumplir con uno de los
requerimientos parciales para obtener
el grado de Maestro en Ingeniería

DIRECTOR DE TESIS
DR. JUAN MARTÍN PRECIADO RODRÍGUEZ

CODIRECTOR
DR. LUIS FELIPE ROMERO DESSENS

HERMOSILLO, SONORA, MÉXICO. MARZO 2022

Universidad de Sonora

Repositorio Institucional UNISON

UNIVERSIDAD DE SONORA

División de Ingeniería
Posgrado en Ingeniería Industrial
Maestría en Ingeniería en Sistemas y Tecnología

Hermosillo, Sonora a 14 de diciembre de 2021.

VICENTE SOLIS SANDOVAL

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado vigente, otorgamos a usted nuestra aprobación de la fase escrita del examen de grado, como requisito parcial para la obtención del Grado de Maestro(a) en Ingeniería: Ingeniería en Sistemas y Tecnología.

Por tal motivo este jurado extiende su autorización para que se proceda a la impresión final del documento de tesis: **ANÁLISIS DEL COMPORTAMIENTO DEL PRECIO DE LA UVA DE MESA SONORENSE EN ESTADOS UNIDOS MEDIANTE ESTRATEGIAS DE MINERÍA DE DATOS** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

DR. JUAN MARTIN PRECIADO RODRIGUEZ
Director(a) de tesis y Presidente del jurado

DR. FEDERICO MIGUEL CIRETT GALAN
Secretario(a) del Jurado

M.C. CARLOS ANAYA EREDIAS
Vocal del Jurado

DR. LUIS FELIPE ROMERO DESSENS
Vocal del Jurado

Pereira, Colombia, a 16 de diciembre de 2021.

VICENTE SOLIS SANDOVAL

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado de la Universidad de Sonora, otorgo a usted mi aprobación de la fase escrita del examen profesional, como requisito parcial para la obtención del Grado de Maestro en Ingeniería: Ingeniería en Sistemas y Tecnología.

Por tal motivo, como sinodal externo y vocal del jurado, extiendo mi autorización para que se proceda a la impresión final del documento de tesis: **ANÁLISIS DEL COMPORTAMIENTO DEL PRECIO DE LA UVA DE MESA SONORENSE EN ESTADOS UNIDOS MEDIANTE ESTRATEGIAS DE MINERÍA DE DATOS** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

M.C. JUAN LUIS ARIAS VARGAS
UNIVERSIDAD CATÓLICA DE PEREIRA
Sinodal Externo y Vocal del Jurado

RESUMEN

La Cadena Productiva de la Uva de Mesa Sonorense (CPUMS) es una de las cadenas productivas agroindustriales más importantes para el estado de Sonora, ya que cuenta con un volumen de producción anual de 320,990.10 toneladas en el año 2020 (Cierre de producción agrícola 2020 a nivel estatal - Sonora, 2020), representando el 84.47% a nivel nacional teniendo un valor de producción de $9,296,739.69 (miles de pesos) (Anuario Estadístico de la Producción Agrícola, 2020) en donde se estima que el 80% de la producción es exportada a Estados Unidos, generando divisas por 731,857 miles de dólares, colocándose como el producto agrícola con mayor valor de exportación para el estado (Estimación del volumen y valor de los principales productos agrícolas exportados 2020 - Sonora, 2020).

El mercado de Estados Unidos presenta diversas ventajas para la CPUMS como lo son los precios de venta altos, en una ventana comercialización de entre 15 a 20 días, cuando la uva sonorense se encuentra como único producto de este tipo, ya que se coincide con la salida de Chile y la entrada de California, lo que representa una ventaja competitiva para la CPUMS; sin embargo, este mercado se encuentra caracterizado por un incremento en estándares de calidad, inocuidad y responsabilidad social, lo que provoca un aumento en la estructura de costos para los productores, sumado a precios relativamente estables que provocan reducción de los márgenes de utilidad.

Dado que es un sistema abierto de producción, depende de diversos factores como el clima, desastres naturales, economía, política y oportunidad, haciendo que el periodo de comercialización se vuelve incierto que, sumado a lo mencionado anteriormente, provoca que la sostenibilidad del sistema se encuentra en riesgo.

Es por eso por lo que en este trabajo se propone una metodología basada en técnicas de minería de datos que permita identificar la dinámica de ventana de mercado en cuanto a la fecha de inicio y duración de dicho periodo, así como caracterizar el comportamiento de los precios a través de modelos de series de tiempo, con la finalidad de apoyar a la toma de decisiones y planeación de la producción de los agricultores sonorenses y de México.

ABSTRACT

The Sonora Table Grape Productive Chain (CPUMS) is one of the most important agro-industrial productive chains for the state, since it has an annual production volume of 320,990.10 tons in 2020 (Cierre de producción agricola 2020 a nivel estatal - Sonora, 2020), representing 84.47% a national level representing a production value of $ 9,296,739.69 (thousands of pesos) (Anuario Estadístico de la Producción Agrícola, 2020) where is estimated that 72.6% of the production is exported to the United States generating foreign exchange for 733.9 thousands of dollars, placing itself as the agricultural product with the highest export value (Estimación del volumen y valor de los principales productos agrícolas exportados 2020 - Sonora, 2020).

The United States market presents several advantages for the CPUMS such as high prices and for 15 or 20 days of commercialization, where Mexico is the only supplier, since it coincides with the exit of Chile from the market and the entry of California , which provides more competitive prices, however, this market is characterized by an increase in quality standards, safety and social responsibility, which causes an increase in the cost structure for producers, added to relatively stable prices causes that profit margins are reduced.

Since it is an open production system, it depends on various factors such as climate, natural disasters, economy, politics, and opportunity, making the commercialization period uncertain, which, added to the, the sustainability of the system is in risk.

That is why this work proposes a methodology based on data mining techniques that allow identifying market windows in presence and extinction, as well as characterizing price behavior through time series models, with the purpose of supporting decision-making and production planning for Sonoran and Mexican farmers.

AGRADECIMIENTOS

Agradezco de todo corazón a mi madre, padre y hermano que siempre estuvieron ahí para mi cuando lo necesite, brindándome apoyo y consejo en cada uno de ellos momentos.

A Rolando Valenzuela, Ivan Alexis, Roberto Guzmán y Diana Nuñez por ser de las personas que más me brindaron apoyo dentro de la maestría y que más que compañeros de clase, amigos.

A Michelle Irene, mi novia que siembre estuvo para mí en esos momentos de estrés y duda, que entendía cómo me sentía y me brindaba todo el apoyo y comprensión.

Al Dr. Martin Preciado, mi director de tesis y contacto en la empresa, por siempre estar apoyándome y guiándome, inclusive en esos momentos donde no había tiempo, por brindarme la oportunidad de aprender cosas nuevas y creer en mí.

Al Dr. Luis Felipe, por brindarle la oportunidad de realizar un proyecto en mi área de interés, por brindar apoyo siempre que se le necesito y por guiar a tomar mejores decisiones.

A todos mis compañeros y profesores de la maestría, por ser una inspiración y motivación en siempre hacer las cosas mejores.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa de Fortalecimiento de la Calidad Educativa (PFCE) por su apoyo económico brindado en mi estudio de posgrado, a mi familia por brindarme todo el apoyo y motivación para sacar adelante este proyecto y a mi director de tesis quien fue mi guía y apoyo en todo este proceso.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

En el presente capítulo se presenta la empresa en la cual fue desarrollado el proyecto, así como la descripción de la problemática a tratar, además se plantean los objetivos generales y específicos, así como la hipótesis y la delimitación del problema.

1.1. Presentación

El proyecto fue realizado en el Centro de Investigación en Alimentación y Desarrollo A.C. (CIAD), con dirección en la carretera Gustavo Enrique Astiazarán Rosas, N0. 46, Col. La Victoria, CP. 83304 en Hermosillo, Sonora, México.

La misión del CIAD, es “Contribuir al desarrollo sustentable y al bienestar de la sociedad en las áreas de alimentación, nutrición, salud, desarrollo regional y recursos naturales mediante la generación, aplicación y difusión de conocimiento científico-tecnológico, la innovación y la formación de recursos humanos de alto nivel”; y tiene como objetivo “Elevar la productividad y calidad en la generación de conocimiento científico mediante la investigación transdisciplinaria en grupos y redes” (CIAD., A.C., 2021).

El proyecto fue adscrito en la Coordinación de Desarrollo Regional, específicamente en la línea de investigación “Organización industrial, Mercado y Cadenas productivas”, la cual se basa en los postulados teórico-metodológicos de la organización industrial y el enfoque sistémico de cadenas productivas (CIAD., A.C., 2021).

Dentro de las cadenas productivas agroindustriales más estudiadas por este equipo de trabajo, está la de la Uva de Mesa Sonorense (CPUMS), Ésta es una de las cadenas productivas agroindustriales más importantes para el estado de Sonora, ya que cuenta con un volumen de producción anual de 320,990.10 toneladas en el año 2020, representando el 84.47% a nivel nacional teniendo un valor de producción de $9,296,739.69 (miles de pesos) en donde el 72.6% de la producción es exportada a Estados Unidos generando divisas por 733.9 millones de dólares colocándose como

el producto agrícola con mayor valor de exportación (“Estadística uva de mesa sonorense”, 2020).

De acuerdo con (Montaño y Preciado, 2017), la CPUMS enfrenta condiciones de vulnerabilidad y competitividad derivadas, entre otras cosas, por centrar su estrategia de mercado solo en Estados Unidos en el cual posee una ventana de comercialización de 15 o 20 días, en este periodo generalmente sólo México como proveedor del mercado estadounidense produce uva de mesa, lo cual le permite acceso al mercado de precios atractivos. El desempeño obtenido durante la ventana de mercado está determinado por el cumplimiento de las exigencias que impone el mercado estadounidense, el cual está caracterizado por ser exigente al imponer amplias restricciones de calidad, inocuidad y responsabilidad social, incrementando los costos de producción a los productores a fin de cumplir con dichas exigencias (Aranda Figueroa, 2016). Es a partir de estas evidencias, que surge la necesidad de profundizar en la comprensión del comportamiento de los precios y la ventana de mercado en Estados Unidos, a partir de un análisis de series de tiempo lo que permita coadyuvar a la determinación de la sostenibilidad de la CPUMS, en proyectos de investigación posteriores.

1.2. Planteamiento del problema

El mercado de Estados unidos presenta diversas ventajas para la CPUMS como lo son los precios altos de venta y que en ciertos periodos de comercialización, México se encuentra como único proveedor de este producto, lo que brinda precios más competitivos, sin embargo, este mercado se encuentra caracterizado por un incremento en estándares de calidad, inocuidad y responsabilidad social, lo que provoca un aumento en la estructura de costos para los productores, que sumado a precios relativamente estables, provoca que los márgenes de utilidad se vean reducidos. Una de las ventajas competitivas que está en riesgo, es la ventana de comercialización de altos precios que por un lado coincide con salida del país de Chile como proveedor de este mercado y el inicio de la exportación de Sonora, así mismo, los 15 o 20 días donde solo México exporta este producto a Estados Unidos es limitada

por el inicio de la comercialización por parte de California. Dado que es un sistema abierto de producción, este depende de muchos factores como lo son el clima, desastres naturales, la economía o política, el periodo de comercialización se vuelve incierto en cuanto al inicio y la duración de la ventana de mercado, teniendo como consecuencia que la sostenibilidad del sistema se encuentra en riesgo.

Es por eso por lo que la presencia y extensión de la ventaja competitiva del CPMUS, requiere ser analizada de forma tal que reduzca la incertidumbre y al mismo tiempo permita tomar decisiones para reducir el riesgo de la sostenibilidad del sistema. Es a partir de estos hechos, que es preciso comprender el mercado estadounidense de la uva de mesa, a través de analizar del comportamiento del precio y oferta de la uva utilizando modelos de análisis de series de tiempo que permitan identificar el inicio y duración de ventana de comercialización de precios altos.

1.3. Objetivo general

Desarrollar un modelo explicativo de series de tiempo a través de técnicas de minería de datos, que permita identificar el comportamiento de los precios de venta y la ventana de mercado de la uva de mesa sonorense en Estados Unidos para llevar a cabo futuras investigaciones y estudiar la sostenibilidad del sistema de la CPUMS.

1.4. Objetivos específicos

- Conformar una base de datos con los precios de venta de la uva de mesa sonorense en el mercado estadounidense a partir de información del Departamento de Agricultura de los Estados Unidos (USDA) del año 2000 al 2020.
- Desarrollar un método para identificar las ventanas de mercado.
- Desarrollar un modelo de series de tiempo que permita identificar los parámetros que mejor expliquen el comportamiento de los precios.
- Determinar el nivel de error del modelo desarrollado.
- Caracterizar el comportamiento de los precios de venta de la uva de mesa sonorense en el mercado estadounidense.

1.5. Hipótesis

El desarrollo de un modelo de series de tiempo basado en técnicas de minería de datos permitirá caracterizar el comportamiento de los precios de venta, así como identificar y delimitar la ventana de mercado de la uva de mesa sonorense en Estados Unidos.

1.6. Alcances y delimitaciones

Se estudiará el comportamiento de los precios de la uva de mesa sonorense y se identificará y delimitará la ventana de mercado en EE. UU a través de la modelación de series de tiempo. La base de datos utilizada se conformó a partir del sistema “Agronometrics” con información de la USDA y la temporalidad comprende del año 2000 al 2020.

1.7. Justificación

El proyecto será realizado debido a la situación de competitividad y vulnerabilidad en la que se encuentra la cadena productiva de la uva de mesa sonorense, que al ser el producto con mayor valor de exportación para el estado y que genera una derrama económica importante, justifica estudiar el comportamiento de los precios de venta y delimitar la ventana de mercado a través del análisis de series de tiempo lo que permita determinar la sostenibilidad de la CPUMS, en proyectos de investigación posteriores y apoyará a la toma de decisiones de los productores.

2. MARCO DE REFERENCIA

En esta sección se describen los términos relacionados con el descubrimiento de conocimiento en base de datos, minería de datos, series de tiempo y algoritmos de minería de datos que dan sustento a la metodología utilizada y a la implementación de este trabajo de investigación.

2.1 Descubrimiento de conocimiento en bases de datos

Cada generación de avances técnicos ya sea en inteligencia artificial, gráficos por computadora, conectividad electrónica, redes sociales, entre otros (Moreno *et al.*, 2017), se promete ayudar a las personas a comprender y administrar mejor una gran cantidad de actividades, desde el análisis financiero hasta los datos de monitoreo de las misiones espaciales, el control del espacio aéreo, entre otros. Ciertamente, esta informatización del mundo moderno ha avanzado enormemente nuestra capacidad para recopilar, transmitir y transformar datos, produciendo niveles de acceso a los datos sin precedentes (Woods *et al.*, 1999). Un ejemplo de esto es que en el año 2010 se generaron 2 zettabytes de información y en el 2020 un total de 59, teniendo un aumento del 2950% y se prevé que en 4 años más esta cantidad aumente en un 253% más, como se puede ver en la Figura 2.1 (Statista, 2021).

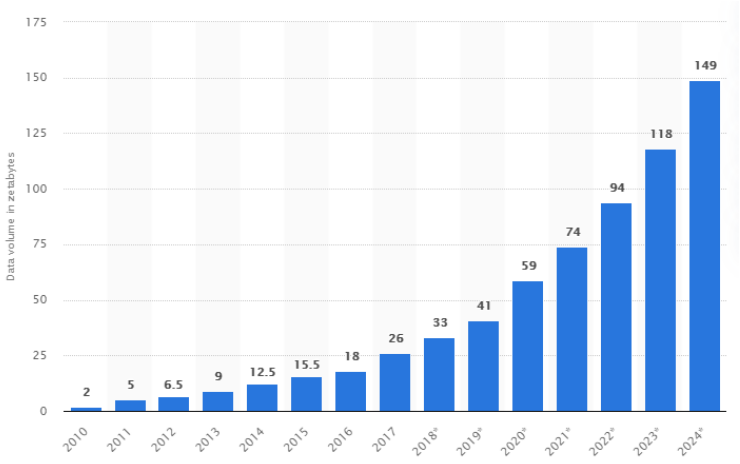

Figura 2.1. *Volumen de datos / información creados, capturados, copiados y consumidos en todo el mundo desde 2010 hasta 2024 (Statista, 2021).*

Esto ha generado una problemática conocida como sobrecarga de datos o información (Fayyad, Piatetsky-Shapiro and Smyth, 1996) la cual ocurre cuando la cantidad de entrada a un sistema excede su capacidad de procesamiento (Speier, Valacich y Vessey, 1999). Científicos e investigadores se han enfrentado a este problema debido la existencia y generación constante de millones de datos que por su volumen quedan en espera de ser convertidos en información útil (Fayyad, Haussler y Stolorz, 1996).

Poco a poco nos hemos ido acostumbrando al hecho de que hay enormes volúmenes de datos llenando nuestras computadoras, redes y vidas. Las agencias gubernamentales, las instituciones científicas y las empresas han dedicado enormes recursos económicos y humanos, con el propósito de recopilar y almacenar datos, sin embargo, la realidad es que solo se utilizará una pequeña cantidad de éstos, porque en muchos casos, los volúmenes son simplemente demasiado grandes para gestionarlos o no presentan la estructura adecuada para ser analizados de manera eficiente (Kantardzic, 2011) .

La generación de grandes de bases de datos aunado a la dificultad para procesarlas, generó un gran interés en el desarrollo de métodos y tecnologías que hiciera posible este proceso de manera más eficiente (Matheus, Chan y Piatetsky-Shapiro, 1993) y reducir el desfase entre la reproducción de datos y el descubrimiento de conocimiento (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

A partir de estos hechos, así como el avance de la tecnología, surge el concepto “Descubrimiento de Conocimiento en Bases de Datos” (KDD, por sus siglas en inglés) el cual es definido como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles en los datos (Fayyad, Piatetsky-Shapiro and Smyth, 1996). Otra definición es la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos (J. Frawley, Piatetsky-Shapiro y J. Matheus, 1992), algunos autores se refieren a este concepto como el proceso no trivial de encontrar e interpretar patrones en los datos mediante la aplicación repetida de métodos y algoritmos de minería de datos y la

interpretación de los patrones generados (Becerra-Fernandez, Zanakis y Walczak, 2002). Todas estas definiciones están dirigidas al desarrollo de métodos, técnicas y herramientas que apoyan a los analistas en el proceso general de descubrir información y conocimiento útil en bases de datos, siendo un proceso interactivo e iterativo entre un humano y una base de datos que involucra fuertemente el conocimiento previo del experto en análisis de dominio (Stumme, Wille y Wille, 1998).

Siendo entonces el principal objetivo del KDD extraer conocimiento de los datos, mediante el procesamiento e identificación automática de los patrones más significativos y presentarlos como conocimiento apropiado para lograr el objetivo del usuario (Matheus, Chan y Piatetsky-Shapiro, 1993).

Cuando se habla de la extracción de algo no trivial, hace a la necesidad de realizar inferencias, que implica un proceso computacional que va más allá del cálculo de estadísticos descriptivos y los patrones descubiertos deben ser válidos novedosos con cierto grado de certeza y potencialmente útiles para generar algún beneficio para el usuario (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

2.2 Metodologías para el proceso KDD

En este apartado se explicarán algunas de los esquemas o metodologías de manera detallará para llevar a cabo un proceso KDD según diversos autores.

El proceso del KDD es multidisciplinario, ya que involucra diversos campos de investigación como: aprendizaje automático, reconocimiento de patrones, bases de datos, estadística, inteligencia artificial, adquisición de conocimiento de sistemas expertos, visualización de datos, computación de alto rendimiento en donde estas disciplinas convergen con el objetivo de obtener información de alto valor de los datos para el apoyo en la toma de decisiones (Preciado Rodriguez, Romero Dessens y Ojeda Benítez, 2011; Gamarra, Guerrero y Montero, 2016).

La definición de pasos del proceso KDD también puede tener un fuerte impacto en los resultados finales de la adquisición de conocimiento de una base de datos, un ejemplo

de esto es que los algoritmos de minería de datos (DM) presentan dificultades para encontrar información útil si los atributos seleccionados no pueden representar completamente las características de los datos o que el proceso de preparación de los datos no haya sido el adecuado (Tsai *et al.*, 2014).

2.2.1 Modelo Fayyad

Este proceso ha sido descrito por diversos autores, todos con enfoques ligeramente diferentes y etapas, uno de ellos es el propuesto por (Fayyad, Piatetsky-Shapiro y Smyth, 1996) el cual es iterativo e interactivo en donde el usuario se involucra en la toma de decisiones en varios de los pasos, a continuación en la Figura 2.2 se presenta el esquema gráfico de este modelo.

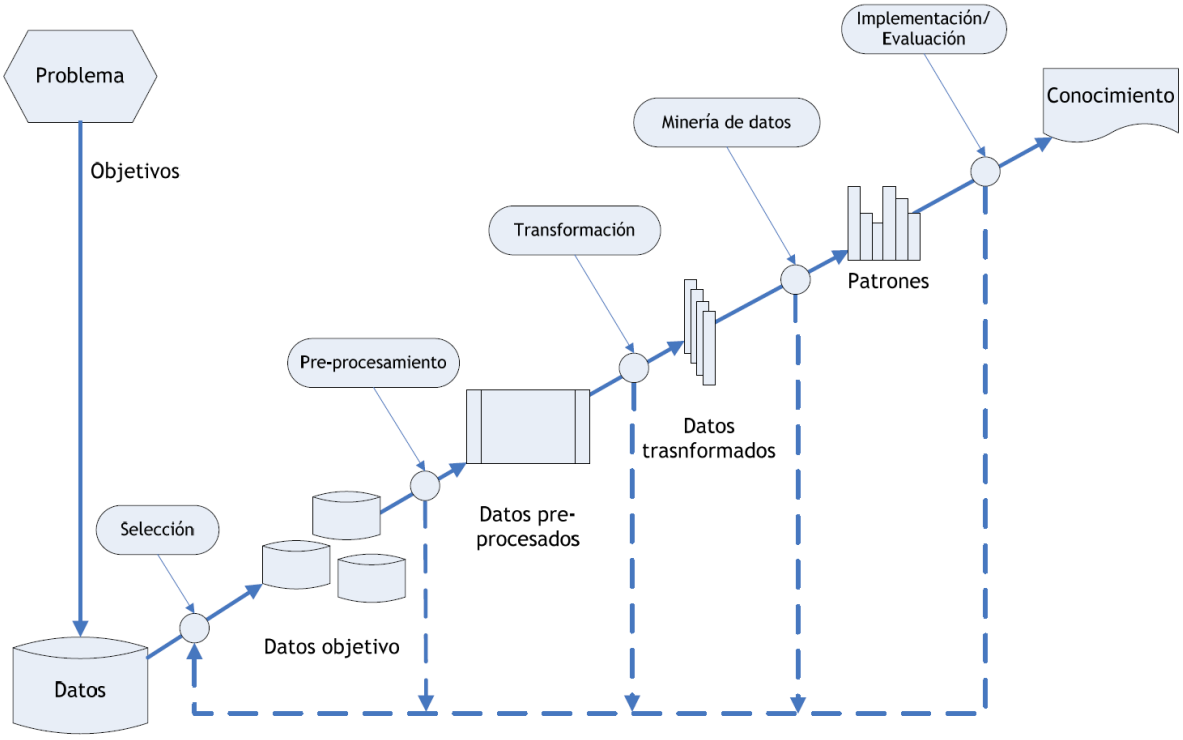

Figura 2.2. *Proceso KDD (Fayyad, Piatetsky-Shapiro and Smyth, 1996 en Preciado, 2011).*

A continuación, se explicarán de manera breve cada una de las etapas:

1. **Identificación y entendimiento del problema**: El primer paso consiste en una comprensión del dominio de la aplicación del problema y el conocimiento previo relevante y la identificación del objetivo del proceso KDD desde el punto de vista del cliente.
2. **Creación del conjunto de datos**: Lo siguiente consiste en la creación o selección de una base de datos objetivo o de solo un subconjunto de esta.
3. **Limpieza y procesamiento previo de datos**: Este paso consiste en la limpieza de los datos y el preprocesamiento de los datos, a grandes rasgos consiste en llevar a cabo operaciones básicas que incluyen eliminar el ruido si es apropiado, recopilar la información necesaria para modelar o contabilizar el ruido, decidir estrategias para manejar los campos de datos faltantes, entre otros.
4. **Reducción y proyección de datos**: El cuarto paso es la reducción y proyección de los datos, encontrar características útiles para representar los datos según el objetivo de la tarea con los métodos de transformación o reducción de la dimensionalidad.
5. **Identificar un método o técnica analítica en particular, de la minería de datos, de acuerdo con el objetivo de la aplicación del proceso KDD.**
6. **Exploración de análisis, modelo y selección de hipótesis**: En este paso se realiza un análisis exploratorio de los datos, la selección de la hipótesis a probar, los modelos de minería que serán utilizados y los parámetros a tomar en cuenta.
7. **Minería de datos**: Consiste en la utilización de técnicas o algoritmos de minería de datos para encontrar patrones de interés en los datos, incluyendo reglas o árboles de decisión, regresión y agrupamiento.
8. **Interpretación de los patrones minados**: Como penúltimo paso se interpretan los patrones minados, en donde posiblemente se tendrá que regresar a cualquiera de las etapas anteriores realizando iteraciones entre los pasos. Además, este paso podría involucrar la visualización de los patrones y modelos obtenidos o la visualización de los datos en los modelos extraídos.

9. **Aplicación del conocimiento descubierto**: El último paso consiste en la aplicación del conocimiento descubierto directamente, incorporarlo en otro sistema para acciones posteriores o simplemente documentarlo y reportarlo a las partes interesadas. Esta etapa también incluye la verificación y resolución de posibles conflictos con el conocimiento previamente adquirido o extraído.

2.2.2 Modelo Williams

Otro modelo es el presentado por Williams y Huang (1996) el cual sintetiza las nueve etapas mencionadas por Fayyad, Piatetsky-Shapiro and Smyth (1996) en cuatro principales de alto nivel que se encuentran en tareas sencillas las cuales se pueden ver en la Figura 2.3.

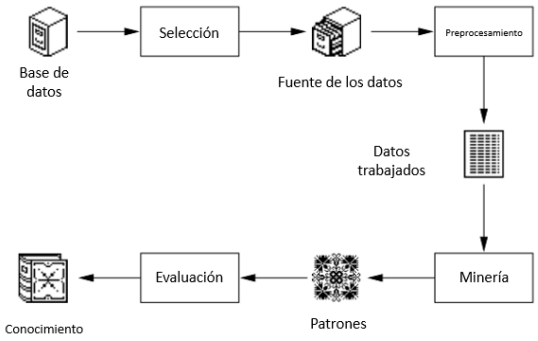

Figura 2.3 *El proceso KDD de cuatro etapas (Williams y Huang, 1996).*

A continuación, se describirán cada una de las etapas:

1. **Selección:** La etapa inicial consiste en seleccionar del conjunto de datos originales aquellos que sean de interés y los parámetros que serán utilizados para descubrir el conocimiento.
2. **Preprocesamiento:** El propósito de esta etapa es limpiar los datos tanto como sea posible y estructurarlos de forma adecuada para el procesamiento con las herramientas de minería a utilizar.
3. **Minería:** Una vez que el conjunto de datos de trabajo tiene el formato adecuado, inicia el proceso de minería, donde se emplea una variedad de herramientas

para explorar diferentes tipos de patrones en los datos de trabajo. La etapa de minería en sí es a menudo cíclica entre las tres etapas que la conforman: identificación del modelo, estimación de parámetros y validación del modelo.

4. **Evaluación:** Consiste en evaluar aquellos patrones encontrados para identificar aquellos que den lugar a conocimientos útiles y novedosos.

2.2.3 Modelo de Ho

Un modelo más reciente es el planteado por (Ho, 2017) que parte como base del proceso KDD presentado por Fayyad, Piatetsky-Shapiro and Smyth (1996) y del proceso estándar de la industria para la minería de datos (CRISP-DM, por sus siglas en inglés) (el cual se explica más adelante en este capítulo) para crear una metodología mixta o hibrida entre estos dos modelos como se puede ver en la Figura 2.4.

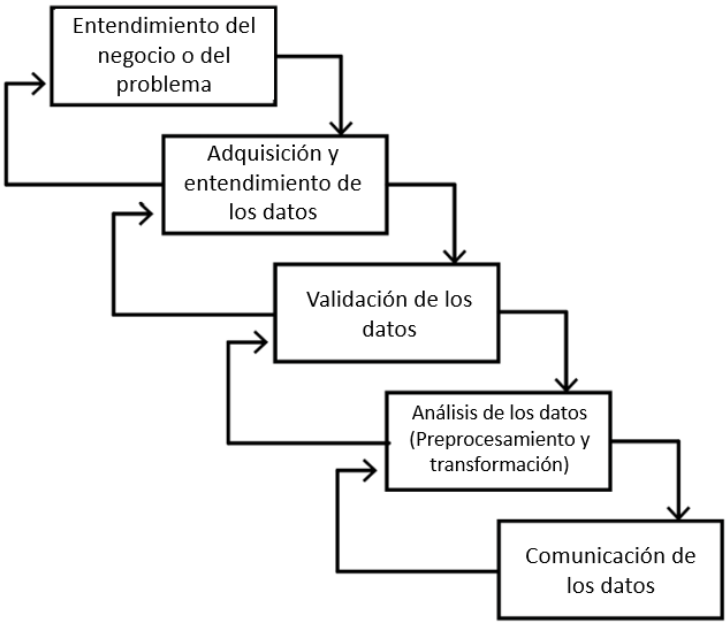

Figura 2.4 *Etapas de proceso de descubrimiento de conocimientos (Ho, 2017).*

1. **Entendimiento del negocio o del problema:** Implica trabajar en estrecha colaboración con las partes interesadas para definir un problema y determinar los objetivos del problema, la terminología, las preguntas e identificar las partes

interesadas clave. Los objetivos y preguntas del proyecto se traducen en datos que se pueden entender como objetivos técnicos.

2. **Adquisición y comprensión de datos:** Este paso incluye la recopilación, entendimiento de los datos y la decisión de qué herramientas o métodos se deben utilizar para resolver el problema. A menudo se presenta un ciclo de retroalimentación en esta fase, ya que existe la necesidad de conocimientos de dominio adicionales para comprender los datos. Esta fase se puede dividir en cuatro pasos principales: 1) recopilación, 2) descripción, 3) exploración 4) verificación de la calidad.
3. **Validación de los datos**: Los datos se verifican para comprobar que estén completos y evitar que sean redundantes o falten valores. Uno de los aspectos más importantes de esto es verificar la utilidad y patrones de los datos con respecto al problema inicial.
4. **Análisis de los datos:** En este paso, se deben encontrar patrones dentro del conjunto de datos. La evaluación incluye comprender patrones, validarlos e interpretarlos. En algunos casos, puede ser minería de datos o modelado de datos. Una vez que se establece un patrón, se deben verificar los resultados para ver si los conocimientos son nuevos e interesantes para el objetivo del proyecto.
5. **Visualización de los datos y comunicación del conocimiento:** Una vez que se interpretan los resultados, estos se comunican a las partes interesadas. Este es un paso crítico en este proceso, ya que es la forma en la que se proporciona valor a la organización convirtiendo los patrones descubiertos en conocimiento practico.

La principal diferencia de este modelo es que se proporciona una descripción más general orientada a la investigación de cada paso, además de introducir un paso de análisis de datos y reducir la importancia de la minería de datos.

2.3 Minería de datos

El termino de Minería de Datos (MD), usualmente se ha utilizado como sinónimo del Knowledge Discovery in Databases (Descubrimiento de conocimiento en bases de datos), sin embargo, el termino MD se refiere a un paso en particular del proceso KDD (Fayyad y Stolorz, 1997) que consiste en aplicar algoritmos de descubrimiento y análisis de datos que, bajo limitaciones aceptables de eficiencia computacional, producen una enumeración particular de patrones o modelos sobre los datos (Fayyad, Piatetsky-Shapiro y Smyth, 1996).

Otros autores definen la MD como el proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenados ya sea en bases de datos, un almacén de datos o algún otro repositorio de información, a través de la construcción de un modelo predictivo o descriptivo (Chen *et al.*, 2015) o como un método de análisis empresarial que van más allá de los recuentos, las técnicas descriptivas, la generación de informes y los métodos basados en reglas empresariales (Shmueli, C. Bruce y R. Patel, 2017)

Los dos objetivos principales de la minería de datos en la práctica tienden a ser la predicción o clasificación (modelos supervisados) y la descripción (modelos no supervisados) (Jothi, Rashid y Husain, 2015).

2.3.1 Aprendizaje automático

El aprendizaje automático explora el estudio y la construcción de algoritmos que pueden aprender y hacer predicciones sobre los datos. Samuel (1988), definió el aprendizaje automático como un "campo de estudio que brinda a las computadoras la capacidad de aprender sin estar programadas explícitamente", también se entiende que el aprendizaje automático es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante. Se les considera el caballo de batalla en la nueva era del “big data”, en donde un algoritmo de aprendizaje automático es un proceso computacional que utiliza datos de entrada para lograr una tarea deseada sin estar

literalmente programado (es decir, "codificado") para producir un resultado en particular. Estos algoritmos están, en cierto sentido, "programados por software" en el sentido de que alteran o adaptan automáticamente su arquitectura mediante la repetición (es decir, la experiencia o a veces conocido como entrenamiento) para que sean cada vez mejores en el logro de la tarea deseada. El proceso de adaptación se denomina formación o entrenamiento, en el que se proporcionan muestras de datos de entrada junto con los resultados deseados, en donde usualmente conforman el 80% del total del conjunto de datos, mientras que el otro 20% se utiliza para validar que tan buenos resultados brinda el algoritmo. Luego, el algoritmo se configura de manera óptima para que no solo produzca el resultado deseado cuando se le presenten las entradas de entrenamiento, sino que pueda generalizar para producir el resultado deseado a partir de datos nuevos que nunca se habían visto. Esta formación es la parte de "aprendizaje" del aprendizaje automático (El Naqa y Murphy, 2015).

El ideal del aprendizaje automático es emular la forma en que los seres humanos (y otras criaturas sensibles) aprenden a procesar señales sensoriales (de entrada) para lograr un objetivo. Este objetivo podría ser una tarea en el reconocimiento de patrones, en la que el alumno quiere distinguir las manzanas de las naranjas. Cada manzana y cada naranja es única, pero todavía podemos (normalmente) distinguir una de la otra. En lugar de codificar una máquina con muchas representaciones exactas de manzanas y naranjas, se puede programar para aprender a distinguirlas a través de la experiencia repetida con manzanas y naranjas reales (El Naqa y Murphy, 2015).

De manera general el aprendizaje automático se puede entender como el desarrollo de los algoritmos y modelos con los cuales una computadora puede aprender a interpretar datos de manera automática, prediciendo valores dada una entrada o clasificando una serie de datos, mientras que la minería de datos sería el proceso de implementación de estos algoritmos para obtener conocimiento nuevo (Buczak y Guven, 2016). Otro autor define al aprendizaje automático como el campo de investigación dedicado al estudio formal de los sistemas de aprendizaje. Este es un campo altamente interdisciplinario que toma prestadas y se basa en ideas de

estadística, informática, ingeniería, ciencia cognitiva, teoría de la optimización y muchas otras disciplinas de la ciencia y las matemáticas (Ghahramani, 2004).

Una vez que el modelo ya ha aprendido de los datos de aprendizaje, se procede a evaluar el desempeño del modelo que ha sido creado y usualmente los pasos involucrados en tal evaluación son (i) calcular los resultados de los métodos en los datos de prueba (ii) calcular los errores de estos resultados con respecto a los datos de referencia bajo el uso de una función de pérdida, (iii) calcular una medida de error de los errores, y (iv) utilizar un procedimiento de selección de modelo para examinar la distribución de la medida del error y / o encontrar el mejor modelo que se utilizaría en la aplicación final, a este proceso se le conoce como validación cruzada (Bergmeir y Benítez, 2012).

2.3.2 Modelos supervisados

El aprendizaje supervisado o predictivo es un modelo de aprendizaje automático para adquirir la información de la relación entrada-salida de un sistema en función de un conjunto determinado de muestras de formación de entrada-salida emparejadas. La salida de este sistema se considera como la etiqueta de los datos de entrada o la supervisión, una muestra de entrenamiento de entrada-salida también se denomina datos de entrenamiento etiquetados o supervisados que en ocasiones, también se lo conoce como aprender con un profesor (Haykin, 1999), aprendizaje a partir de datos etiquetados o aprendizaje automático inductivo (Kotsiantis, 2007). El objetivo del aprendizaje supervisado es construir un sistema artificial que pueda aprender a través de una función de mapeo entre la entrada y la salida, donde esta función permita predecir o estimar la salida del sistema con nuevas entradas, y se dividen en: modelos de estimación y modelos de clasificación.

La clasificación es el proceso de encontrar un modelo (o función) que describe y distingue clases de datos o conceptos, con el propósito de poder usar el modelo para predecir la clase de objetos cuya etiqueta de clase se desconoce (Han y Kamber, 2006). En estos tipos de modelos los grupos o clases objetivo son conocidos con anterioridad; éstos tienen el objetivo de clasificar los casos dentro de tales clases. Así

mismo, también calculan el puntaje de propensión que marca la posibilidad de ocurrencia de un grupo objetivo o evento (P., 2011).

Los modelos de estimación aplican funciones de aprendizaje supervisado para predecir valores continuos desconocidos o futuros de otras variables de interés teniendo como objetivo producir un modelo, expresado como un código ejecutable, que se puede utilizar para realizar clasificación, predicción, estimación u otras tareas similares (Kantardzic, 2011). En resumen este tipo de modelos se caracterizan por la realización de inferencias sobre los datos actuales para predecir las variables de salida (Han y Kamber, 2006).

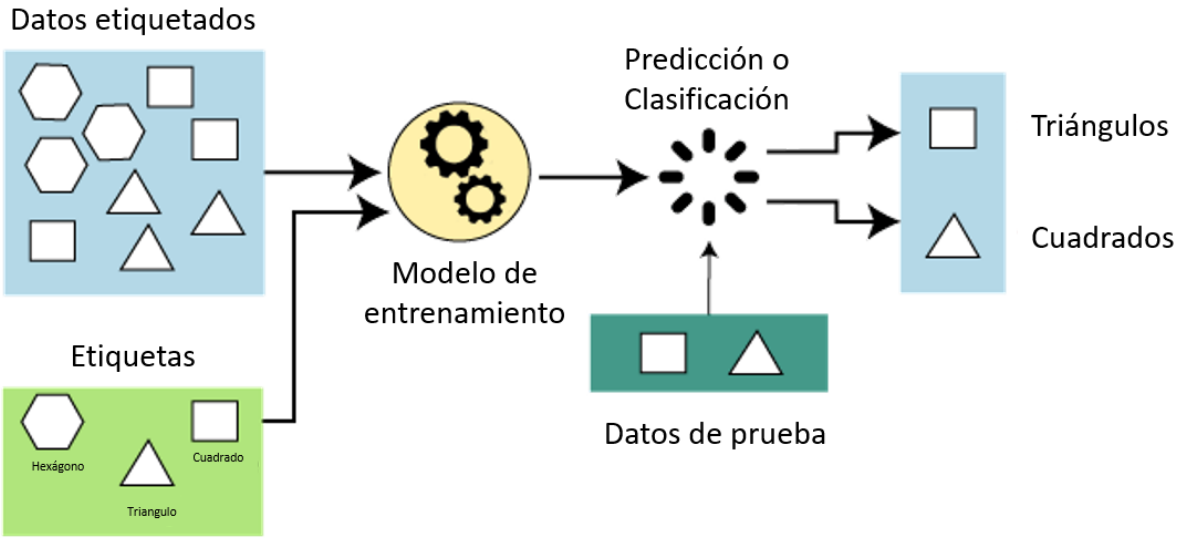

Figura 2.5. *Funcionamiento de un modelo de aprendizaje supervisado (javatpoint, 2019)*

En la Figura 2.5 se explica en funcionamiento de un modelo de aprendizaje supervisado, supóngase que se tiene un conjunto de datos de diferentes tipos de formas que incluyen cuadrado, rectángulo, triángulo y polígono. Ahora el primer paso es que se requiere entrenar el modelo para cada forma.

- Si la forma dada tiene cuatro lados y todos los lados son iguales, entonces se etiquetará como un cuadrado.
- Si la forma dada tiene tres lados, entonces se etiquetará como un triángulo.

- Si la forma dada tiene seis lados iguales, se etiquetará como hexágono.

Ahora, después del entrenamiento, se prueba el modelo usando el conjunto de prueba, y la tarea del modelo es identificar la forma. La máquina ya está entrenada en todo tipo de formas, y cuando encuentra una nueva, la clasifica sobre la base de varios lados y predice el resultado (javatpoint, 2019).

Aunque los límites entre predicción y descripción no son muy claros (algunos de los modelos predictivos pueden ser explicativos, en la medida en que sean comprensibles e interpretables por el ser humano) la distinción es útil para comprender el objetivo de descubrimiento general (Fayyad, Piatetsky-Shapiro and Smyth, 1996).

2.3.3 Modelos no supervisados

En los modelos no supervisados, la máquina simplemente recibe entradas, pero no obtiene resultados de destino supervisados. Sin embargo, es posible desarrollar un marco formal para el aprendizaje no supervisado basado en la noción de que el objetivo de la máquina es construir representaciones de la entrada que se pueden utilizar para la toma de decisiones, comunicar de manera eficiente las entradas a otra máquina, etc. En cierto sentido, se puede pensar en el aprendizaje no supervisado como encontrar patrones en los datos por encima y más allá de lo que se consideraría puro ruido no estructurado. Dos ejemplos clásicos muy simples de aprendizaje no supervisado son el agrupamiento y la reducción de dimensionalidad (Ghahramani, 2004).

Los modelos descriptivos suelen aplicar las funciones de aprendizaje no supervisado con el objetivo de obtener una comprensión del sistema analizado al descubrir patrones y relaciones en grandes conjuntos de datos, pudiendo ser interpretados por el ser humano (Kantardzic, 2011). Este tipo de tareas caracterizan las propiedades generales de los datos en la base de datos (Han y Kamber, 2006).

2.3.4 Modelos de agrupamiento

En los modelos de clustering, llamados también de agrupamiento, los grupos de datos o sus clasificaciones no son conocidas con anterioridad. En su lugar, los algoritmos

analizan los patrones de datos de entrada e identifican el agrupamiento natural de los registros o casos. Por último, los modelos de asociación detectan dependencia entre eventos discretos, productos o atributos. (Preciado, 2011).

Existen dos tipos de modelos de agrupamiento principales los cuales son el clustering particional y el clustering jerárquico (Larrañaga, Inza y Moujahid, 2012), los cuales se explicarán a continuación.

Modelo de agrupamiento jerárquico

La técnica de agrupación jerárquica convierte los datos en grupos sobre un dendrograma, que contiene un árbol de agrupaciones en diferentes escalas (Habib, Hayat y Zucker, 2016). De manera diferente, el algoritmo de agrupamiento de k-medias desarrolla los grupos basados en el centro del grupo y del que cada uno tiene un miembro dependiendo del valor de aptitud más cercano, y los centros del grupo se actualizarán hasta que no haya cambios en ninguno de los centroides del mismo (Khawaja et al., 2017). La agrupación en clústeres jerárquica y la agrupación en clústeres de k-medias se utilizan ampliamente debido a su eficiencia, escalabilidad y simplicidad (Sun y Huang, 2020).

El análisis de conglomerados jerárquico se puede dividir en dos tipos: aglomerativos y divisivos (Saxena et al., 2017).

La agrupación aglomerada, también conocida como enfoque ascendente o agrupación aglomerativa jerárquica es un método de agrupamiento que no requiere la especificación previa del número de agrupaciones. Los algoritmos ascendentes tratan cada dato como un grupo único desde el principio y luego aglomeran sucesivamente pares de grupos hasta que todos los grupos se hayan fusionado en uno solo que contenga todos los datos. En el caso de la agrupación divisiva, también conocida como enfoque de arriba hacia abajo, tampoco requiere especificar previamente el número de clústeres. El agrupamiento de arriba hacia abajo requiere un método para dividir un grupo que contiene todos los datos y procede dividiendo los grupos de forma recursiva

hasta que los datos individuales se hayan dividido en un grupo único (Murtagh, 1983; ML | Hierarchical clustering (Agglomerative and Divisive clustering), 2019).

A continuación, se presentan algunos de los métodos más utilizados para la realización de clúster (Riyadi et al., 2017; Saxena et al., 2017).

Vinculación simple

Este tipo de agrupación en clústeres se llama a menudo como la conexión o el método mínimo. El vínculo entre dos grupos se realiza mediante un solo par de elementos, es decir, esos dos elementos están determinados por la distancia más cercana de cualquier miembro de un grupo a cualquier miembro de otro grupo, que define la similitud, si los datos son similares, la similitud entre un par de grupos es igual a la mayor similitud entre un miembro de un grupo y un miembro de los otros grupos.

Vinculación completa

Es uno de los métodos de agrupación que utilizan la distancia máxima entre los datos. Esta medida es similar a la medida de enlace único, la diferencia es un enlace único utilizando la distancia mínima.

Enlace promedio

Las reglas que utilizan el método de grupo de pares no ponderados mediante el promedio aritmético para superar las limitaciones de la vinculación única y completa que propone medir el promedio entre los datos. Se supone que este método representa un compromiso natural entre las medidas de vinculación para proporcionar una evaluación más precisa de la distancia entre los conglomerados.

Método WARD

También llamado método de suma incremental de cuadrados usa las distancias dentro de los conglomerados (cuadrados) y la distancia entre los conglomerados (al cuadrado) para encontrar la distancia que distinga de manera óptima cada uno de los grupos.

Una forma de saber cuál método es el más adecuado para la construcción de los grupos es mediante el coeficiente aglomerativo, que mide la cantidad de estructura de agrupamiento encontrada (los valores más cercanos a 1 sugieren una estructura de agrupamiento más fuerte), lo que permite encontrar ciertos métodos de agrupamiento jerárquico puedan identificar estructuras de agrupamiento más adecuadas, una forma automatizada de realizar esta prueba es a través del paquete “Agnes” en el lenguaje de programación R (Hierarchical Cluster Analysis, 2020).

Las técnicas de agrupamiento no jerárquico son métodos en el análisis de agrupamiento que requiere definir el número de grupo antes de realizar el proceso de agrupamiento. Posteriormente, el algoritmo permitirá que los objetos se agrupen en función del centroide más cercano. El centroide se calcula mediante la media entre los objetos de cada grupo, siendo un problema de optimización teniendo como objetivo minimizar la distancia entre cada uno. El procedimiento de k-medias se puede definir como: 1. Establecer el número de grupos con k grupos; 2. Elección aleatoria de los centroides iniciales; 3. Calcular la distancia para volver a calcular la media de cada grupo que se establecerá como el nuevo centroide. 4. Repetir los pasos 2 y 3 hasta que no haya más reasignaciones para cada objeto o se optimice el valor de la función objetivo (Fulcher y Jones, 2014).

Modelo de agrupamiento particional

El agrupamiento particional divide un conjunto de datos en varios grupos en función de cierto criterio conocido como medida de aptitud o cohesión del clúster. La medida de aptitud afecta directamente la naturaleza de la formación de grupos. Una vez que se selecciona una medida de aptitud adecuada, la tarea de partición se convierte en un problema de optimización (Nanda y Panda, 2014). En el clustering particional el objetivo es obtener una partición de los objetos en grupos de tal forma que todos los objetos pertenezcan a alguno de los k clústeres posibles y que por otra parte todos sean disjuntos (Larrañaga, Inza y Moujahid, 2012). Al particionar la agrupación en clústeres, se debe especificar la cantidad de agrupaciones, lo que puede llevar a una agrupación incorrecta del conjunto de datos dado, mientras que la agrupación

jerárquica y basada en la densidad elige la cantidad de agrupaciones por sí mismos (Ogbuabor y F. N, 2018).

A diferencia de los métodos de agrupamiento jerárquico, el agrupamiento particional tiene como objetivo agrupaciones sucesivas utilizando algunos procesos iterativos. La agrupación en clústeres particional asigna un conjunto de puntos de datos en k-clústeres mediante el uso de procesos iterativos. En estos procesos, n datos se clasifican en k grupos. La función de criterio predefinida asigna la fecha en k-ésimo conjunto de números de acuerdo con el cálculo de maximización y minimización en k conjuntos (Kutbay, 2018).

2.3.5 Validación del modelo de aprendizaje automático

Una cuestión central en el aprendizaje supervisado se refiere a la precisión del modelo resultante. Aquí, un problema clave es el sobreajuste. Es muy fácil construir un modelo que se adapte perfectamente al conjunto de datos disponible, pero que luego no pueda generalizar bien a datos nuevos y no vistos (Berrar, 2018). El objetivo es maximizar su precisión predictiva en los nuevos puntos de datos, no necesariamente su precisión en los datos de entrenamiento. De hecho, si se trabaja demasiado para encontrar la variación que mejor se ajuste a los datos de entrenamiento, existe el riesgo de que se ajuste el ruido en los datos, al memorizar varias peculiaridades de los utilizados en el entrenamiento, en lugar de encontrar una regla predictiva general, este fenómeno suele ser llamado "Sobreajuste" (Dietterich, 1995; Nitish *et al.*, 2014).

¿Cómo se evalúa la capacidad de generalización de un modelo? Idealmente, se evaluaría el modelo usando nuevos datos que se originan en la misma población que los datos usados para construir el modelo. En la práctica, los nuevos estudios de validación independientes a menudo no son factibles (Simon, 2003; Berrar, 2018), generalmente para que el modelo puede aprender las relaciones en los datos y evitar un sobreajuste se suele utilizar de un 70% a 90% de los datos para enseñar al modelo de aprendizaje automático y un 30% a 10% para validar los resultados obtenidos, considerando que preferentemente se tiene que utilizar el 100% (Guyon, 1997).

Una de las formas mayormente utilizada para la validación de los modelos predictivos o de clasificación, o también usado para el submuestreo de los datos, es la validación cruzada el cual es un método para evaluar la capacidad de generalización de los modelos predictivos y evitar el sobreajuste (Zhang y Yang, 2015; Roberts et al., 2017). Uno de los motivos principales por el cual es utilizado este método es que en la práctica el obtener nuevos datos para ver si un modelo funciona muchas veces no es posible (Zhang y Yang, 2015).

El funcionamiento de una validación cruzada de n veces, los datos se dividen en n partes iguales. La primera parte se utiliza como conjunto de datos de prueba, el resto se utiliza como conjunto de datos de calibración. Luego, la segunda parte se usa para los datos de prueba y el resto se usa para una nueva calibración. Este procedimiento se repite n veces y se promedian las predicciones de los n datos de prueba. Es esencial que ningún conocimiento de los modelos se transfiera de un pliegue a otro. No existen reglas claras sobre cuántos pliegues utilizar para la validación cruzada, por lo que la forma más sencilla y clara de realizar la validación cruzada es dejar una muestra a la vez (Dieterle, 2019)

En la Tabla 2.1 se enlistarán algunas de las medidas de error comúnmente utilizadas en los modelos de aprendizaje supervisado de predicción.

#	Métrica	Definición	Formula	Comentario
1	Error (E)	Cantidad en la que una observación difiere de su valor real.	$E = A - P$	Intuitivo y fácil de aplicar, siempre y cuando se eleve al cuadrado y se obtenga la raíz cuadrada para hacerlo positivo.

2	Error promedio (ME)	El promedio de todos los errores en un conjunto.	$ME = \frac{\sum_{i=1}^{n} E_i}{n}$	Puede que no sea útil en los casos en que las predicciones positivas y negativas se anulan entre sí.
3	Error promedio porcentual (MPE)	Promedio calculado de errores porcentuales.	$MPE = \frac{\sum_{i=1}^{n} E_i/A_i}{n/100}$	Indefinido siempre que un único valor real sea cero
4	Error absoluto medio (MAE)	Mide la diferencia entre dos variables continuas.	$MAE = \frac{\sum_{i=1}^{n} \lvert E_i \rvert}{n}$	Puede usarse para comparar series de diferentes escalas.
5	Error de porcentaje absoluto medio (MAPE)	Mide el alcance del error en términos porcentuales.	$MAPE = \frac{100}{n} \sum_{i=1}^{n} \lvert E_i \rvert / \lvert A_i \rvert$	• No se puede utilizar si existen valores cero reales. • No existe un límite superior para el porcentaje de error en las predicciones que son demasiado altas. • No simétrico (afectado negativamente si un valor predicho es mayor o menor que el

				valor real correspondiente)
6	Error cuadrático medio (MSE)	Mide el promedio de los cuadrados de los errores.	$MSE = \frac{\sum_{i=1}^{n} E_i^2}{n}$	• Depende de la escala. • Valores atípicos muy ponderados. • Muy dependiente de la fracción de datos utilizados (baja fiabilidad).
7	Raíz del error cuadrático medio (RMSE)	Raíz cuadrada del error cuadrático medio	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} E_i^2}{n}}$	• Depende de la escala. • Un valor más bajo para RMSE es favorable. • Sensible a valores atípicos. • Muy dependiente de la fracción de datos utilizados (baja fiabilidad)

Tabla 2.1 *Lista de métricas de error comúnmente utilizadas en modelos de pronóstico (Fourier, 1999; Hyndman y Koehler, 2006; Shcherbakov et al., 2013; M.Z. y Amir H., 2020).*

2.4 Metodologías para proyectos de minería de datos

El simple entendimiento de los algoritmos utilizados para el análisis de datos no es suficiente para un proyecto de minería de datos exitoso (Ho, 2017), lo que significa que el éxito de este tipo de proyectos depende en gran medida de la persona o el equipo en particular que lo lleve a cabo. La MD necesita un enfoque de trabajo estándar que ayude a traducir los problemas en tareas específicas, ya sean

transformaciones de datos, técnicas o algoritmos adecuados, medios para evaluar la efectividad de los resultados y documentar la experiencia (Wirth, 2000). A partir de esto es que nacen una serie de metodologías enfocadas a la realización de proyectos de MD. Al descubrimiento metodológico de relaciones y patrones útiles en los datos a través de un conjunto de actividades iterativas se conocería como el proceso de minería de datos (Kotu y Deshpande, 2015).

En la siguiente sección se mencionarán algunas de las principales metodologías más utilizadas en este tipo de trabajos o proyectos.

2.4.1 Cross Industry Standard Process for Data Mining (CRISP-DM)

CRISP – DM es una metodología para proyectos de minería de datos que fue concebida a finales de 1996, por cuatro líderes en el mercado de minería de datos con el fin de crear un modelo de proceso que estandarizara una metodología para la realización de proyectos de DM y ayudara a las organizaciones a lanzar sus propios proyectos de minería de datos (Shearer *et al.*, 2000).

CRISP-DM es una metodología integral que proporciona a cualquier persona, desde principiantes hasta expertos, el modelo de proceso completo para llevar a cabo un proyecto de minería de datos. CRISP-DM divide el ciclo de vida del proyecto en seis fases: comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación e implementación como se puede ver en la Figura 2.6 (Shearer *et al.*, 2000; Gupta *et al.*, 2019).

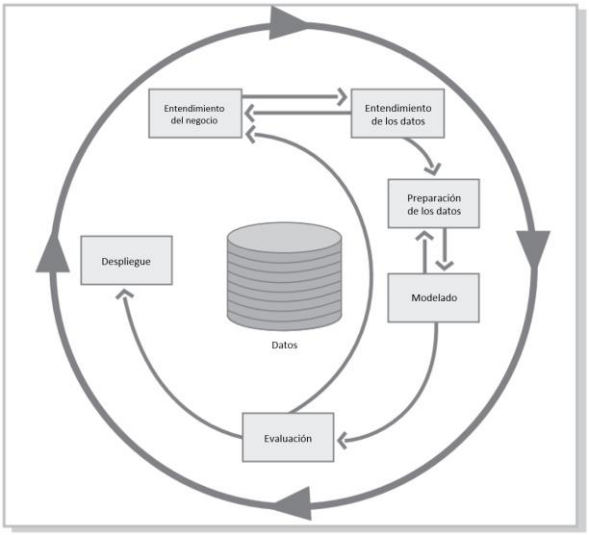

Figura 2.6 *Fases del modelo de referencia CRISP-DM (Pete et al., 2000).*

A continuación se esté explicaran las etapas de este modelo (Pete *et al.*, 2000; Wirth, 2000).

1. **Comprensión comercial:** Incluye la determinación de los objetivos empresariales, la evaluación de la situación actual, el establecimiento de objetivos de minería de datos y el desarrollo de un plan de proyecto.

2. **Comprensión de datos:** La fase de comprensión de datos inicia con una recopilación inicial de datos y continúa con actividades para familiarizarse con los datos, identificar si tienen problemas de calidad o detectar subconjuntos interesantes para formar hipótesis sobre información oculta.

3. **Preparación de datos:** Una vez que se identifican las fuentes de datos disponibles, debe seleccionarse una o varias de éstas, limpiarse y transformarse al formato adecuado/requerido. La limpieza y transformación de los datos, así como la preparación para el modelado debe ocurrir en esta fase.

4. **Modelado:** En esta fase se seleccionan e implementan las herramientas de software (algoritmos) de minería de datos, como la visualización (trazar datos y establecer relaciones), modelos predictivos o el análisis de conglomerados entre otros. Una vez

que se obtiene una mayor comprensión de los datos (a menudo mediante el reconocimiento de patrones que se activa al ver la salida del modelo), se pueden aplicar modelos más apropiados según el tipo de datos.

5. **Evaluación:** Los resultados del modelo deben evaluarse en el contexto de los objetivos comerciales establecidos en la primera fase (comprensión comercial). Esto conducirá a la identificación de otras necesidades, a menudo a través del reconocimiento de patrones o a través de la utilización medidas de la calidad del modelo, así como validación cruzada haciendo que frecuentemente se vuelva a fases anteriores de CRISP-DM. Obtener comprensión empresarial es un procedimiento iterativo en la minería de datos, donde los resultados de diversas herramientas de visualización, estadísticas e inteligencia artificial muestran al usuario nuevas relaciones que proporcionan una comprensión más profunda de las operaciones organizativas.

6. **Despliegue o implementación:** La fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible. En muchos casos será la persona interesada en el desarrollo del proyecto o el cliente, no el analista de datos, quien llevará a cabo los pasos de implementación. En cualquier caso, es importante comprender de antemano qué acciones deberán llevarse a cabo para poder hacer uso de los modelos creados.

2.5 Series de tiempo

Desde los precios del mercado de valores, la propagación de una epidemia, la grabación de una señal de audio hasta el monitoreo del sueño es común que los datos del mundo real se registren considerando alguna escala de tiempo. Cuando el acopio de datos se realiza de acuerdo a una escala de tiempo se conoce como una serie de tiempo (Wei, 2013; Gamboa, 2017). Una definición más formal es dada por (Esling y Agon, 2012) en donde una serie de tiempo T es una secuencia ordenada de n variables de valor real en donde $T = (t_1,..., t_n)$, $t_i \in R$, por tanto, una serie de tiempo puede

definirse como un conjunto de valores en instantes de tiempo contiguos, en donde estos pueden ser univariantes o multivariantes (vectores).

El objetivo principal de la minería de datos en el análisis de series de tiempo es extraer conocimiento significativo del comportamiento del fenómeno o problema a través del tiempo para realizar una predicción, clasificación, agrupamiento o explicar su comportamiento (Esling y Agon, 2012).

En la predicción de la series de tiempo, el objetivo es investigar las relaciones en el conjunto secuencial de datos pasados para conocer el valor o la tendencia a futuro (Puchalsky *et al.*, 2018; Yang y Liu, 2019). En la clasificación de las series de tiempo el objetivo es asignar etiquetas a cada serie de un conjunto. La principal diferencia en comparación con la tarea de agrupación es que las clasificaciones o etiquetas se conocen de antemano, el algoritmo se entrena en un conjunto de datos de ejemplo para aprender las características que las distinguen entre sí, después cuando se ingresa un conjunto de datos sin etiquetar en el sistema, puede determinar automáticamente a qué clase pertenece cada serie (Esling y Agon, 2012). En el caso de la agrupación de series de tiempo el objetivo es agrupar el conjunto de datos en grupos naturales, sin la necesidad de que ya exista una clasificación de antemano, es decir encontrar los conglomerados más homogéneos que sean lo más distintos posible de otros maximizando la variación entre grupos y minimizarla dentro de ellos. Por tanto, el algoritmo debería localizar automáticamente qué grupos están intrínsecamente presentes en los datos (Esling y Agon, 2012; Aghabozorgi, Seyed Shirkhorshidi y Ying Wah, 2015).

2.5.1 Características generales de una serie de tiempo

Dentro de las series de tiempo se cuenta con diversos tipos de modelos o enfoque, los cuales pueden ser univariantes o multivariantes cuando abarcan simultáneamente múltiples dimensiones dentro del mismo rango de tiempo (Esling y Agon, 2012), sin embargo, también se encuentran aquellas series de tiempo que siguen un comportamiento lineal, no lineal, estacional y no estacional.

Una de las características de las series de tiempo es la estacionalidad que es cuando el comportamiento de las variables se repite después de un periodo de tiempo regular (Wei, 2013). Un ejemplo de esta característica es cuando durante todos los meses del año en su cuarta semana hay una alza de ventas, pasando esta fecha, las ventas bajarían lo que generaría un patrón repetitivo y cíclico de corto período; otra característica que se debe de tomar en cuenta al momento de modelar o analizar series de tiempo son aquellas denominadas cíclicas en donde existe un patrón cíclico cuando los datos exhiben subidas y bajadas que no son de un período fijo como lo seria en el caso de la tendencia (Hyndman J, 2011) y por último, se encuentra el componente aleatorio el cual no responde a ningún patrón de comportamiento, sin o que es el resultado de factores fortuitos o aleatorios que inciden de forma aislada en una serie de tiempo. Por lo que una serie de tiempo se puede denotar como $X_t = T_t + E_t + I_t$ donde T_t es la tendencia, E_t es el componente estacional e I_t es el componente aleatorio (Madin Rivera, 2021).

Una serie de tiempo estacionaria significa que las propiedades estadísticas de su proceso no cambian con el tiempo. No significa que la serie no cambie con el tiempo, solo que la forma en que cambia no modifica sus parámetros (Palachy, 2019).

Otra definición de una serie de tiempo estacionaria es aquella cuyas propiedades estadísticas como la media, la varianza, la autocorrelación, etc., son todas constantes en el tiempo. La mayoría de los métodos de pronóstico estadístico se basan en el supuesto de que la serie de tiempo puede volverse aproximadamente estacionaria mediante el uso de transformaciones matemáticas como las diferencias o logaritmos (Nau, 2020a). En el caso de una serie de tiempo no estacionaria es aquella en la que su tendencia y/o variabilidad cambia con el tiempo.

2.5.2 Modelos ARIMA

Uno de los modelos de series de tiempo más importantes y ampliamente utilizados es el modelo autorregresivo integrado de promedios móviles (ARIMA). La popularidad de este modelo se debe a sus propiedades estadísticas, así como a la conocida

metodología Box-Jenkins (Geurts, Box y Jenkins, 1977; E. O. Box *et al.*, 2008), que se utiliza principalmente para el modelo de pronóstico en aplicaciones de series de tiempo financieras y de ciencia de datos. (Zhang, 2003).

En 1977, Box y Jenkins desarrollaron un cuerpo metodológico destinado a identificar, estimar y diagnosticar modelos dinámicos de series temporales en los que la variable tiempo juega un papel fundamental. Una parte importante de esta metodología está pensada para liberar al investigador de la tarea de especificación de los modelos dejando que los propios datos temporales de la variable a estudiar nos indiquen las características de la estructura probabilística. Estos tipos de modelos se caracterizan porque funcionan al estar basadas en series de tiempo que cumplen con un proceso estocástico estacionario con media cero, varianza constante e independente del tiempo (De arce y Mahía, 2003).

A continuación, se presentarán algunos de los conceptos clave para el análisis de series de tiempo a través de la metodología Box y Jenkins.

Proceso estacionario

Una serie de tiempo estacionaria es aquella cuyas propiedades no dependen del momento en el que se observa la serie. Por lo tanto, aquellas que cuentan con tendencia, o con estacionalidad, no lo son; ya que afectarán el valor de la serie en tiempos diferentes. Por otro lado, una serie de ruido blanco es estacionaria; no importa cuando la observe, debería verse muy similar en cualquier momento.

Una serie de tiempo con comportamiento cíclico, sin tendencia, ni estacionalidad, puede ser confundo de identificar si es estacionara o no, sin embargo, lo es. Esto se debe a que los ciclos no tienen una duración fija, por lo que antes de observar la serie no podemos estar seguros de dónde estarán los picos y valles de los ciclos. En general, una serie de tiempo estacionaria no tendrá patrones predecibles a largo plazo. Los gráficos de tiempo mostrarán que la serie es aproximadamente horizontal (aunque

es posible algún comportamiento cíclico), con varianza constante (Kwiatkowski *et al.*, 1992).

En términos generales, se dice que una serie de tiempo es estacionaria si no hay un cambio sistemático en la media (sin tendencia), en la varianza y si se han eliminado las variaciones estrictamente periódicas (Chatfield, 2003). Una forma habitual de verificar si la serie de tiempo es estacionaria es utilizando la prueba de Dickey-Fuller para estacionariedad (Fattah *et al.*, 2018).

Diferenciación

Un tipo especial de filtrado, que es particularmente útil para eliminar una tendencia en los datos, es diferenciar una serie de tiempo determinada hasta que se vuelva estacionaria. El calcular las diferencias entre observaciones consecutivas se conoce como diferenciación, que puede ayudar a estabilizar la media de una serie de tiempo al eliminar (o reducir) la tendencia y la estacionalidad (Rob J Hyndman y Athanasopoulos, 2018b).

Este método es una parte integral de la metodología propuesta por (E. O. Box *et al.*, 2008). En los datos no estacionales, la diferenciación de primer orden suele ser suficiente para lograr una estacionariedad aparente, de modo que la nueva serie $\{y_1, ..., Y_{N-i}\}$ se forma a partir de la serie original $\{x_1, ..., x_N\}$ por $y_t = x_{t+1} – x_t$ (Chatfield, 2003).

Proceso estocástico

Un proceso estocástico puede ser descrito como un fenómeno estadístico que evoluciona en el tiempo de acuerdo con leyes probabilísticas. Matemáticamente, puede definirse como una colección de variables aleatorias que están ordenadas en el tiempo y definidas en un conjunto de puntos temporales que pueden ser continuos o discretos se denotara la variable aleatoria en el tiempo t por X(t) si el tiempo es continuo (generalmente -oo <t <oo}, y por X_t, si el tiempo es discreto (generalmente t = O, +1, +2, ...) (Chatfield, 2003).

Un proceso estocástico estacionario está basado en la premisa de que está en un estado particular de equilibrio estadístico, este proceso es estrictamente estacionario si sus propiedades no son afectadas por un cambio en el tiempo de origen, es decir, que la distribución de probabilidad conjunta asociada a k observaciones realizada en cualquier intervalo de tiempo es la misma (Mota López, 2016).

Ruido blanco

Las series de tiempo que no muestran autocorrelación se denominan ruido blanco. Si $\{X_t\}$ es una secuencia de variables aleatorias no correlacionadas de una distribución fija, cada una con media cero y varianza σ^2, entonces $\{X_t\}$ se denomina ruido blanco, esto se indica mediante la notación $\{Xt\} \sim \text{WN}\ 0, \sigma^2$ (Brockwell y Davis, 2002; Wei, 2013).

A continuación, se muestra el gráfico de un proceso de ruido blanco, en donde se observa que la serie oscila alrededor del cero sin patrón de comportamiento, lo que se explica al no existir correlación entre sus observaciones.

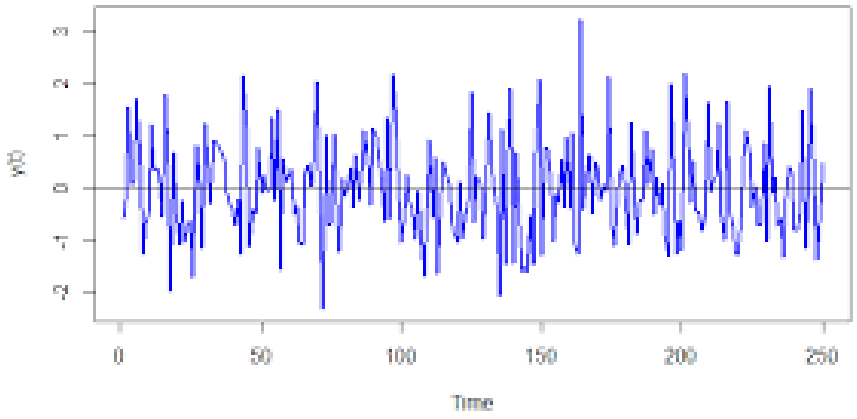

Figura 2.7. *Proceso de ruido blanco (RPubs, 2016)*

Para series de ruido blanco, esperamos que cada autocorrelación sea cercana a cero. Por supuesto, no serán exactamente iguales a cero ya que existe alguna variación aleatoria. el 95% de los picos en el Función de Autocorrelación (FAC) estén dentro de $\frac{\pm 2}{\sqrt{t}}$ donde t es la longitud de la serie de tiempo. Es común trazar estos límites en un

gráfico del FAC (las líneas discontinuas azules arriba). Si uno o más picos grandes están fuera de estos límites, o si sustancialmente más del 5% de los picos están fuera de estos límites, entonces la serie probablemente no sea ruido blanco (Rob J Hyndman y Athanasopoulos, 2018a). La prueba Ljung-box es usualmente utilizada para validar si una serie de tiempo es ruido blanco y si el modelo se ajusta correctamente.

Función de autocorrelación y autocorrelación parcial

La función de autocovarianzas de un proceso estocástico estacionario recaba toda la información sobre la estructura dinámica lineal del mismo, sin embargo, como esta depende de las unidades de medida de la variable, generalmente se utiliza la función de autocorrelación. El coeficiente de autocorrelación entre Z_t y Z_{t+k} mide el grado de asociación lineal que existe entre esas observaciones, y se define como:

$$P_z\,(k) = \frac{Cov\,(Z_t, Z_{t+k})}{\sqrt{Var\,(Z_t)}\,\sqrt{Var\,(Z_{t+k})}} = \frac{Y_z\,(k)}{Y_z(0)}$$

La gráfica del coeficiente de autocorrelación $P_z\,(k)$ en función del retardo k es llamada la función de autocorrelación $P_z\,(k)$ (FAC) del proceso (Mota López, 2016).

Debido a condiciones de estabilidad, las funciones de autocorrelación de procesos autorregresivos estacionarios de orden finito siempre son sucesiones que convergen a cero, pero no llegan a él. Esto complica el distinguir entre procesos de diferentes órdenes cuando se usa la función de autocorrelación. Para lidiar con este problema, se hace uso de la función de autocorrelación parcial la cual es la correlación que queda si el impacto posible de todas las otras variables aleatorias ha sido eliminado, de tal manera que mide la correlación entre dos variables separadas por k periodos cuando no se considera la dependencia creada por los retados intermedios existentes entre estas. La autocorrelación parcial en el análisis de series de tiempo es denotada matemáticamente como $\mathrm{Corr}(\mathrm{Zt}\,,\mathrm{Zt}+\mathrm{k}|\mathrm{Zt}+1,\ldots,\mathrm{Zt}+\mathrm{k}-1)$(Mota López, 2016).

Autorregresivo (AR)

Definimos un modelo como autorregresivo si la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores añadiéndose, como en los modelos estructurales, un término de error o entendiendo como un modelo de regresión que utiliza las dependencias entre una observación y varias observaciones retrasadas (p) (De arce y Mahía, 2003).

En un modelo de autorregresión, pronosticamos la variable de interés utilizando una combinación lineal de valores pasados de la variable (Mota López, 2016; Rob J Hyndman y Athanasopoulos, 2018c). Los modelos autorregresivos asumen que Y_t es una función lineal de los valores anteriores y viene dada por la ecuación: $y_t = \alpha_1 y_{t-1}\, \varepsilon_t$. Literalmente, cada observación consta de un componente aleatorio (choque aleatorio, ε) y una combinación lineal de las observaciones anteriores. α_1 en esta ecuación es el coeficiente de auto regresión (Fattah *et al.*, 2018).

Medias móviles (MA)

Un promedio móvil es una serie de tiempo construida tomando promedios de varios valores secuenciales de otra serie de tiempo. Este término se utiliza para describir este procedimiento porque cada promedio se calcula eliminando la observación más antigua e incluyendo la siguiente. El promedio "se mueve" a través de la serie de tiempo hasta que se calcula z_t en cada observación para la cual están disponibles todos los elementos del promedio (Hyndman, 2009).

Un modelo de medias móviles es aquel que explica el valor de una determinada variable en un período t en función de un término independiente y una sucesión de errores correspondientes a períodos precedentes, ponderados convenientemente. (De arce y Mahía, 2003).

Estos procesos representan series de tiempo de memoria corta, describen fenómenos en los que los eventos producen un efecto inmediato que dura periodos cortos de tiempo. Un proceso de medias móviles tiene la forma: $\mathrm{Z}_t = \mathrm{a}_t - \theta_1 \mathrm{a}_{t-1} - \cdots \theta_q \mathrm{a}_{q-1}$ donde $\mathrm{a}_j \sim \mathrm{WN}(0, \sigma^2), \mathrm{j} = \mathrm{t} - \mathrm{q}, \ldots, \mathrm{t}$ (Mota López, 2016).

El valor actual de un proceso de promedios móviles es una combinación lineal de la perturbación actual con una o más perturbaciones anteriores. El orden de la media móvil indica el número de períodos anteriores incluidos en el valor actual (Hyndman, 2009).

ARMA

Un proceso estacionario e invertible puede ser representado de la forma autorregresiva o de medias móviles. En la práctica, pueden suscitarse problemas en su representación al contener demasiados parámetros, aunque el modelo sea de orden finito, por lo que se procede a la unión de ambos modelos en uno solo que recibe el nombre de proceso autorregresivo de medias móviles ARMA, pues un modelo con demasiados parámetros reduce la eficiencia de la estimación. Un proceso ARMA determina a Z_T en función de su pasado hasta el retardo p, de la innovación contemporánea y el pasado de la innovación hasta el retardo q: $\mathrm{Z}_t = \varphi_1 \mathrm{Z}_{t-1} + \cdots + \varphi_p \mathrm{Z}_{t-p} + a_t + \theta_1 \mathrm{a}_{t-1} + \theta_q \mathrm{a}_{t-q}$ donde $\{Z_T\}$ es estacionaria y $\{a_t\} \sim WN(0, \sigma^2)$ (Mota López, 2016).

Las condiciones de estacionariedad del modelo ARMA(p, q) vienen impuestas por la parte autorregresiva, dado que la parte de medias móviles finita siempre es estacionaria, mientras que las condiciones de invertibilidad del modelo se comprueban a partir de la parte de medias móviles, pues la parte autorregresiva finita siempre lo será.

ARIMA

El modelo autorregresivo integrado de promedios móviles (ARIMA) es un modelo generalizado autorregresivo (ARMA) que combina el proceso autorregresivo (AR)(p) y

los procesos de media móvil (MA)(q) y con su integración, construye un modelo compuesto de la serie de tiempo. Como indican las siglas, ARIMA (p, d, q) captura los elementos clave del modelo (De arce y Mahía, 2003; Siami-Namini, Tavakoli y Siami Namin, 2018)

El modelo ARIMA se puede describir con tres parámetros (Yenidogan *et al.*, 2018):

p = número de términos autorregresivos.

d = número de diferencias.

q = número de términos de media móvil.

Los modelos ARIMA proporcionan otro enfoque para el pronóstico de series de tiempo. El suavizado exponencial y los modelos ARIMA son los dos enfoques más utilizados para la predicción de series de tiempo y proporcionan enfoques complementarios al problema. Mientras que los modelos de suavizado exponencial se basan en una descripción de la tendencia y la estacionalidad de los datos, los modelos ARIMA tienen como objetivo describir las autocorrelaciones en los datos (Rob J Hyndman y Athanasopoulos, 2018d).

Al hacer estacionaria una serie de tiempo midiendo las diferencias de observaciones en diferentes momentos (d), hace referencia a que el modelo es capaz de tratar con datos de series de tiempo no estacionarias debido a su paso de "integración" que implica diferenciar la serie de tiempo para convertirla en una estacionaria (De arce y Mahía, 2003). En este proceso, el orden de integración es el número de diferencias que hay que tomar para lograr la estacionariedad. En la práctica, (d) casi siempre toma los valores de 0, 1 y a lo máximo 2 (Mota López, 2016).

En general, si una serie $\{Z_T\}$ es integrada de orden d, se representa con el modelo: $\Phi_p(\mathrm{B})\Delta^d Z_t = \delta + \Theta_q(B)\mathrm{a}_t$ donde el polinomio autorregresivo estacionario $\Phi_p(\mathrm{B})$ y el polinomio invertible de medias móviles $\Theta_q(B)$ no tienen raíces comunes. Este modelo recibe el nombre de Modelo Autorregresivo Integrado de Medias Móviles de orden

(p,d,q) o de forma sintetizada ARIMA(p,d,q), donde p es el orden del polinomio autorregresivo estacionario, d es el orden de integración de la serie, es decir, el número de diferencias que hay que tomar a la serie para que sea estacionaria, y q es el orden del polinomio de medias móviles invertibles (Mota López, 2016).

El comportamiento de la serie temporal puede verse afectado por el efecto acumulativo de algunos procesos. Por ejemplo, el estado de las existencias se modifica constantemente por el consumo y la oferta, pero el nivel medio de las existencias depende esencialmente del efecto acumulativo de los cambios instantáneos durante el período entre inventarios. Si bien los valores de las acciones a corto plazo pueden fluctuar con grandes contingencias en torno a este valor promedio, el nivel de la serie a largo plazo se mantendrá sin cambios. Una serie de tiempo determinada por el efecto acumulativo de una actividad pertenece a la clase de procesos integrados. Incluso si el comportamiento de una serie es errático, las diferencias de una observación a la siguiente pueden ser relativamente bajas o incluso oscilar alrededor de un valor constante para un proceso observado en diferentes intervalos de tiempo. Esta estacionariedad de la serie de diferencias para un proceso integrado es una característica crucial vista desde el lado del análisis estadístico de la serie de tiempo. Los procesos integrados son el arquetipo de series no estacionarias. Un proceso integrado se define por la ecuación $Y_t = Y_{t-1} + \varepsilon_t$ donde la perturbación aleatoria ɛt es un ruido blanco (Fattah *et al.*, 2018).

El paso más importante para estimar el modelo ARIMA es identificar los valores de (p, d, q). Con base en la gráfica de tiempo de los datos, si, por ejemplo, la varianza crece con el tiempo, deberíamos usar transformaciones estabilizadoras de varianza y diferenciación. Luego, usando la función de autocorrelación (FAC) para medir la cantidad de dependencia lineal entre las observaciones en una serie de tiempo que están separadas por un retardo p, y la función de autocorrelación parcial (FACP) para determinar cuántos términos autorregresivos son necesarios , podemos identificar los valores preliminares de orden autorregresivo p, el orden de diferenciación d y el orden de media móvil q (Siami-Namini, Tavakoli y Siami Namin, 2018). Uno de los detalles

principales en este tipo de modelos es que la selección de los parámetros (p, d, q) usualmente es llevado a cabo de manera manual y depende en gran medida de la experiencia de la persona que analice los gráficos FAC y FACP, ya que como Box y Jenkins propusieron son herramientas básicas para identificar el orden del modelo ARIMA (Fattah *et al.*, 2018).

(Nau, 2020b) nos presenta una serie de reglas para estimar los parámetros (p, d, q) de un modelo ARIMA:

1. Si la serie tiene autocorrelaciones positivas con un número elevado de rezagos, probablemente necesite un orden de diferenciación superior.
2. Si la autocorrelación de retardo-1 es cero o negativa, o las autocorrelaciones son pequeñas y sin patrón, entonces la serie no necesita un orden superior de diferenciación. Si la autocorrelación de retardo-1 es -0,5 o más negativa, la serie puede estar sobre diferenciada.
3. El orden óptimo de diferenciación suele ser el orden de diferenciación en el que la desviación estándar es la más baja.
4. Un modelo sin órdenes de diferenciación asume que la serie original es estacionaria. Un modelo con un orden de diferenciación asume que la serie original tiene una tendencia promedio constante. Un modelo con dos órdenes de diferenciación total supone que la serie original tiene una tendencia variable en el tiempo.
5. Un modelo sin órdenes de diferenciación normalmente incluye un término constante (que permite un valor medio distinto de cero). Un modelo con dos órdenes de diferenciación total normalmente no incluye un término constante. En un modelo con un orden de diferenciación total, se debe incluir un término constante si la serie tiene una tendencia promedio distinta de cero.
6. Si el PACF de la serie diferenciada muestra un corte brusco y / o la autocorrelación de retardo-1 es positiva, es decir, si la serie aparece levemente "subdiferenciada", entonces considere agregar un término AR al modelo. El retraso en el que se corta el PACF es el número indicado de términos AR.

7. Si el ACF de la serie diferenciada muestra un corte brusco y / o la autocorrelación de retardo-1 es negativa, es decir, si la serie aparece ligeramente "sobre diferenciada", entonces considere agregar un término MA al modelo. El retraso en el que se corta el ACF es el número indicado de términos de MA.
8. Es posible que un término AR y un término MA cancelen los efectos del otro, por lo que si un modelo AR-MA mixto parece ajustarse a los datos, pruebe también un modelo con un término AR menos y un término MA menos.
9. Si hay una raíz unitaria en la parte AR del modelo, es decir, si la suma de los coeficientes AR es casi exactamente 1, debe reducir el número de términos AR en uno y aumentar el orden de diferenciación en una.
10. Si hay una raíz unitaria en la parte MA del modelo, es decir, si la suma de los coeficientes MA es casi exactamente 1, debe reducir el número de términos MA en uno y reducir el orden de diferenciación por una.
    a. Raíces unitarias: si una serie está sumamente subdiferenciada o sobre diferenciada, es decir, si es necesario agregar o cancelar un orden completo de diferenciación, esto a menudo se indica mediante una "raíz unitaria" en los coeficientes AR o MA estimados del modelo. Se dice que un modelo AR (1) tiene una raíz unitaria si el coeficiente AR (1) estimado es casi exactamente igual a 1. Cuando esto sucede, significa que el término AR (1) está imitando precisamente una primera diferencia, en cuyo caso debe eliminar el término AR (1) y agregar un orden de diferenciación en su lugar. En un modelo AR de orden superior, existe una raíz unitaria en la parte AR del modelo si la suma del coeficiente AR es exactamente igual a 1. En este caso, debe reducir el orden del término AR en 1 y agregar un orden de diferenciación. Una serie de tiempo con una raíz unitaria en los coeficientes AR no es estacionaria, es decir, necesita un orden superior de diferenciación.
11. Si los pronósticos a largo plazo parecen erráticos o inestables, puede haber una raíz unitaria en los coeficientes AR o MA.

Otra forma de identificar los parámetros de un modelo ARIMA es la función auto.arima del paquete “forecast” de Rstudio la cual devuelve el mejor modelo ARIMA según el valor AIC, AICc o BIC. La función realiza una búsqueda sobre el modelo posible dentro de las restricciones de orden proporcionadas (auto.arima function - RDocumentation, 2020).

AIC es una estimación de una constante más la distancia relativa entre la función de verosimilitud verdadera desconocida de los datos y la función de verosimilitud ajustada del modelo, de modo que un AIC más bajo significa que se considera que un modelo está más cerca de la verdad. BIC es una estimación de una función de la probabilidad posterior de que un modelo sea verdadero, bajo una determinada configuración bayesiana, de modo que un BIC más bajo significa que se considera que es más probable que un modelo sea el modelo verdadero. A pesar de varias sutiles diferencias teóricas, su única diferencia en la práctica es el tamaño de la sanción; BIC penaliza más la complejidad del modelo. La única forma en que deberían estar en desacuerdo es cuando AIC elige un modelo más grande que BIC (The Pennsylvania State University, 2019).

El Criterio de información de Akaike (AIC), es útil para determinar el orden de un modelo ARIMA. Se puede escribir como $AIC = -2\log(L) + 2(p+q+k+1)$ donde L es la probabilidad de los datos, k = 1 si $c \neq 0$ y k = 0 si c = 0.

Para los modelos ARIMA, el AIC corregido se puede escribir como: $AICc = AIC + \frac{2(p+q+k+1)(p+q+k+2)}{T-p-q-k-2}$ y el criterio de información bayesiano se puede escribir como: $\mathrm{BIC} = \mathrm{AIC} + [\log(\mathrm{T}) - 2](\mathrm{p} + \mathrm{q} + \mathrm{k} + 1)$. Generalmente los modelos más adecuados se obtienen minimizando el AIC, AICc o BIC, en donde AICc es el criterio de referencia principal.

Es importante señalar que estos criterios de información tienden a no ser buenas guías para seleccionar el orden apropiado de diferenciación (d) de un modelo, sino solo para seleccionar los valores de p y q. Esto se debe a que la diferenciación cambia los datos

sobre los que se calcula la probabilidad, lo que hace que los valores de AIC entre modelos con diferentes órdenes de diferenciación no sean comparables. Entonces, necesitamos usar algún otro enfoque para elegir d, y luego podemos usar el AICc para seleccionar p y q (Rob J. Hyndman y Athanasopoulos, 2018).

2.6 Visualización de datos

Con el avance de la tecnología ha existido la necesidad de mostrar cantidades masivas de datos de una manera que sea fácilmente accesible y comprensible. Las organizaciones generan datos todos los días. Como resultado, la cantidad de datos disponibles ha aumentado dramáticamente y es difícil para los usuarios visualizar, explorar y utilizar esta enorme cantidad de datos. Proporcionar una representación de datos eficaz que se originan en diferentes fuentes, permite a los responsables de la toma de decisiones ver los análisis en forma visual y les facilita su comprensión, les ayuda a descubrir patrones, comprender información y formarse una opinión (N. O. Sadiku *et al.*, 2016).

Comprender los conjuntos de datos es esencial para el proceso científico. Sin embargo, discernir la importancia de los datos mirando solo sus valores es una tarea formidable. Los métodos para presentar estadísticas resumidas visualmente incluyen tablas, cuadros y diagramas gráficos. Este tipo de herramientas son interesantes porque transmiten gráficamente una gran cantidad de información de una manera concisa que permite una rápida interpretación y comprensión de los datos. Hay muchas formas gráficas de presentar estadísticas descriptivas (Potter, 2006).

Las herramientas de visualización de datos se utilizan en la industria para respaldar la toma de decisiones y también en el ámbito académico. En la visualización de análisis de negocios son más útiles para monitorear completamente todas las actividades y también para tomar decisiones a tiempo. En la industria, es muy útil para comprender la posición de mercado de la empresa (Wright, 2008).

A continuación, se mencionaron algunos de los métodos más utilizados para la visualización de datos.

2.7 Datos faltantes

Los datos perdidos (o valores perdidos) se definen como el valor de los datos que no se almacenan para una variable en la observación de interés (Kang, 2013). El problema de la falta de datos es relativamente común en casi todas las investigaciones y puede tener un efecto significativo en las conclusiones que se pueden extraer de los mismos (Graham, 2009).

Hay tres problemas principales que pueden surgir al tratar con datos incompletos. Primero, hay una pérdida de información y, como consecuencia, una pérdida de eficiencia. En segundo lugar, existen varias complicaciones relacionadas con el manejo, cálculo y análisis de datos, debido a las irregularidades en su estructura y a la imposibilidad de utilizar software estándar. En tercer lugar, y más importante, puede haber sesgo debido a diferencias sistemáticas entre los datos observados y no observados. Un enfoque para resolver problemas de datos incompletos es la adopción de técnicas de imputación (Junninen *et al.*, 2004).

Las observaciones que faltan dificultan que los analistas realicen el análisis de datos. Los tipos de problemas que generalmente se asocian con valores perdidos son: 1) pérdida de eficiencia; 2) complicaciones en el manejo y análisis de los datos; 3) sesgo resultante de las diferencias entre datos faltantes y completos (estimaciones de sesgo) y 4) reducción del poder estadístico (estimaciones ineficientes) (Noor *et al.*, 2013)

Independientemente de los motivos por las que faltan datos, el problema no puede simplemente ignorarse y la presencia de falta de registros complica el análisis estadístico. Más específicamente, surgen tres problemas estadísticos cuando los datos se han descarriado: sesgo, error de tipo I y error de tipo II (Little *et al.*, 2016).

2.7.1 Sesgo

Cuando en los datos faltantes se muestran patrones sistemáticos, quiere decir que esa muestra seleccionada ya no es representativa de la población de la cual se extrajo, derivando en que cualquier parámetro estimado obtenido de los datos restantes ya no se generalizará a la población y cualquier conclusión será, hasta cierto punto, inválida.

Los tratamientos modernos de datos faltantes abordan el sesgo inducido, por una influencia sistemática asociada, ajustando las estimaciones de los parámetros para reflejar, lo más fielmente posible, lo que habrían sido si no hubiera datos faltantes. La medida en que esto sea posible depende de las causas de los datos faltantes y de si esas causas están representadas en los datos (Chatfield, Little y Rubin, 2002).

2.7.2 Error tipo 1

Muchos métodos tradicionales para tratar los datos faltantes (por ejemplo, cualquier técnica de imputación única) dan como resultado errores estándares que son más pequeños de lo que deberían ser, dado que los datos no están completos. El problema aquí es similar a lo que sucedería si uno recopilara datos de veinte personas, pero calculara un error estándar usando N = 200. El error estándar se basa en la suposición de que los datos contienen más precisión estadística de la que realmente tienen, por lo que el error estándar estimado es demasiado pequeño. Los errores estándar demasiado pequeños se traducen en una probabilidad superior a la nominal de encontrar un resultado significativo cuando no hay ningún efecto en la población (Little *et al.*, 2016).

2.7.3 Error tipo 2

Este tipo de error consiste en que los datos faltantes encontrados en las bases de datos significan menos precisión estadística, por lo que siempre resultará en análisis menos potentes que un conjunto de datos completo (a menos que los valores faltantes sean completamente redundantes con los valores no perdidos, Savalei y Rhemtulla, 2012). (Little *et al.*, 2016).

Dentro de los métodos que llevan a una mayor pérdida de capacidad de análisis se encuentra la eliminación por lista (Craig K., 2010). Los métodos de imputación única, como la sustitución de medias o la imputación de regresión, conducirán a mejores resultados; sin embargo, estos métodos rara vez conducen a estimaciones no sesgadas y se asocian con mayores tasas de error de tipo I (Craig K., 2010; Little *et al.*, 2016).

2.7.4 Causas o mecanismos de desaparición

Cuando se habla en el contexto de los datos faltantes a la base de datos surgen tres conceptos que se usan para categorizarlos: completamente al azar (MCAR), faltante al azar (MAR) y faltante no al azar (MNAR) (Graham, 2009) los cuales serán explicados a continuación.

2.7.5 Faltantes completamente al azar (MCAR)

Cuando los datos son MCAR, significa que los patrones de datos faltantes tienen cero asociaciones con cualquier variable incluida en el conjunto de datos o con cualquier variable no medida que esté asociada con los valores faltantes en sí mismos. Esta falta de asociación no significa que un dato no sea válido, falta por una razón: bajo MCAR, las razones son esencialmente aleatorias. En este sentido, la etiqueta de este mecanismo comunica claramente, incluso a lectores ingenuos, la razón de su ausencia. En este sentido, MCAR puede ser un proceso verdaderamente aleatorio o un proceso que tiene una asociación cero con todas las variables en una investigación (Little *et al.*, 2016).

Muchos métodos clásicos para tratar los datos faltantes asumen que todos los datos faltantes se deben al mecanismo MCAR (Chatfield, Little y Rubin, 2002). Cuando los datos son MCAR, significa que cada tratamiento de datos faltantes, excepto las técnicas de imputación única, producirá estimaciones de parámetros no sesgadas (Little *et al.*, 2016).

2.7.6 Faltantes al azar (MAR)

Cuando los datos son MAR, significa que las causas de los datos faltantes están en los datos observados. Aquí, los valores de los datos faltantes pueden considerarse como un efecto aleatorio que puede predecirse mediante otras variables del conjunto de datos. El mecanismo MAR, por lo tanto, también es un mecanismo ignorable en la medida en que los valores de los datos faltantes se pueden recuperar con precisión utilizando un modelo moderno de datos faltantes. En este sentido, tanto MCAR como MAR se consideran ignorables porque las consecuencias para las conclusiones son insignificantes; es decir, cuando se emplean herramientas de datos faltantes modernas

y basadas en principios, las estimaciones de los parámetros de cualquier modelo estadístico que se ajuste a los datos serán insesgadas cuando intervengan uno o ambos mecanismos (Little *et al.*, 2016)

2.7.7 Faltantes no al azar (MNAR)

Aquí, los datos faltan por una razón sistemática, pero el investigador no puede acceder a esta razón sistemática de la falta. Es decir, no existe un predictor de la falta que pueda utilizar un investigador. Otra forma de describir este mecanismo es que faltan datos por una razón, pero la variable que representa la razón no está en el conjunto de datos observados. Esto significa que la información que está disponible en el conjunto de datos no es suficiente para recuperar las relaciones entre variables. La circunstancia del MNAR es una situación que no se puede ignorar porque tendrá consecuencias sobre la validez de las conclusiones. La información necesaria para corregir los datos por las razones simplemente no está disponible porque falta. Debido a que los datos observados no pueden predecir los valores perdidos, si estos valores perdidos fueran imputados, las imputaciones se construirían como si el mecanismo de perdidas fuera MCAR. Sin embargo, los resultados estarían sesgados porque la razón de la ausencia es irrecuperable (Bergmeir y Benítez, 2012).

2.8 Tratamiento de datos faltantes

El proceso que remplaza valores faltantes de una base de datos a través de predicciones u otros valores observados se le conoce como imputación (Robinson y Hamann, 2011; Little *et al.*, 2016) y en la Tabla 2.2 se mencionaran los principales métodos para llevar a cabo esta tarea.

Método	Descripción	Efecto	Autor
Eliminación por lista	Consiste en la eliminación de todos los registros en los cuales faltan valores	Los errores estándar, intervalos de confianza y p valores son mayores. La muestra deja de ser representativa de la	(Little *et al.*, 2016)

		población. Estimaciones de parámetros sesgadas.	
Eliminación por pares	Es un método en el que se incluyen datos para una variable pertinente a una evaluación específica, incluso si faltan valores para el mismo registro en otras variables.	Cada estadístico se basa en un tamaño de muestra diferente, lo que puede resultar en una matriz de correlación imposible, lo que puede causar serios problemas de estimación.	(Marsh, 1998; Little *et al.*, 2016)
Sustitución por promedio	Como su nombre lo indica aquellos valores faltantes dentro de una variable son sustituidos con la media de esta.	Los valores imputados tienen una varianza cero porque todos toman el mismo valor. Como tal, las estimaciones de la varianza de la variable y las relaciones con otras variables (por ejemplo, covarianzas, correlaciones, coeficientes de regresión) se subestiman.	(Little *et al.*, 2016)
imputación por regresión	Cada valor faltante se reemplaza con su valor predicho basado en una ecuación de regresión que utiliza todas las	Se subestimarán las varianzas (porque falta la varianza del error) y se sobrestimarán las relaciones de las	(Little *et al.*, 2016)

	demás variables como predictores. La información de las variables completas se utiliza para completar las variables incompletas.	variables (correlaciones, covarianzas, regresiones).	
Imputación de regresión estocástica	Al igual que con la imputación de regresión, cada valor faltante se reemplaza con su valor predicho, pero tiene un término de error residual distribuido normalmente agregado a la puntuación predicha imputada. La varianza residual agregada agrega variabilidad a la variable imputada y agrega ruido a las relaciones estimadas entre variables.	La información perdida debido a la falta de datos no se contabiliza, por lo que los errores estándar serán demasiado pequeños. Se obtienen tasas de error de tipo I infladas (en todos los mecanismos).	(Little *et al.*, 2016)
Imputación de expectativa-maximización	El algoritmo EM produce estimaciones de máxima verosimilitud (ML) de la matriz de covarianza a partir de datos incompletos. El algoritmo se repite en dos pasos. En	La incertidumbre en los datos no se tiene en cuenta en un conjunto de datos de imputación única. Las tasas de error de tipo I se elevarán.	(Craig K., 2010; Kang, 2013; Little *et al.*, 2016)

	el paso de expectativa (E), los valores faltantes se completan con sus valores esperados después de condicionar los datos observados. El siguiente paso de maximización (M) aplica la estimación ML de datos completa a los datos imputados del paso (E) para derivar estimaciones actualizadas de las covarianzas y las medias. Estas estimaciones actualizadas se utilizan en el siguiente paso (E) para construir las ecuaciones de regresión utilizadas para completar los datos faltantes.		
Imputación Hot deck	La imputación hot deck funciona al encontrar participantes que coinciden con el caso con datos faltantes sobre otras variables.	Dependiendo de cómo se realice realmente la imputación, este método puede conservar o no las correlaciones entre variables. Este método no es paramétrico y menos	(Jerez *et al.*, 2010)

		sensible a la especificación incorrecta del modelo. No es Bueno para muestras pequeñas.	
Última observación llevada a cabo	Los valores faltantes en los datos son remplazados con la última observación anterior.	Las estimaciones de los parámetros estarán sesgadas en direcciones impredecibles, según las características particulares de los datos. La magnitud del sesgo no es predecible. Las tasas de error de tipo I aumentan al no tener en cuenta el efecto de la información faltante.	(Little *et al.*, 2016)
Imputación múltiple	En este método se imputan varios valores para cada punto de datos que falta. La información perdida debido a la falta de datos en la estimación de los parámetros del modelo se cuantifica utilizando la variabilidad en las estimaciones de los parámetros entre	Este algoritmo realiza las imputaciones de los datos de manera imparcial y eficiente, el error estándar es insesgado por lo que da como resultado errores de tipo 1 precisos.	

	conjuntos de datos imputados.		
Máxima probabilidad de información completa	Estima los parámetros del modelo y los errores estándar directamente utilizando todos los datos disponibles y tiene como objetivo identificar los parámetros de población que tienen más probabilidades de haber generado los datos observados (es decir, maximizar la probabilidad de observar datos de muestra). FIML es una solución de datos faltantes basada en modelos, lo que significa que su precisión depende de la precisión del modelo que se estima.	Cuando se cumplen los supuestos, FIML conduce a estimaciones asintóticamente insesgadas y eficientes	
Interpolación lineal	Este método consiste en conectar dos puntos de datos con una línea recta,	Para supone que tienen una dependencia lineal de	(Noor *et al.*, 2015)

	para completar los datos faltantes.	un parámetro de las distribuciones.	

Tabla 2.2 *Métodos de imputación para datos faltantes (Elaboración propia)*

2.9 Estudios previos

El trabajo presentado por (Jawadi y Ftiti, 2019) intenta dar respuesta a que pasaría si el precio del petróleo de Arabia Saudita se desploma y si su actividad económica se diversificara con el fin de ver cuál sería el impacto en su economía, debido a que esta se encuentra centrada principalmente en este recurso, haciéndola altamente dependiente, esto es debido a que el precio del petróleo en la última década ha tenido un comportamiento altamente volátil principalmente por dos factores, la economía mundial y la demanda del petróleo.

En su trabajo realiza un análisis de series de tiempo con la suposición de que la relación del precio del petróleo y la economía es no lineal y variable con el tiempo por lo que proponen un modelo de umbral de encendido / apagado que puede proporcionar una medida diferente de la reacción del PIB saudí al cambio del precio del petróleo haciendo uso de técnicas como los modelos autorregresivos de umbral denominados TAR propuestos por (Tong y Lim, 1980), análisis de rotura estructural y test de linealidad.

Su primer hallazgo fue confirmar la dependencia de la economía de Arabia Saudita del sector del petróleo, pero también se mostró que el precio del petróleo exhibe un efecto de umbral que varía según el régimen o el estado del mercado. La identificación de estos estados es particularmente útil para proteger la economía contra un mayor colapso del petróleo. En segundo lugar, se mostró el beneficio de la transformación económica a través de la opción de diversificación, ya que solo la inversión en acciones puede impulsar la economía de ese país.

Otro caso de estudio previo corresponde al trabajo presentado por (Puchalsky *et al.*, 2018) al evaluar el rendimiento de las redes neuronales Wavelet (WNN) combinado

con cinco técnicas de optimización metaheurística para obtener la mejor predicción de series de tiempo al considerar dos estudios de caso en el sector agroindustrial. El primero adopta el precio del saco de soja y el segundo aborda el problema de la demanda de un grupo diferenciado de productos de una empresa alimentaria, donde las tendencias no lineales son la principal característica en ambas series de tiempo.

En este trabajo WNN fue seleccionado debido a la pequeña dimensión de insumos asociada a ambos estudios de caso, principalmente con el caso de demanda de productos. Además, el éxito de los enfoques WNN en el tratamiento de la tendencia y la estacionalidad en series de tiempo se confirmó en los trabajos de (Shahabi, Tian y Zhao, 2000; Martínez y Gilabert, 2009).

Las técnicas de optimización que fuera utilizadas son: evolución diferencial, colonia de abejas artificial, optimización de enjambre de gusanos luminosos, algoritmo de búsqueda gravitacional y algoritmo competitivo imperialista. Éstas se evaluaron considerando pronósticos a corto y largo plazo, y se consideró un horizonte de predicción de 30 días para el caso del precio del saco de soja, mientras que para el caso de demanda de productos se consideraron 12 meses a futuro.

Respecto a la evaluación de los métodos de optimización, para verificar cuál era el más eficiente en la predicción de ambas series de tiempo, se adoptó el Error Cuadrático Medio (MSE) y el Error Porcentual Absoluto Medio (MAPE) como criterios de aproximación y para analizar el modelo propuesto, los resultados se compararon con el método de Máquina de Aprendizaje Extremo (ELM) y el algoritmo de entrenamiento de propagación hacia atrás (BP) clásico asociado a un WNN.

En ambos casos analizados en este estudio, ELM superó a la mayoría de los métodos durante los procedimientos de validación. Sin embargo, durante las pruebas a corto y largo plazo, los métodos de optimización metaheurística utilizados para el entrenamiento superaron los resultados de ELM en casi todos los casos, lo que demuestra la relevancia de este enfoque.

3. METODOLOGÍA

En este capítulo se presenta la metodología que guio el desarrollo de la investigación, la cual fue del tipo cuantitativa con alcance descriptivo, fue elaborada para estudiar el comportamiento de los precios de la uva de mesa sonorense en el mercado de Estados Unidos a través de la utilización estrategias de minería de datos, específicamente la modelación de series de tiempo con el propósito de identificar y delimitar la ventana de mercado para este producto.

La presente propuesta es una adaptación de los trabajos realizados por Pete *et al.*, (2000) y Fayyad, Piatetsky-Shapiro y Smyth (1996). Por ello se propone una metodología de cinco etapas de las cuales cada una incluye una serie de tareas a desarrollar como se muestra en la Figura 3.1. en la página 54.

A continuación, se describirán a detalle cada una de las cuatro etapas y tareas que conforman esta metodología propuesta, con el fin de tener un mejor entendimiento de las actividades que se deben de realizar, y que pueda ser replicado en un momento dado en otro contexto. Es importante recalcar que el proceso a seguir es iterativo pudiendo ir de una fase a otra en caso de ser necesario con el fin de mejorar los resultados obtenidos.

Figura 3.1. *Modelo de la metodología propuesta para el estudio del comportamiento de los precios a través de estrategias de minería de datos (elaboración propia adaptado de los trabajos de Pete et al., (2000) y Fayyad, Piatetsky-Shapiro y Smyth (1996)).*

3.1 Comprensión del negocio

La primera etapa consiste en el acercamiento con la problemática a tratar, es por ello por lo que se debe de realizar una serie de reuniones con las personas interesadas quienes permitirán el acceso a la información necesaria para identificar, de manera clara y concisa la situación actual, además de clarificar de cuáles hechos surge el interés del desarrollo del proyecto. El resultado de esta etapa consiste en un documento estructurado donde se podrá identificar claramente la problemática, los objetivos del proyecto, requerimientos por parte del cliente, limitaciones, restricciones y el plan de trabajo.

3.1.1 Antecedentes del proyecto

Esta tarea es probablemente una de las más importantes de todo el proceso, debido a que a partir de esta se derivan todas las demás tareas, consiste en llevar a cabo una serie de reuniones con la parte interesada o el cliente en donde se tendrán que identificar y entender el problema que generó el interés en el desarrollo del proyecto, así como sus antecedentes. Para la realización de esta tarea se utilizará el formato presentado en la Tabla 3.1, el cual permite llevar un registro de las minutas de cada una de las reuniones realizadas, donde se registrará la fecha, asistentes y comentarios claves, lo que servirá de apoyo para la siguiente tarea.

Fecha	Motivo de la reunión	Asistentes	Comentarios

Tabla 3.1. *Formato de minuta (Elaboración propia).*

3.1.2 Definición de objetivos generales y específicos

A partir de los registros de las minutas, será necesario transformar lo discutido verbalmente en una serie de objetivos generales bien definidos los cuales serán desglosados en una serie de objetivos específicos y sus tareas a fin de lograr cumplir con el proyecto, cabe recalcar que este paso tiene que estar avalado por las partes interesadas.

3.1.3 Requerimientos del cliente

Esta tarea consiste en definir cuáles son los requerimientos mínimos de calidad, seguridad, confidencialidad, comprensión, fidelidad, tiempo de entrega del proyecto, entre otros por parte del cliente para considerar el proyecto como aceptable.

3.1.4 Alcances y límites

Una vez definidos los requerimientos del cliente se procederá a definir, en conjunto con la parte interesada, cuáles son los alcances y límites en el proyecto, un ejemplo sería que solo se trabajara con datos de cierto año o producto.

3.1.5 Plan de trabajo

Para esta tarea, se hará uso de la herramienta del “Diagrama de Gantt” para la realización del plan de trabajo del proyecto a realizar, y consiste en transformar los objetivos generales y específicos en acciones concretas, asignando un responsable y definiendo tiempos de realización, así como plantear los supuestos acerca de cómo serán realizados, por ejemplo, que materiales y técnicas serán aplicadas en caso de ser posible. El formato por utilizar para esta tarea será el que se presenta en la Figura 3.2. en la página 57.

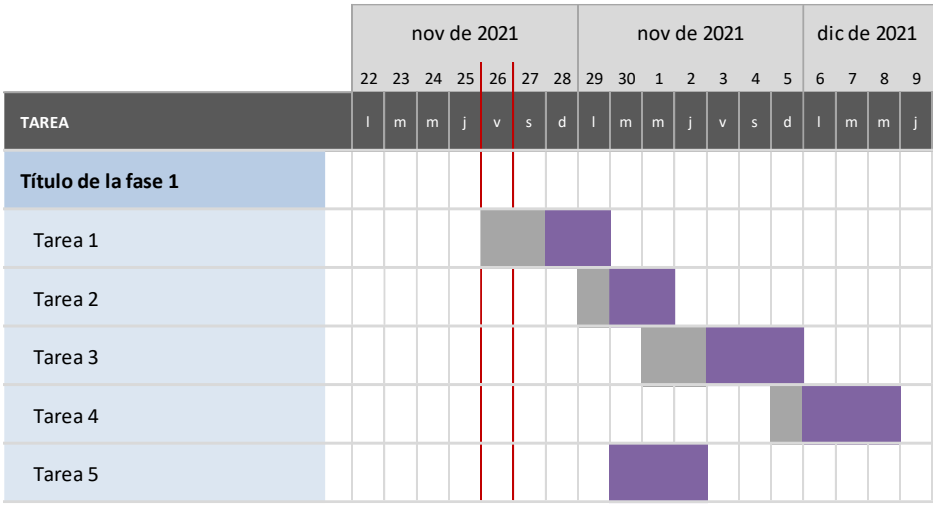

***Figura 3.2**. Diagrama de Gantt (Elaboración propia)*

3.2 Comprensión de los datos

Esta etapa consiste en la conformación de la base de datos, así como la realización de análisis descriptivos y exploratorios de los datos con el fin de obtener una mayor comprensión de estos, para posteriormente validarlos.

3.2.1 Identificación y validación de las fuentes de datos

La primera tarea de esta etapa consiste en identificar las posibles fuentes de datos e información a partir de las cuales posiblemente se recolectarán los datos para posteriormente validar cuales de estas fuentes son oficiales o de confianza,

3.2.2 Recolección de datos iniciales

Esta tarea consiste en descargar archivos que contengan los datos de interés para el estudio en el formato necesario de la fuente de información seleccionada.

3.2.3 Descripción de los datos

Consiste en describir de manera general el contenido de las bases de datos que se descargaron, cuantos archivos son, en que formato están almacenados, el tipo de información que contienen, tipo de variable y cantidad de registros y cualquier otra información que se considere relevante, haciendo uso del formato que se ve en la Tabla 3.2 en la página 58.

Tipo de variable	Tipo de dato	Descripción	Tipo de variable	Tipo de dato	Descripción
X1			X4		
X2			X5		
X3			Xn		

Tabla 3.2. *Formato para descripción de variables (Elaboración propia).*

3.2.4 Análisis exploratorio

Esta etapa consiste en realizar un análisis exploratorio de los datos haciendo uso de estadística descriptiva como la media, moda, promedio, entre otros y hacer uso de herramientas de visualización de datos como gráficas de frecuencia, de dispersión o de caja y bigote a fin de entender la manera como se comportan e identificar valores erróneos, en esta tarea se utilizará el formato visto en la Tabla 3.3. La herramienta para realizar todas estas actividades y análisis quedará a consideración del analista quien deberá de basarse en su dominio de la herramienta y las capacidades de esta, un ejemplo es el uso de Excel, Orange, R o Python.

Nombre de variable	Registros	Dato en blanco	Promedio	Desv. Est.	Min	Max
x1						
x2						
x3						
xn						

Tabla 3.3. *Formato para análisis exploratorio (Elaboración propia).*

3.2.5 Validación de los datos

En esta última tarea se verificará la calidad e integridad de los datos identificando valores faltantes, erróneos. El por qué estos datos se encuentran de esta manera y que tan comunes son, corresponden a esta última tarea, que se considera de suma importancia para la etapa de preprocesamiento y limpieza ya que sirve como base para

identificar los valores o problemas a tratar. A manera de resumen se utilizará el formato de la Tabla 3.4.

Problema	Posible causa	Posible solución
Problema 1		
Problema 2		
Problema n		

Tabla 3.4. *Formato validación de datos (Elaboración propia).*

3.3 Preparación de los datos

Consiste en la preparación, limpieza y reformateo de la base de datos con el fin de implementar los algoritmos para el modelado y la evaluación del desempeño del modelo creado.

3.3.1 Selección de variables de interés

Antes de comenzar a trabajar con los datos y dejarlos listos para la etapa de modelado, es necesario seleccionar del conjunto de variables que pudieran ser de interés o que sirvan para cumplir con el objetivo del proyecto. En la realización de esta tarea es de suma importancia el conocimiento previo del tema debido a que será de gran ayuda al momento de elegir aquellas variables de interés.

A partir de las variables seleccionadas se procederá a crear una nueva base de datos lo que ayudará con el proceso de visualización de estos y de su análisis, así como a determinar la capacidad de procesamiento requerida.

3.3.2 Limpieza de datos

En esta tarea será llevará a cabo la limpieza de los datos incompletos o erróneos identificados en la tarea 3.2.6; existen diversas actividades que se pueden realizar, desde eliminar todos aquellos valores incompletos o erróneos, sin embargo, dependerá en gran medida de la cantidad de registros con los que se cuente, se debe

tener en cuenta que con su eliminación se pudiera eliminar la mayoría de los registros dejando poca información para el análisis.

3.3.3 Estandarización de datos

Esta tarea consiste en unificar los tipos de valores, por ejemplo, si en una base de datos se encuentra un formato de fecha como día, mes y año, y también se encuentran dentro de esa misma columna formatos como mes, día y año, se debe aplicar un solo formato; otro ejemplo, si en la columna de pesos se cuenta con medidas de kilogramos y libras se tendría que definir solo un tipo de unidad.

3.3.4 Reestructuración y formateo de base de datos

Esta tarea consiste en generar nuevas columnas de valores a partir de otras columnas de la misma base de datos, un ejemplo es calcular el precio por kilo de algún producto, partiendo de la suposición de que en la base de datos solo se tiene información del total de kilos y su precio de venta entonces se procedería a realizar una división entre estos valores. Otra de las tareas consiste en dar el formato adecuado a los datos ya sea una fecha o un identificador.

3.4 Modelado

Esta etapa consiste en llevar a cabo un análisis de los diferentes algoritmos que ayuden al cumplimiento del objetivo; una vez seleccionado el modelo, se deberá de generar la prueba de diseño, dividiendo el conjunto de entrenamiento y prueba con el fin de identificar y evaluar el modelo.

3.4.1 Análisis y selección de modelos

Consiste en analizar los diferentes tipos de modelos y sus algoritmos a fin de seleccionar el que mejor se ajuste a las necesidades del proyecto y a los datos disponibles para trabajar. Debido a que algunos requieren cierto tipo o estructura de datos, es necesario iterar entre las etapas anteriores para adecuar el formato al algoritmo de extracción de conocimiento; un ejemplo de esto sería el análisis de series

de tiempo que requieren de una columna que indique las fechas de ocurrencia de los registros.

Una vez seleccionado que tipo de modelo o técnica se utilizará, se deberá hacer una tabla comparativa de los diversos algoritmos siguiendo el formato de la Tabla 3.5 para identificar cuáles se ajustan a nuestros datos, intereses y capacidades.

Nombre del algoritmo	Características	Requisitos

Tabla 3.5. *Tabla comparativa de algoritmos de minería de datos (Elaboración propia)*

3.4.2 Generar diseño de prueba

Esta tarea consiste en generar un procedimiento o mecanismo para probar la calidad y validez del modelo, así como los resultados obtenidos a través un plan para entrenar y probar el modelo, para posteriormente alcanzar la calibración de este. Una actividad importante es la segmentación del conjunto de datos en un parte para entrenamiento y otra para prueba del modelo.

3.4.3 Construcción del modelo

Una vez definido los datos utilizados para el entrenamiento del modelo, se procede a implementar el o los algoritmos seleccionados en la etapa anterior a través de la herramienta o técnica que sea adecuada para el proyecto y que sean del dominio para la persona que se encuentra implementando el modelo.

3.4.4 Evaluación del desempeño

Esta tarea consiste en evaluar el desempeño obtenido de los modelos generados según el conocimiento de dominio del tema, los criterios de éxito del proyecto y con base al diseño de prueba deseado; para esta tarea, también se utilizará el conjunto de datos de evaluación para calcular el error, según las métricas de evaluación de resultados, que dependerán en gran medida según el tipo de modelo y algoritmo que

haya sido utilizado. El modelo general de la evaluación de desempeño se puede ver en la Figura 3.3.

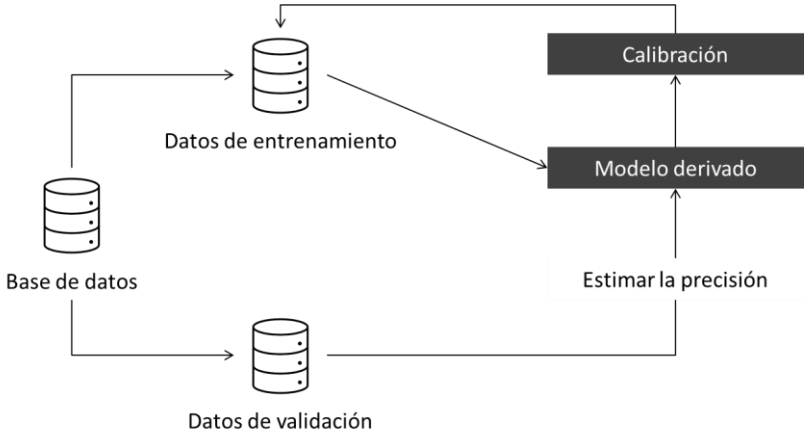

Figura 3.3. *estimación de la precisión (Han y Kamber, 2006)*

3.5 Evaluación de resultados

Esta etapa, consiste en la elaboración de un documento, con los resultados y conclusiones obtenidas, para que sea evaluado por las personas interesadas en la realización del proyecto, y en caso de representar un conocimiento valioso para la organización o institución, sea integrado dentro de su proceso productivo o toma de decisiones. Es importante recalcar que es un proceso interactivo en el cual se puede volver a etapas anteriores a fin de mejorar los resultados obtenidos.

3.5.1 Reporte de resultados

Esta tarea consiste en la elaboración de un reporte escrito en el cual se plasmen los resultados obtenidos y la forma en la que se llegaron a éstos; en el documento se debe de argumentar cómo los resultados cumplen con los objetivos y requerimientos de las partes interesadas o del cliente. También, se deberá de contar con la aceptación por parte de ellos validando el trabajo realizado. En esta tarea se debe utilizar un formato que contenga los siguientes puntos:

- Título y fecha de entrega.
- Plan de trabajo.

- Objetivos generales y específicos.
- Métodos, algoritmos y técnicas utilizadas.
- Resultados obtenidos.
- Conclusiones.
- Recomendaciones.
- Firma del responsable.

3.5.2 Próximos pasos

Una vez terminado toda la documentación del proyecto y entregada al cliente o la parte interesada es necesario determinar las acciones futuras como el volver a realizar algunas de las etapas a fin de mejorar el resultado ya obtenido o comprobar algún otro supuesto.

4. IMPLEMENTACIÓN

4.1 Comprensión del negocio

En esta etapa se detallará la problemática a tratar y el objetivo a cumplir para satisfacer las necesidades del cliente, en este caso el CIAD, y se plasmará el plan de trabajo y los límites.

4.1.1 Antecedentes del proyecto

En esta etapa, se llevaron a cabo una serie de reuniones con el encargado del proyecto por parte del CIAD en las cuales se hablaron acerca de los intereses que tienen para la realización del proyecto; a lo largo de estas reuniones también plantearon los objetivos, tiempos y el plan de trabajo de todo el proyecto. En la Tabla 4.1 se muestran las reuniones llevadas a cabo, asistentes y el tema.

Fecha	Motivo de la reunión	Asistentes	Comentarios
sábado, 22 de agosto de 2020	Presentación del proyecto	Dr. Luis Felipe romero Dessens Dr. Juan Martin Preciado Ing. Vicente Solis Sandoval	Se presentaron a los interesados y los objetivos del proyecto.
martes, 15 de septiembre de 2020	Conocer los antecedentes del proyecto	Dr. Juan Martin Preciado Ing. Vicente Solis Sandoval	Se conoció lo que hace CIAD y en de donde nace el objetivo del proyecto.
domingo, 20 de septiembre de 2020	Definición de objetivos generales y específicos	Dr. Juan Martin Preciado Ing. Vicente Solis Sandoval	Se plasmó en un documento.
viernes, 2 de octubre de 2020	Determinación del plan de trabajo	Dr. Juan Martin Preciado Ing. Vicente Solis Sandoval	Se plasmó en un documento.

Tabla 4.1 *Bitácora de reuniones para la comprensión del negocio (elaboración propia).*

4.1.2 Definición de objetivos generales y específicos

Con base a las reuniones que se tuvieron con el encargado del proyecto por parte del CIAD, se llegó al siguiente objetivo general: desarrollar un modelo explicativo de series de tiempo a través de técnicas de minería de datos, que permita identificar el comportamiento de los precios de venta y ventanas de mercado de la uva de mesa sonorense en Estados Unidos, y para llevar a cabo futuras investigaciones, así como estudiar la sostenibilidad del sistema de la CPUMS.

Dentro de los objetivos específicos se encuentran los siguientes:

- Conformar una base de datos con los precios de venta de la uva de mesa sonorense en el mercado estadounidense a partir de información recopilada del Departamento de Agricultura de los Estados Unidos (USDA – Agricultural Marketing Service) del año 1998 al 2020.
- Desarrollar un modelo de series de tiempo que permita identificar los parámetros que mejor expliquen el comportamiento de los precios y de la ventana de mercado.
- Determinar el nivel de error del modelo desarrollado.
- Caracterizar el comportamiento de los precios de venta y de la ventana de mercado de la uva de mesa sonorense en el mercado estadounidense.

4.1.3 Requerimiento del cliente

Dentro de los requerimientos del cliente se encuentra que los datos que serán utilizados para el desarrollo del modelo deberán de provenir únicamente de fuentes oficiales; respecto a la calidad del proyecto, debe de ser un trabajo de investigación bien sustentado académicamente con el propósito de alcanzar su replicabilidad en proyectos futuros dentro de la organización. Por último, el tiempo de entrega deberá de ser según lo establecido en el plan de trabajo.

4.1.3 Alcance y limites

Se acordó que el proyecto solo comprendería los precios registrados de las exportaciones de Sonora a Estados unido de Uva de Mesa del año 2000 al 2020.

4.1.4 Plan de trabajo

Despues de llevar a cabo varias juntas con el CIAD, se acordó el siguiente plan de trabajo para la realización del proyecto, y plazos establecidos dentro de su organización, mismo que se muestra en la Figura 4.1.

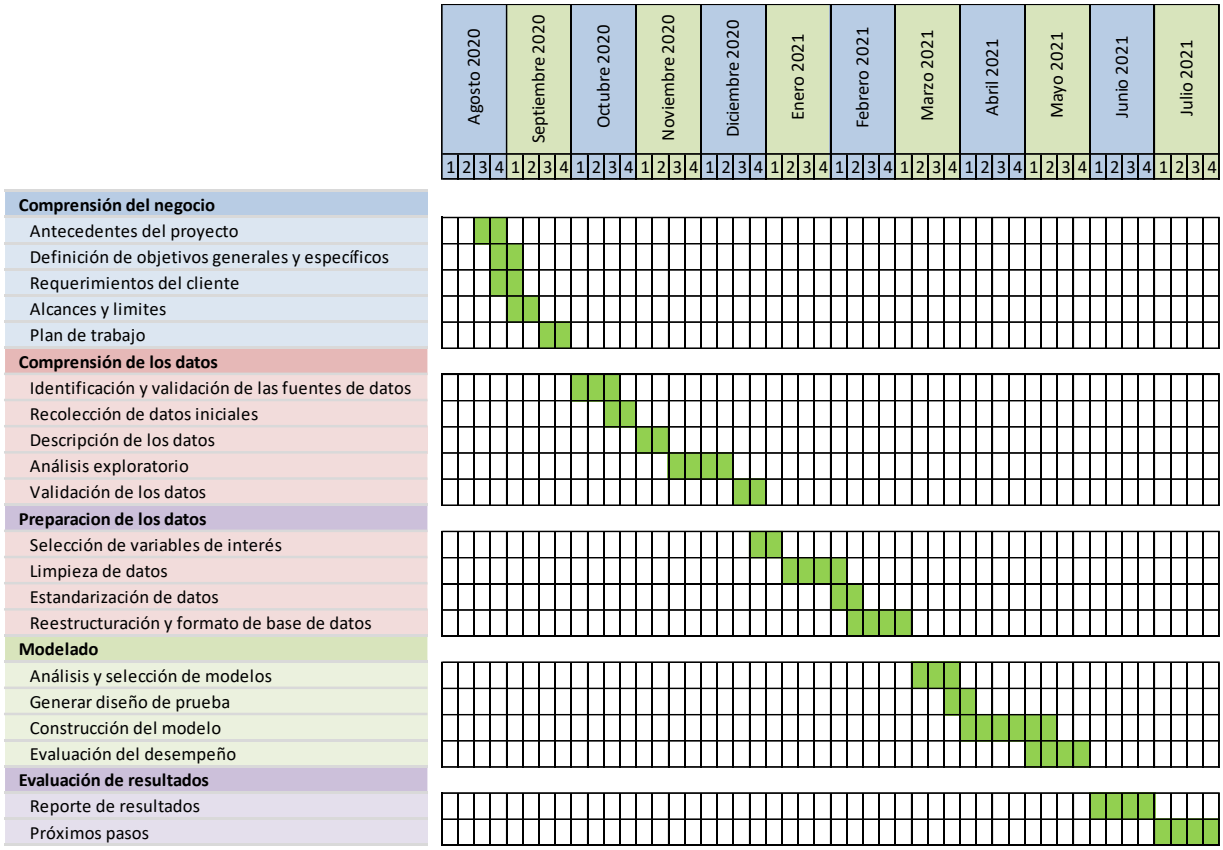

Figura 4.1 *Plan de trabajo para el proyecto (Elaboración propia)*

4.2 Comprensión de los datos

En esta etapa se identificarán las posibles fuentes de datos que brinden la información necesaria para la realización del estudio, se estudiarán los datos recopilados en su comportamiento y se identificaran problemas de calidad para tratarlos en una etapa posterior.

4.2.1 Identificación y validación de las fuentes de datos

Es la búsqueda de una fuente de datos fiable y que tuviera la información acerca del país, estado de origen, destino, el peso y el precio de venta de la uva de mesa en

específico. Se encontró que el Departamento de Agricultura de Estados Unidos (USDA, por sus siglas en inglés) cuenta con un portal para la generación de reportes de los movimientos relacionados con la agricultura de ese país, llamado “Agricultural Marketing Service” (https://www.ams.usda.gov/market-news/custom-reports) dentro del cual se cuenta con la opción de descargar diversas bases de datos de los movimientos y transacciones que ocurren en diferentes tipos de formato como HTML, PDF y XLSX.

También se encontró un portal de información llamado “Agronometrics” (https://www.agronometrics.com/) el cual se encuentra conectado directamente a las bases de datos comerciales de la USDA y que recibe actualizaciones semanales de la información y de igual manera brinda la opción de descargar reportes personalizados en diversos formatos.

4.2.2 Recolección de datos iniciales

En esta tarea se procedió a realizar la descarga de las base de datos de ambos sistemas para compararlas y definir cuál de las dos se utilizará en el estudio; cabe mencionar que en ambos casos se filtró la información a solo el producto de “Uva de mesa” y a que la información se encontrara de manera “Diaria”, de esta manera desde el portal “Agricultural Marketing Service”, los archivos se descargaron de manera trimestral debido a que el sistema de consulta cuenta con una limitante referente al lapso que puedes descargar, resultando en 92 archivos de tipo Excel (.xls) que conforman una serie de tiempo con registros diarios que van del “01 de enero de 1998” hasta el “31 de diciembre del 2020”. Por otro lado, de la base de datos de “Agronometrics” se descargaron dos archivos con registros diarios para el precio por libra y otro referente al volumen de comercialización desde el “01 de enero del 2000” hasta el “31 de diciembre del 2020”.

4.2.3 Descripción de los datos

En esta tarea se describirán ambas bases de datos recolectadas y se analizarán la calidad de cada una de ellas a fin de ver cuál de estas es más adecuada para el estudio.

La base de datos descargada directamente de la USDA contaba con 92 archivos en total de tipo Excel con extensión .xls; todos los archivos comparten el mismo tipo de variables y columnas entre ellas, así como en total se cuenta con 912,204 registros (filas) y 29 variables (columnas), las cuales se describen en la Tabla 4.2 con una descripción de su contenido y el tipo de datos de la variable.

Tipo de variable	Tipo de dato	Descripción	Tipo de variable	Tipo de dato	Descripción
Commodity Name	Texto	Nombre del producto comercializado, en este caso todos los registros de uva de mesa (GRAPE en el sistema).	Color	Texto	Color de la uva de mesa.
City Name	Texto	Corresponde a la ciudad de la terminal o central de abasto.	Environment	Texto	Ambiente en el que fue cultivada la uva de mesa (No aplica).
Type	Texto	Indica si la uva de mesa es orgánica o no.	Unit of Sale	Texto	Si fue vendida por libras o kilogramos.
Package	Texto	Peso y presentación de la uva de mesa comercializada.	Quality	Texto	Nivel de calidad de la uva de mesa.
Variety	Texto	Variedad de la uva de mesa.	Condition	Texto	Nivel de condición de la uva de mesa.
Sub Variety	Texto	Subvariedad de la uva de mesa.	Appearance	Texto	Nivel de apariencia externa de la uva de mesa.
Grade	Texto	Determina si la uva de mesa es producida en Estados Unidos o no.	Storage	Texto	Almacenamiento o factores externos que afectan al producto (No aplica) .
Date	Texto	Fecha de cotización.	Crop	Texto	Discrimina si el producto es nuevo.
Low Price	Numero	Cotización mínima alcanzada.	Repack	Texto	Si fue reempaquetada o no.
High Price	Numero	Cotización máxima alcanzada.	Trans Mode	Texto	Modo de transporte del país de origen a Estados Unidos.
Mostly Low	Numero	Cotización generalmente mínima.	Offerings	Texto	Comentarios acerca de la oferta que se hizo por la mercancía.
Mostly High	Numero	Cotización generalmente máxima.	Market Tone	Texto	Estado en el que se encontraba el mercado al momento de realizar la transacción.
Origin	Texto	País de origen de la mercancía.	Price Comment	Texto	Comentarios acerca del precio de la mercancía.
Origin District	Texto	Distrito de origen de la mercancía .	Comments	Texto	Comentarios generales.
Item Size	Texto	Tamaño de la uva de mesa.			

Tabla 4.2. *Descripción de la base de datos de la USDA (elaboración propia).*

La base de datos que fue descargada de Agronometrics contaba con la información referente a los volúmenes y los precios, sin embargo, la estructura en la que se encontraba era diferente, ya que sólo contaba con las fechas de la transacción, el valor o volumen y una columna por cada país que haya tenido algun registro, además, en el

caso de los precios, se contaba con una columna haciendo referencia al promedio y para el volumen una columna con la suma total del día.

En el caso de la base de datos de precio se contaba con un total de 9 variables (columnas), para el volumen 18 (columnas) y, en ambos casos se tendrían un total de 7671 registros (filas).

4.2.4 Análisis exploratorio

En todas las etapas que tienen que ver con el trabajo de las bases de datos se utilizaron los programas informativos de Excel 365, Jupyter Notebook en Python 3 y Rstudio 2021.09.0+351 en el lenguaje de programación R 3.0.1+.

Se comenzó por analizar la base de datos de la USDA, el primer paso que se realizó para la descripción de los datos fue abrir uno de los archivos trimestrales para verificar de manera visual si los datos descargados se encontraban en la manera correcta, sin embargo, se encontró un error al momento de abrirlo en Excel como se puede ver en la Figura 4.2.

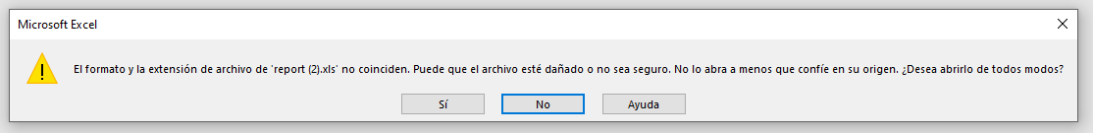

Figura 4.2. *Error en la apertura de la base de datos en Excel (elaboración propia).*

Al intentar abrirlo desde Jupyter Notebook de igual manera lanzaba un mensaje de error el cual era el siguiente "XLRDError: Unsupported format, or corrupt file: Expected BOF record; found b'<html><b'" , se realizó una verificación en el tipo de archivo que se había descargado y resultó que el problema residía en que el contenido del archivo descargado estaba en formato "HTML" y la extensión del archivo era .xls que corresponde a un archivo de Excel como se puede ver en la Figura 4.3. en la página 70.

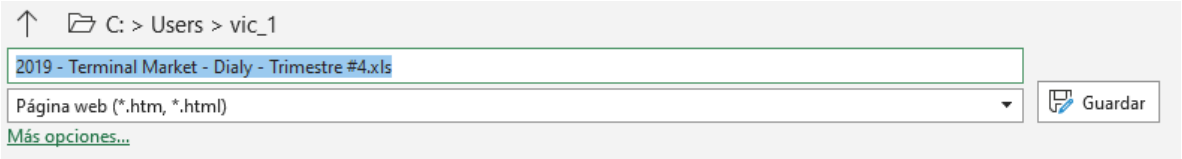

Figura 4.3. *Error en el formato y extensión de la base de datos (elaboración propia).*

Lo que se hizo para corregir este error fue abrir cada uno de los archivos en Excel, aunque diera error, y guardarlos cambiando el tipo de archivo a un “Libro de Excel” con extensión .xlsx lo que permitiría manipular de manera correcta y sin errores los datos; sin embargo, antes de continuar con la descripción de los datos era necesario conformar un archivo único que facilitaría su análisis, para ello se procedió a realizar la carga de los archivos en Jupyter Notebook donde se haría uso de los paquetes “pandas” y “os” en Python para combinar los 92 archivos en una sola base de datos.

Explorando el contenido del archivo único se identificó que existían ciertos problemas de calidad en la base de datos como:

1. Las fechas se encontraba en formatos diferentes haciendo que el trabajarlo como estampa de tiempo no sea posible.
2. Gran cantidad de las variables disponibles se encontraban sin registros o con muy pocos.
3. La variable de low y high price contaban con datos extremos superiores (495) e inferiores (.25) lo que generaba ruido en la base de datos.
4. Existían más de 1 valor por día debido a que existen diversas variedades de uva, punto de llegada y salida por país.
5. La variable que hace referencia al volumen de la transacción en peso se contaba en formato de texto debido a que dentro del mismo registro se menciona su medida de peso y el empaque.
6. No existe una unidad de medida unitaria como precio por libra / kilo.

Una vez identificados estos problemas de calidad de los datos, se procedió a explorar la segunda base de datos con el fin de seleccionar cuál de ellas será la utilizada para realizar el estudio.

Se encontró que en la base de datos de Agronometrics no contaba con los problemas mencionados anteriormente, las estampas de tiempo se encontraban de manera correcta y también esta contaba con una unidad de medida unitaria como lo es precio por libra de uva de mesa, en la Figura 4.4 se utilizó un gráfico de líneas para analizar el comportamiento de esta variable y a simple vista no se encontró un dato atípico extremo como en el caso anterior.

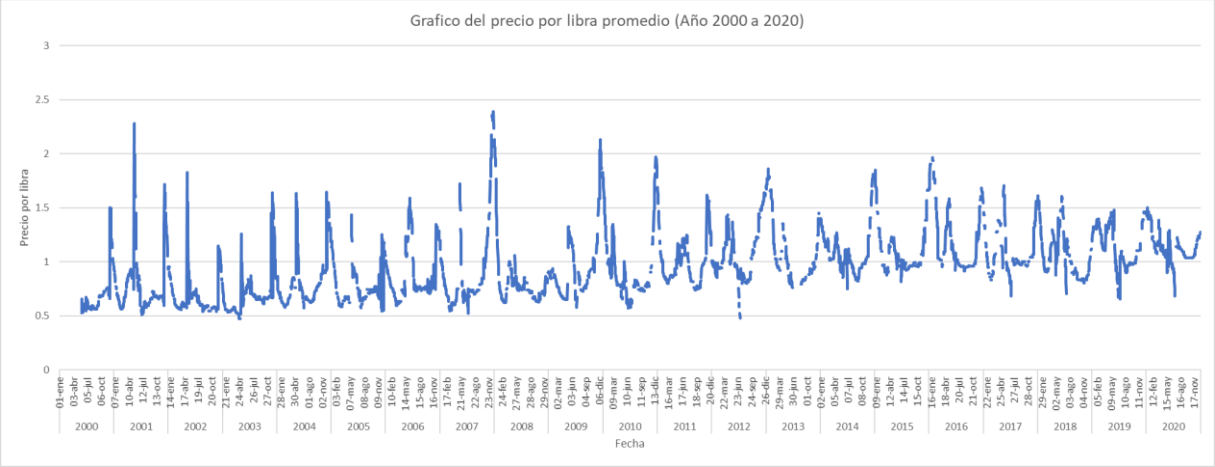

Figura 4.4. *Gráfico de línea del precio por libra promedio de la uva de mesa del año 2000 al 2020, base de datos Agronometrics (elaboración propia).*

Debido a que la base de datos está estructurada en forma de matriz y que al momento de generar los datos se seleccionó como variable adicional el origen de las transacciones, no se cuentan con más de un registro por día, haciendo que no sea necesario realizar operaciones adicionales para corregir este punto.

Analizando la base de datos referente a los volúmenes de importación a Estados Unidos de los diferentes países, se construyó una gráfica de líneas (Figura 4.5) con la finalidad de poder ver su comportamiento general e identificar posibles errores de calidad.

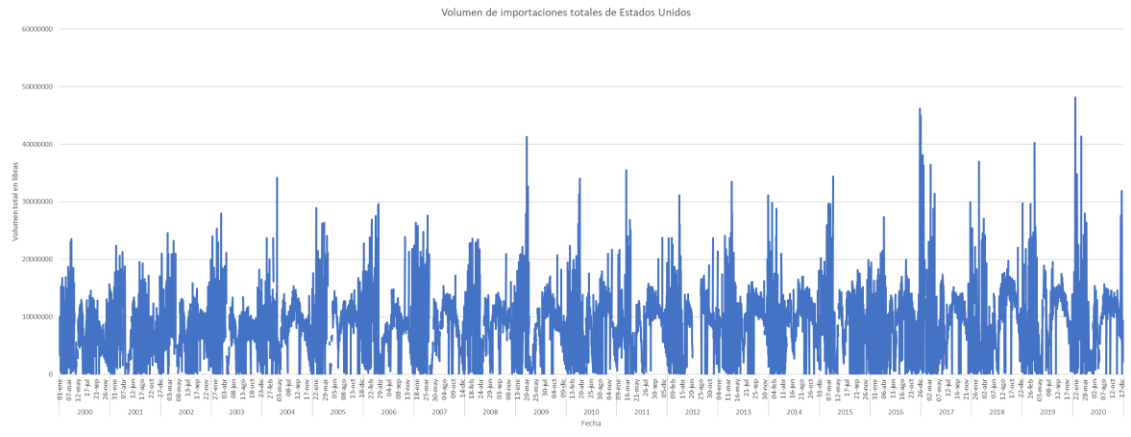

Figura 4.5. *Gráfico de línea del volumen de exportaciones a Estados Unidos de la uva de mesa del año 2000 al 2020, base de datos Agronometrics (elaboración propia).*

Con base a la gráfica anterior se consideró que los datos proporcionados por Agronometrics cuentan con una mejor estructura y menos errores de calidad por lo que se decidió utilizar esta base de datos para el análisis.

Para este proyecto, una de las principales tareas consiste en delimitar las ventanas de mercado o de oportunidad en donde México puede ingresar con posibles mejores precios debido a que hay poca oferta por parte de sus otros competidores; una forma de comprobar que se cuenta con la información necesaria es elaborando una matriz en donde por día se promedie el porcentaje de participación de cada uno de los países considerando los 21 años de datos, resultando en la Figura 4.6.

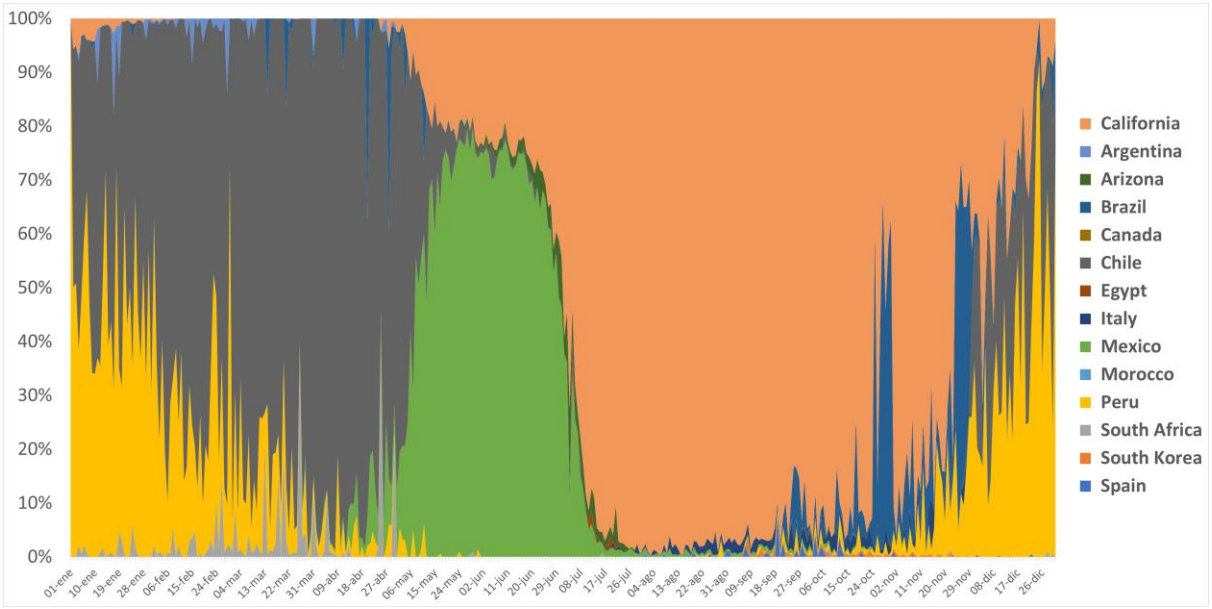

Figura 4.6. *Porcentaje promedio de participación en el mercado de uva de mesa en Estados Unidos por país (elaboración propia).*

La sección que se encuentra de color verde en la Figura 4.6 son los datos de México, el gris oscuro representa a Chile, el naranja representa a California, Estados Unidos y el resto de los colores representa la participación de distintos países. Como se puede ver existe un periodo en donde México cuenta con la mayoría de la participación del mercado, que coincide con la salida de chile y el ingreso de California, Estados Unidos; sin embargo, se puede ver como California cuenta aproximadamente con un 25% de la participación la mayor parte del periodo de comercialización de México, el cual va alrededor del 30 de abril al 10 de junio. Sin embargo, esta gráfica contiene información de manera global, por lo que se utilizará un mapa de calor por año para identificar de manera visual periodos del año en donde la oferta no sea tan alta, para esto a la columna de volumen total se le restó la participación de México y da lugar a la Figura 4.7.

Figura 4.7. *Mapa de calor de oferta al mercado de Estados Unidos (elaboración propia).*

En la Figura 4.7 no se puede distinguir la participación de los demás países ya que solo considera el volumen global, por lo que la forma de interpretar esta figura consiste en que entre más verde sea el recuadro menos saturado se encuentra el mercado ese día con respecto a su año; por el contrario, entre más rojo, mayor saturación, por lo que se puede ver claramente como en los meses de enero a mayo y diciembre existe una oferta relativamente baja. Sin embargo, los meses de agosto a noviembre existe una alta oferta, que coincide con California, Estados Unidos, mientras que los meses de inicios de mayo a julio existe poca oferta, lo que representa la posible ventana de mercado de México; solo existe un año en donde la ventana se recorre de agosto a octubre, otra observación relevante es que la ventana de mercado aparentemente ha disminuido con el pasar de los años.

En el caso de los precios, se realizó una gráfica de líneas (Figura 4.8) con el fin de poder ver si los datos se comportaban de manera normal y si serian de utilidad para poder compararlo con las posibles ventanas de mercado de México, por cuestión de visualización la frecuencia utilizada fue mensual.

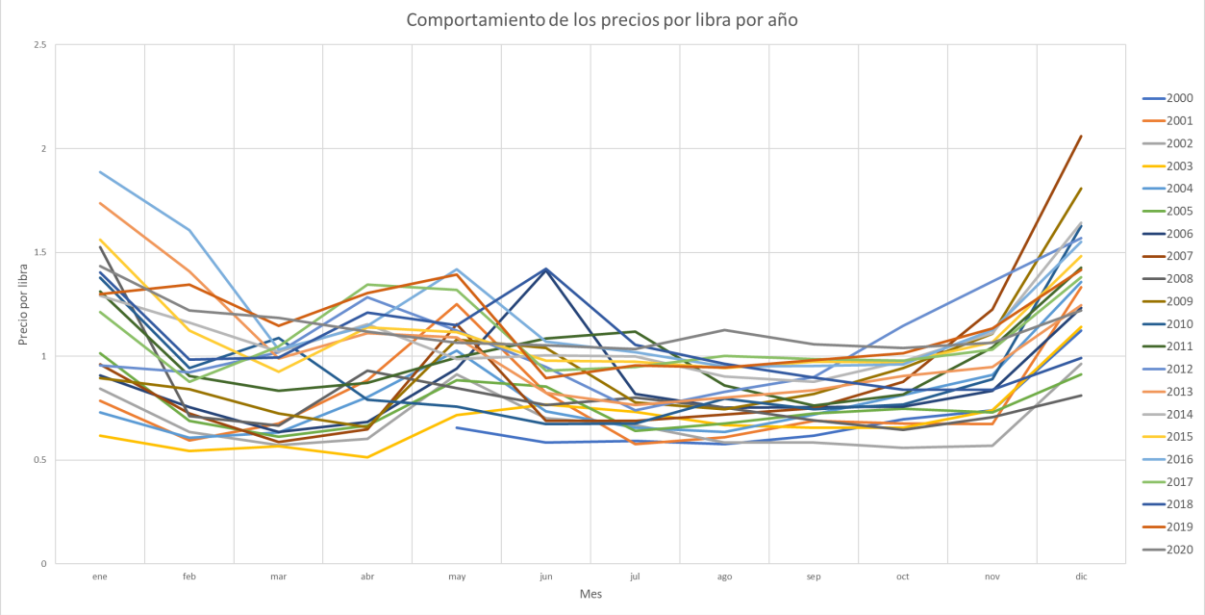

Figura 4.8. *Gráfico de línea del precio por libra promedio de la uva de mesa del 2000 al 2020 global (elaboración propia)*

Analizando la Figura 4.8, se puede ver que en el periodo de comercialización de julio a noviembre los valores del precio son bajos y se comportan de manera relativamente estable, esto posiblemente debido a las regulaciones del comercio nacional de la uva de mesa y a que no se incurre en gastos de exportación, ya que es la temporada dominada por Estados Unidos. En donde existe la mayor cantidad de variabilidad de los datos se encuentra en el mes de enero, mayo y diciembre, en donde mayo coincide con la participación de la comercialización de México, se deberán de evaluar los casos específicos por año de México para ver si no existen valores erróneos, un aspecto a resaltar es que el periodo de mayo comienza con uno de los precios más altos alcanzados por año y comienza a decaer entre más se acerca a julio, este mismo efecto también sucede a principios y finales de año.

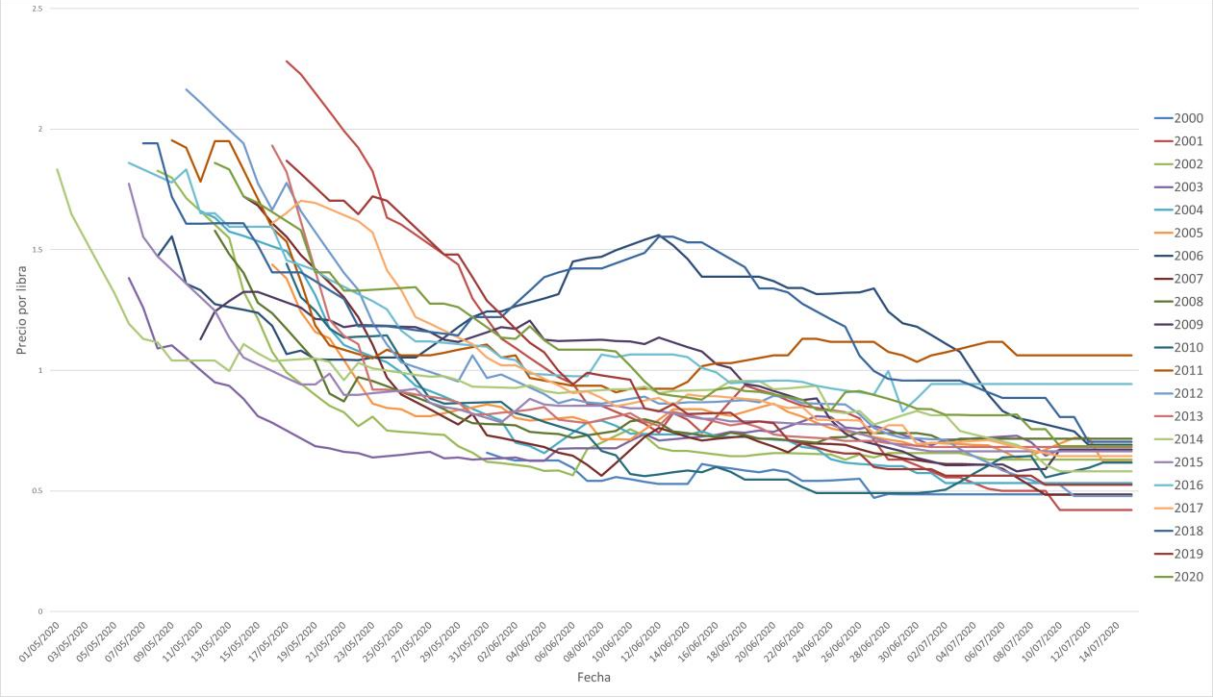

Figura 4.9. *Gráfico de línea del precio por libra promedio de la uva de mesa del 2000 al 2020 solo México (elaboración propia).*

En la Figura 4.9 se graficaron por año solamente los precios correspondientes a México; se puede observar, cómo aparentemente siguen el mismo comportamiento y no cuentan con datos atípicos. Sin embargo, uno de los problemas que se puede ver en las figuras anteriores referente al precio es que los cortes que se encuentran en las líneas significan que son fechas sin registro de valores.

Con el fin de reforzar la comprensión de las diferentes variables y encontrar aspectos relevantes de la base de datos, específicamente en el caso de México, se realizó la Tabla 4.3 en donde se brinda la estadística básica del precio por libra y el volumen por año.

Estadística básica (Precio México) por año

Año	Registros	Promedio	Desviación poblacional	Mínimo	Máximo
2000	22	0.5640	0.0499	0.4717	0.6592
2001	37	1.0331	0.4929	0.4218	2.2816
2002	41	0.8242	0.3366	0.5655	1.8280
2003	46	0.7624	0.1602	0.6237	1.3835
2004	37	0.9043	0.3336	0.5318	1.6602
2005	41	0.8285	0.1824	0.6214	1.4379
2006	46	1.2218	0.2230	0.6917	1.5617
2007	39	0.8488	0.3349	0.4842	1.7207
2008	38	0.8618	0.2266	0.6999	1.5785
2009	43	0.9983	0.2386	0.5806	1.3254
2010	42	0.7261	0.2531	0.4914	1.4424
2011	43	1.1780	0.2983	0.9095	1.9527
2012	44	1.0169	0.4230	0.4789	2.1637
2013	32	0.8909	0.2828	0.6817	1.9323
2014	50	0.9838	0.2457	0.5806	1.8870
2015	39	0.9270	0.2386	0.6638	1.7736
2016	39	1.1749	0.2879	0.8301	1.8598
2017	38	1.0098	0.3192	0.6441	1.7033
2018	46	1.2894	0.2861	0.7037	1.9414
2019	36	0.9952	0.4115	0.5247	1.8688
2020	42	1.0983	0.3043	0.6861	1.8598

Estadística básica (Volumen México) por año

Año	Registros	Promedio	Desviación poblacional	Mínimo	Máximo
2000	64	3,039,063	2,200,148	40,000	6,430,000
2001	73	2,284,521	2,117,526	10,000	7,160,000
2002	71	3,087,606	2,661,894	20,000	7,300,000
2003	82	3,599,146	2,844,172	10,000	9,500,000
2004	79	2,588,354	2,886,700	20,000	8,960,000
2005	86	3,794,302	3,906,642	10,000	11,180,000
2006	109	1,918,899	2,143,986	10,000	7,200,000
2007	105	2,763,524	3,420,929	10,000	11,020,000
2008	97	2,983,814	3,388,698	10,000	11,120,000
2009	94	2,567,340	2,690,701	10,000	8,220,000
2010	80	3,860,500	3,891,776	30,000	13,820,000
2011	71	3,741,268	3,228,641	20,000	9,590,000
2012	87	3,467,816	3,752,023	20,000	11,740,000
2013	107	2,757,944	3,572,092	20,000	10,360,000
2014	80	3,766,375	3,505,311	-40,000	11,600,000
2015	87	3,584,713	4,069,583	10,000	12,320,000
2016	110	2,655,455	3,582,854	20,000	11,360,000
2017	104	3,684,904	4,706,972	30,000	14,860,000
2018	90	3,224,000	3,273,625	10,000	11,900,000
2019	114	3,790,000	5,171,446	10,000	16,820,000
2020	125	3,111,760	4,157,178	30,000	14,130,000

Tabla 4.3. *Estadística básica del precio y volumen en libras de México (elaboración propia).*

Analizando las tablas se identificó que existe un valor negativo dentro del valor mínimo en la variable de volumen, que puede deberse a una devolución del producto o un error de captura; también se aprecia que el número de registros por año es poco. Sumado al análisis de las figuras se observa que se trata de un producto estacional, que solo se comercializa en ciertas temporadas del año y también se puede deber a que el mercado de cotización de Estados Unidos se encuentra generalmente cerrado los sábados, domingos y festivos; fuera de estos detalles el comportamiento de las demás variables se encuentra dentro de lo esperado.

Al término de esta etapa, ya se conoce de manera general el comportamiento de los datos y se procederá a la tarea de validación de datos.

4.2.5 Validación de los datos

En esta tarea se plasmarán todos los problemas de calidad identificados en la etapa de análisis exploratorio y se planteará una posible causa y forma de lidiar con este problema, el resumen de estos se encuentra en la Tabla 4.4 en la página 78.

Problema	Posible causa	Posible solución
Valor negativo en el volumen	Error de dedo o devolución de producto	Eliminar registro o imputación.
Días sin registro	Mercado de cotización cerrado sábados, domingo y festivos.	Imputación o no considerar esos días dentro de la serie temporal.

Tabla 4.4. *Tabla de validación de datos (elaboración propia).*

4.3 Preparación de los datos

En esta etapa se atenderán los problemas de calidad detectados en la etapa anterior y se realizarán las transformaciones necesarias para que se encuentren con la estructura correcta y puedan ser modelados.

4.3.1 Selección de variables de interés

Debido a que la finalidad del análisis es estudiar el comportamiento de los precios y de las ventanas de mercado de México, solo se considerarán los valores que se encuentran en la variable de origen que coincida con este país y se dejarán como alternativa las variables de volumen total y precio promedio.

4.3.2 Limpieza de datos

En esta tarea, lo primero que se realizó fue la eliminación de aquellas variables que no sean de interés como las operaciones realizadas por otros países, así mismo, considerando la Tabla 4.4, en el caso de los valores negativos se eliminarán ya que no hay forma de saber si se trataba de un dato erróneo o no. Referente a la imputación de los datos faltantes o no considerar ciertas fechas, no se realizará nada por el momento ya que no afecta como tal a alguna etapa anterior; dependiendo del modelo que se vaya a desarrollar, se planteará el realizar una imputación u otra acción correctiva.

4.3.3 Estandarización de datos

Para la base de datos de Agronometrics no hay necesidad de realizar ninguna estandarización de datos ya que todo se encuentra en el mismo formato o unidades, lo que facilita su análisis; si se hubiera utilizado la base de datos original de la USDA

se habrían tenido que estandarizar las fechas y unidades de medida que representan las transacciones.

4.3.4 Reestructuración y formateo de base de datos

Debido a que la base que la base de datos contaba con una estructura correcta, no fue necesario realizar una reestructuración o formateo, dependiendo de la etapa de modelado se podría volver a esta etapa a fin de que el modelo a utilizar reciba los datos en la estructura y formato correcto; sin embargo, se creó una nueva variable llamada “Volumen sin México” que considera el volumen exportado de otros países a Estados Unidos sin considerar las de México.

4.4 Modelado

Esta etapa consiste en llevar a cabo un análisis de los diferentes algoritmos que ayuden al cumplimiento del objetivo y verificar si los datos con los que se cuentan son adecuados para su implementación. Una vez que ya se haya validado y seleccionado el modelo, se deberá generar la prueba de diseño, en caso de que sea posible, dividiendo el conjunto de entrenamiento y prueba, con una proporción de entre 70-90% y 30-10% respectivamente con el fin de evaluar el modelo.

4.4.1 Análisis y selección de modelos

Para el cumplimiento del proyecto es necesario desarrollar un modelo que sea capaz de identificar cuando se presentan las ventanas de mercado y su duración a futuro, así como analizar el comportamiento que han tenido estas ventanas y los precios que tienen; tomando esto como base, se utilizó el formato de la etapa de análisis y selección de modelos para identificar los modelos y algoritmos que serán utilizados.

En la Tabla 4.5 se indican los algoritmos o modelos que se consideraron.

Nombre	Objetivo	descripción	Requisitos
ARIMA	Modelado de datos	ARIMA son las siglas en inglés de modelos de media móvil integrados autorregresivos. Es una técnica de pronóstico que proyecta los valores futuros de una serie basándose completamente en su propia inercia. Su principal aplicación se encuentra en el área de la previsión a corto plazo. Funciona mejor cuando sus datos muestran un patrón estable o consistente a lo largo del tiempo con una cantidad mínima de valores atípicos. A veces llamado Box-Jenkins (en honor a los autores originales).	Proceso estacionario (media 0 y varianza constante en el tiempo).
Auto ARIMA	Identificación	Devuelve el mejor modelo ARIMA según el valor AIC, AIC o BIC. La función realiza una búsqueda sobre el modelo posible dentro de las restricciones de orden proporcionadas.	Base de datos como serie de tiempo
Delimitador de ventanas de mercado (precio)	Segmentador	Se desarrolló un algoritmo propio que consiste en la segmentación de la base de datos en ventanas de comercialización excluyendo	Valores numéricos

		aquellos días atípicos fuera del grupo principal.	
Delimitador de ventanas de mercado (Volumen)	Segmentador	Algoritmo propio para identificar la temporada con menos saturación del mercado por año.	Valores numéricos
Agrupamiento jerárquico	Agrupamiento	Este algoritmo busca los datos más parecidos entre sí para crear una cantidad desconocida de grupos naturales.	Base de datos con una estructura de matriz No debe de haber datos faltantes

Tabla 4.5. *Evaluación de modelos a implementar (elaboración propia).*

4.4.2 Generar diseño de prueba

Debido a que los métodos que serán utilizados no serán de regresión en esta etapa del proyecto, no se destinará una parte del conjunto de datos para la validación de los modelos.

4.4.3 Construcción del modelo

La forma en la que se llevará a cabo la implementación de los modelos será la siguiente:

1. Identificar las ventanas de mercado y el comportamiento de los precios de México sin datos atípicos, calcular su duración y estadísticas descriptivas básicas.
2. Calcular intervalos de confianza al 95% para inicio y fin de las ventanas de mercado y del comportamiento de los precios de México.

3. Imputar los valores faltantes de los precios correspondientes en las ventanas de mercado.
4. Obtener estimadores de los parámetros ARIMA de las series de tiempo de los precios por cada una de las ventanas de mercado a través de la función Auto.arima en Rstudio.
5. Se aplicará el método de la distancia y la cantidad de agrupamientos óptimos para el conjunto de datos.
6. Se realizará un agrupamiento jerárquico de la matriz generada con los parámetros ARIMA por año.
7. Se obtendrán las características generales de cada uno de los grupos referentes a precios y volumen.

Identificación de ventana de mercado

Lo primero que se realizó fue la identificación de las ventanas de mercado mediante la aplicación de la variable “Volumen sin México” la cual representa el total de exportaciones de uva de mesa sin considerar a México, con la finalidad de obtener aquellos periodos de comercialización que serían óptimos para ingresar al mercado ya que éste no se encontraría saturado, poniendo a México en una posición ventajosa para la comercialización de su producto y obteniendo aparentemente mejores precios, debido a la oferta y demanda. Al analizar la base de datos con los volúmenes de exportación se observa que existen días en los cuales no se cuenta con registros como se había mencionado en la tabla 4.4, por lo que para lidiar con las fechas vacías se decidió cambiar la frecuencia de diaria a semanal solo para la identificación de las ventanas de mercado; además se optó por transformar el volumen en kilogramos por semana a su porcentaje de participación respecto a todas las semanas. También se obtuvo el promedio de este porcentaje para aplicarlo como medida de comparación para identificar la ventana, y para visualizar su comportamiento, se decidió realizar una gráfica con el año 2020, el cual se puede ver en la Figura 4.10. en la página 83.

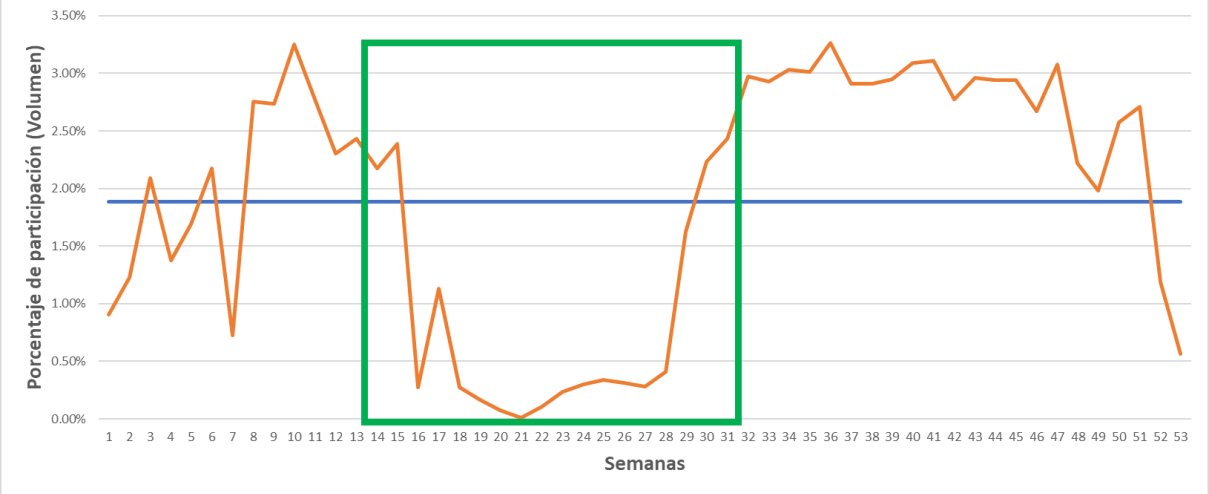

Figura 4.10. *Grafica del porcentaje de volumen de exportación de la uva de mesa a Estados Unidos (Elaboración propia).*

Para complementar el análisis de la Figura 4.10 se realizó la Figura 4.11 para graficar la participación promedio de México y como se observa, suele coincidir durante la ventana de mercado.

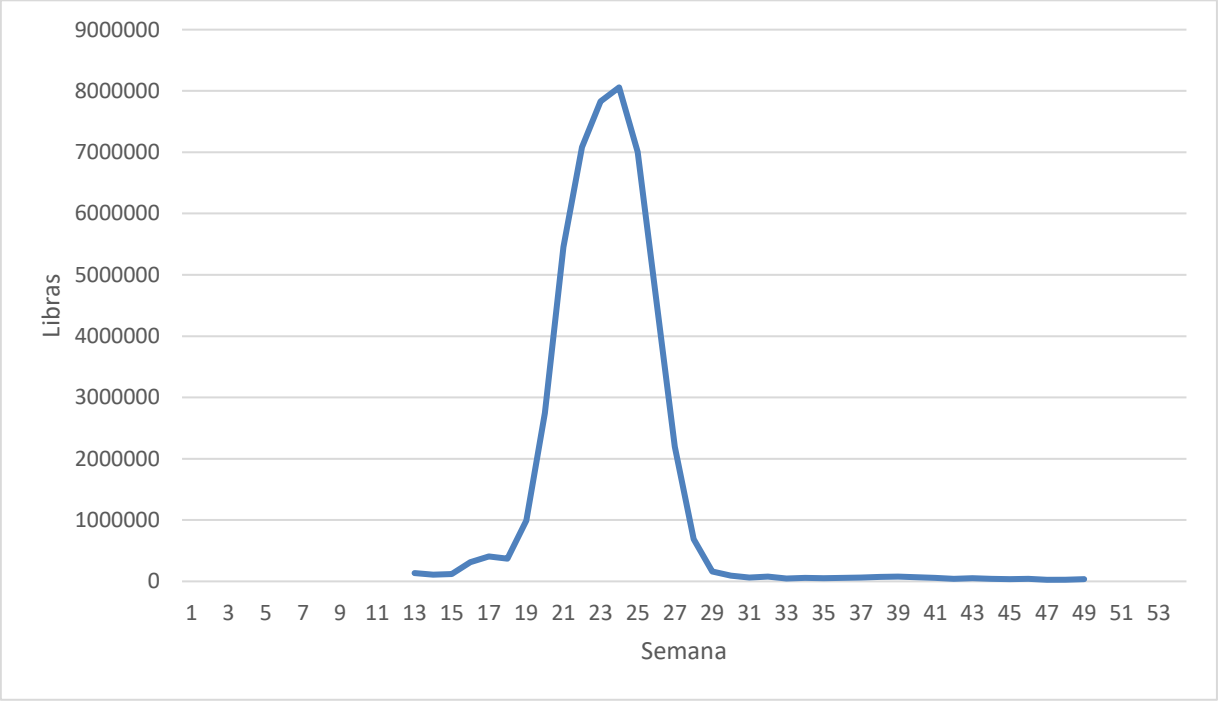

Figura 4.11. *Grafica del promedio de volumen de exportación de la uva de mesa por parte de México a Estados Unidos (Elaboración propia).*

La intención de este algoritmo es detectar la ventana de mercado con la menor participación de otros países, que en el caso de la Figura 4.10 está representado con un recuadro verde, y como se puede apreciar, la utilización del porcentaje promedio de participación delimita de manera adecuada la ventana buscada. El mismo comportamiento se repite en todos los años de nuestra base de datos; el algoritmo desarrollado permite observar que hay 5 semanas consecutivas con un valor debajo del promedio, el primer valor de esta serie de valores se considera el inicio de la ventana, mientras que 5 semanas seguidas se encuentran por arriba del promedio y es posterior al inicio, lo que implica la finalización de la ventana de mercado; los resultados de este algoritmo se plasmas en la tabla 4.6 en donde se presenta por año, el inicio, fin y duración de la ventana.

Año	Duracion en semanas	Semana de Inicio	Semana de finalizacion
2020	14	17	30
2019	14	17	30
2018	15	16	30
2017	14	17	30
2016	13	17	29
2015	13	17	29
2014	13	16	28
2013	14	17	30
2012	20	17	36
2011	15	17	31
2010	15	17	31
2009	13	17	29
2008	12	18	29
2007	11	18	28
2006	12	19	30
2005	12	19	30
2004	12	18	29
2003	13	18	30
2002	12	18	29
2001	12	18	29
2000	11	19	29
Promedio	13	17	30

Tabla 4.6. *Mayor ventana de mercado de la uva de mesa en Estados Unido por año (elaboración propia).*

En la Tabla 4.6 se observan algunos hechos importantes como que la semana de finalización se ha mantenido relativamente igual, sin embargo, el inicio ha empezado antes resultando en ventas de comercialización más duraderas; para poder ilustrar este hecho se construye una gráfica de línea.

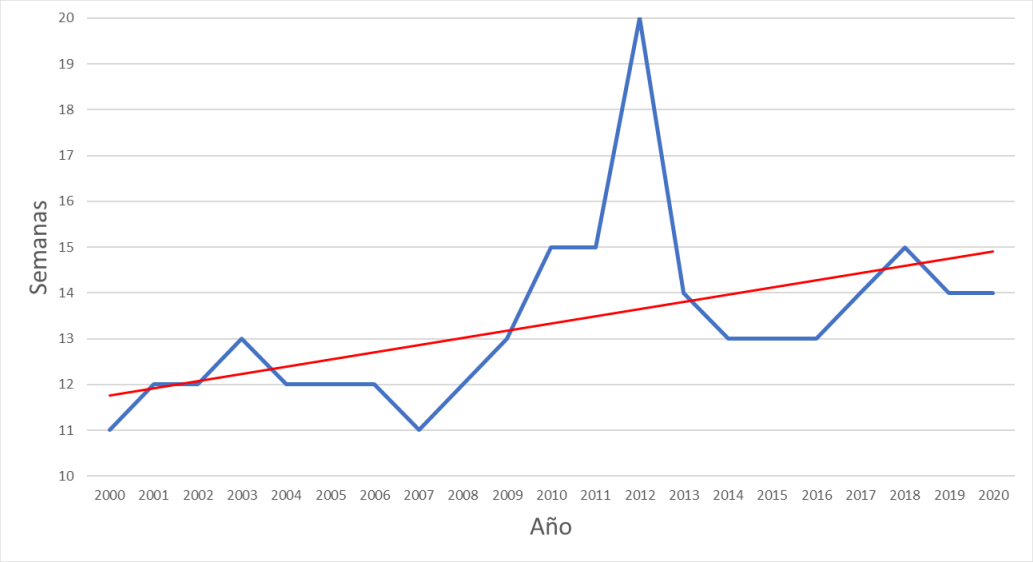

Figura 4.12. *Duración de las ventanas de mercado (Elaboración propia).*

Se puede ver en la Figura 4.12 que la duración de la ventana de mercado se ha ido incrementando y muestra una tendencia al alza.

Una vez obtenidos estos resultados, se procedió a calcular el intervalo de confianza del 95% para cada una de las variables de la tabla anterior, excluyendo el año, a fin de poder identificar entre qué semanas es posible que inicie y finalice la ventana de mercado. Los intervalos de confianza se calcularon haciendo uso de una prueba t de Student, ya que se desconoce la desviación estándar de la población; esta prueba se lleva a cabo a través de la función “t de 1 muestra” en estadística básica de MiniTab, que arroja los valores de la Tabla 4.7. en la página 86.

Intervalo de confianza	Promedio	Desv.Est.	Limite inferior	Limite superior
Semana de finalización	29.81	1.63	29.07	30.55
Semana de Inicio	17.48	0.87	17.08	17.87
Semanas de duración	13.33	1.96	12.44	14.23

Tabla 4.7. Intervalos de confianza al 95% para las ventanas de mercado (elaboración propia).

Como se puede ver en la Tabla 4.7, con un 95% de confianza, se encuentra que el inicio de la ventana de mercado se ubica entre la semana 17 – 18 y su finalización entre la semana 29 – 31, con una duración de 12 – 14 semanas. Una vez identificada la mayor ventana de oportunidad para la exportación de la uva de mesa, se procedió a identificar y estudiar el comportamiento de los precios obtenidos por México en el mercado de Estados Unidos.

Los precios de la uva de mesa Mexica durante las ventanas de mercado se identificaron a través de un algoritmo propio desarrollado en Python 3; la forma en la que funciona es identificando cuando hay cierta cantidad fechas con valores faltantes antes de encontrar una con un registro. En la Figura 4.13 se presenta una representación gráfica de su funcionamiento. Este algoritmo trabaja de esta manera debido a que tiene que identificar aquellos registros de precios que pertenecen a la temporada de comercialización principal de México, ya que antes de que ésta comience, tienen muy poco impacto en el análisis debido a la poca cantidad de registros y volumen de comercialización, posiblemente debido a exportaciones pequeñas de uva de mesa que se produjo antes de temporada, o de igual manera, que haya entrado fuera de temporada al mercado posiblemente debido a la logística, distribución o permisos.

Figura 4.13. *Funcionamiento del algoritmo para detectar ventana de mercado (Elaboración propia).*

Para esta investigación se utilizó una ventana de identificación de siete días. Para el inicio de la ventana de mercado real, se consideró que, si hay siete días previos sin valores y menos de siete posteriores, se toma como el inicio; en el caso de que haya siete valores posteriores después del punto de inicio, se identificará como el fin de la temporada principal para México. Una vez realizado este proceso se extraerá este fragmento de información de la base de datos, en donde se incluirá la fecha, día del año, volumen, precio por libra, la duración de la ventana restando al día de finalización el de inicio; también se calculó el precio mínimo, máximo y promedio, y los resultados se pueden ver en la Tabla 4.8.

Año	Inicio	Dia inicio	Fin	Dia fin	Duración	Min	Max	Promedio
2020	27/03/2020	87	23/07/2020	205	119	0.6867	1.8598	1.1615
2019	15/04/2019	105	06/08/2019	218	114	0.5247	1.8688	1.0183
2018	17/04/2018	107	19/07/2018	200	94	0.7373	1.9415	1.3167
2017	13/04/2017	103	14/07/2017	195	93	0.6441	1.7327	1.0437
2016	29/04/2016	120	03/08/2016	216	97	0.8388	1.9771	1.3163
2015	29/04/2015	119	15/07/2015	196	78	0.6638	1.7736	0.9465
2014	24/04/2014	114	18/07/2014	199	86	0.5868	1.8870	1.0596
2013	06/05/2013	126	23/09/2013	222	97	0.6817	1.9323	0.9022
2012	01/05/2012	122	30/06/2012	182	61	0.7258	2.1924	1.1460
2011	03/05/2011	123	21/07/2011	202	80	0.9238	1.9548	1.3901
2010	08/05/2010	128	23/07/2010	204	77	0.4914	1.4782	0.7400
2009	30/04/2009	120	30/07/2009	211	92	0.5868	1.9827	1.0518
2008	03/05/2008	124	07/08/2008	220	97	0.6999	1.5785	0.8763
2007	05/05/2007	125	24/07/2007	205	81	0.4842	1.7266	0.8622
2006	25/04/2006	115	19/07/2006	200	86	0.6917	1.8138	1.3269
2005	07/05/2005	127	29/07/2005	210	84	0.6214	1.4379	0.8338
2004	06/05/2004	127	19/07/2004	201	75	0.5318	1.6618	0.9169
2003	28/04/2003	118	30/07/2003	211	94	0.6237	1.9980	0.7885
2002	04/05/2002	124	26/07/2002	207	84	0.5655	1.8288	0.8441
2001	10/05/2001	130	21/07/2001	202	73	0.4218	2.2817	1.0446
2000	26/04/2000	117	12/07/2000	194	78	0.4717	0.6592	0.5654

Tabla 4.8. *Estadística descriptiva de participación de México en la ventana de mercado, su inicio, fin y duración (elaboración propia).*

Con base a los resultados de la ventana de mercado, se puede ver que conforme ha avanzado el tiempo, la fecha de entrada al mercado se ha ido adelantando, mientras

que en el caso de la finalización esta se ha ido recorriendo, lo que significa que esta ventana de mercado u oportunidad se ha incrementado con el tiempo; a su vez, el precio promedio también lo ha hecho, confirmando que los productores de México se han ido adaptando a las nuevas ventanas de mercado.

Una vez obtenidos estos resultados, se procedió a calcular el intervalo de confianza con del 95% a la fecha de inicio y fin de las temporadas a fin de poder identificar entre qué fechas es posible que comiencen y terminen temporadas futuras. Los intervalos de confianza se calcularon haciendo uso de una prueba t de Student ya que se desconoce la desviación estándar de la población; esta prueba se llevó a cabo a mediante la función “t de 1 muestra” en estadística básica de MiniTab, que arroja los valores de la Tabla 4.9.

Intervalos de confianza	Inicio	Fin	Duración	Min	Max	Promedio
Límite inferior	113	200	82	0.5726	1.6386	0.9094
Límite superior	123	209	94	0.6849	1.9392	1.1050

Tabla 4.9. *Intervalos de confianza para la duración, inicio, fin, mínimo, máximo y promedio de la ventana de mercado.*

También, para para identificar el comportamiento de la ventana, se procedió a utilizar una gráfica de líneas e incluir la tendencia; como se puede ver en la Figura 4.13, la duración de la participación ha incrementado y sigue con una tendencia al alza.

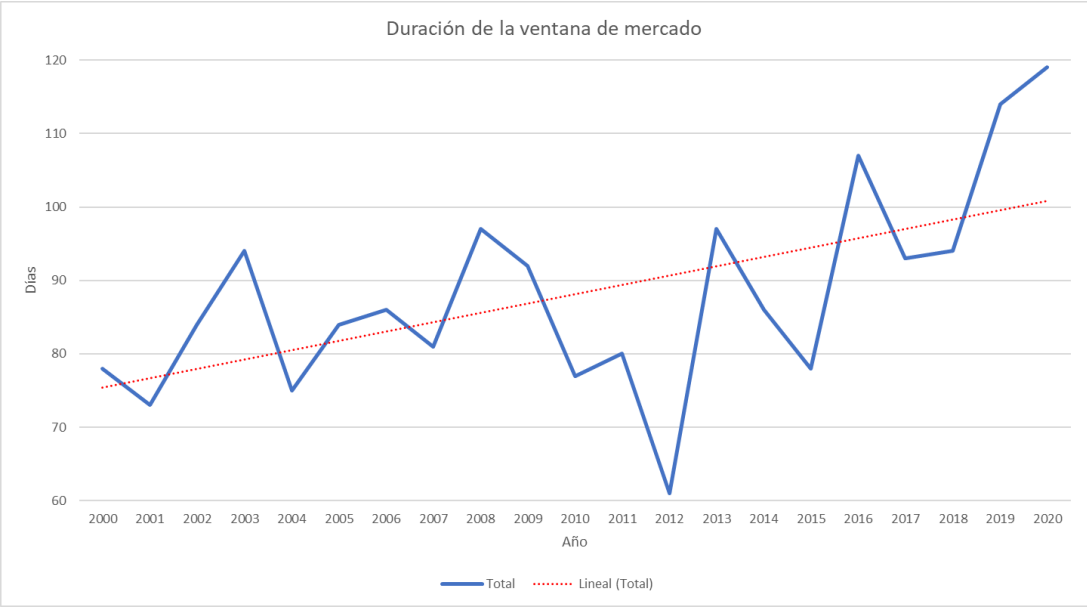

Figura 4.14. *Gráfico de línea de la duración de la ventana de mercado (elaboración propia).*

Una vez ya obtenidas las ventanas de mercado segmentadas, se procedió a imputar el precio de cada una a través de una interpolación lineal con el propósito de que conserven su comportamiento original; en la Figura 4.14 se puede observar del lado izquierdo como se encuentran los precios sin imputar y del lado derecho como resultan una vez hecho este procedimiento.

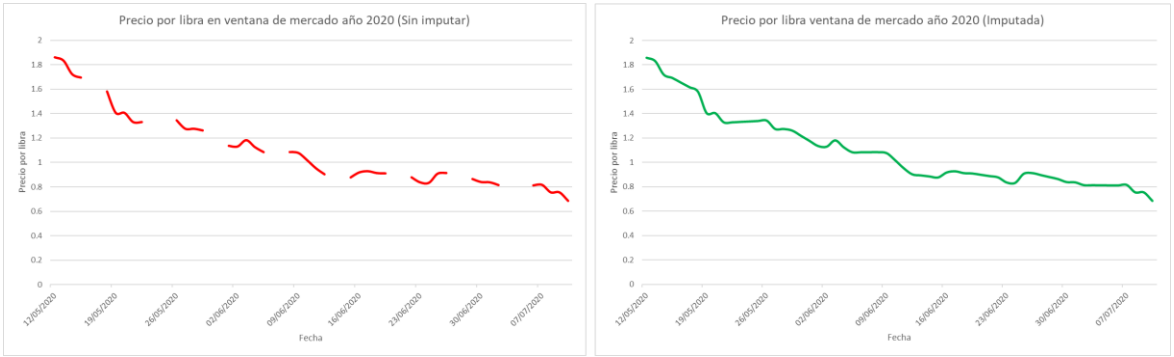

Figura 4.15. *Gráfico de línea del precio por libra imputado y no imputado año 2020 (elaboración propia).*

Una vez que todos los precios dentro de la ventana de mercado han sido imputados se procedió a implementar la función “Auto.arima” para poder identificar los parámetros

de cada una de las series con la finalidad de crear una matriz para poder implementar el algoritmo de agrupamiento.

La serie de tiempo (unidimensional) es una secuencia de observaciones de una variable en particular que consta de cuatro componentes: una tendencia, un elemento aleatorio, variación cíclica y estacional (Geurts, Box y Jenkins, 1977). Por lo tanto, para representar con precisión series de tiempo y definir la similitud entre ellas, es importante considerar los cuatro componentes de la serie temporal. Estos cuatro componentes se pueden capturar modelándolos mediante un modelo estacional de Box-Jenkins, mejor conocido como ARIMA (Hillmer y Wei, 1991).

Si se utilizan los modelos estimados de las series de tiempo para la agrupación, se superan los problemas que surgen debido a series de tiempo de diferentes longitudes y series de tiempo que están creciendo (Kalpakis, Gada y Puttagunta, 2001),

Nuestra noción de similitud es que dos series de tiempo son similares si los modelos físicos subyacentes que las generan son iguales o cercanos. La intuición detrás de esta noción de similitud es que, si los parámetros de los modelos ajustados a las series de tiempo son cercanos, entonces las series de tiempo se comportan de manera similar (probabilísticamente). Dos modelos son similares cuando un modelo se puede ajustar a una secuencia generada por otro modelo. Esta medida de similitud se puede utilizar para agrupar series de tiempo con mayor precisión. Permite hacer inferencias (extraer conocimiento) sobre una serie temporal de otras que pertenecen al mismo grupo y mejora nuestro conocimiento sobre la dinámica del sistema que se está estudiando (Kalpakis, Gada y Puttagunta, 2001).

En la Tabla 4.10 se muestran los resultados de los modelos ARIMA.

Precio							
Precio	**AR**	**I**	**MA**	**Precio**	**AR**	**I**	**MA**
2000	0	1	0	2011	0	2	1
2001	0	2	1	2012	0	2	1
2002	1	2	0	2013	0	2	1
2003	2	2	1	2014	3	1	0
2004	1	2	1	2015	0	2	1
2005	0	1	1	2016	0	2	2
2006	0	1	1	2017	0	1	0
2007	0	2	1	2018	0	1	1
2008	4	2	0	2019	0	2	1
2009	0	1	0	2020	0	2	1
2010	0	2	1				

Tabla 4.10. *Modelos ARIMA para precios imputados por ventana de mercado (elaboración propia).*

Una vez identificados los parámetros de los modelos ARIMA, se cargó la matriz generada a Rstudio, uno de los requisitos para que el modelo de agrupamiento brinde buenos resultados es que los datos se encuentren estandarizados dentro de una escala común, o sea, que cuenten con una media 0 y varianza de 1, lo que se realizó a través de la función “scale” de Rstudio.

Dentro de los modelos de agrupamiento se tiene que identificar que método de distancia se aplicará para la conexión de los nodos; para identificar el valor óptimo se aplicará la prueba Agnes mediante la función “agnes” en R estudio, la cual indica cual es el método más adecuado para el conjunto de datos utilizado. Se tiene que entre más cercano sea el valor a uno, mejor es el agrupamiento.

Método	Valor
ward	0.9400484
complete	0.8887731
average	0.8507654
single	0.8082295

Tabla 4.11. *Evaluación de método de distancia óptimo (elaboración propia).*

Como se puede ver en la Tabla 4.11 el método de distancia más adecuado para el conjunto de datos sería el “Ward” y, con base a éste, se estimará todo el resto del agrupamiento.

Lo siguiente es identificar el número de grupos óptimo; para esto se utilizó el paquete de Rstudio llamado “fviz_nbclust” que identifica que tan cercanos están los puntos dentro de los diversos grupos formados en K = n. El primer método probado en este paquete fue el “Wss” o como normalmente se conoce el método de codo, en el cual el número de grupos se determinará en donde la gráfica comience a volverse horizontal, a continuación, se muestra en la Figura 4.16.

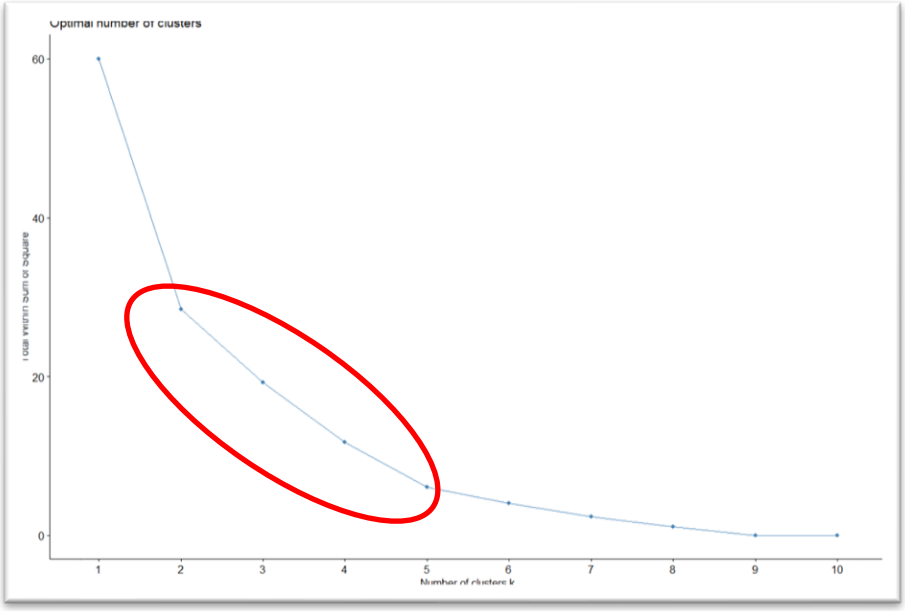

Figura 4.16. *Método Wss para identificar K (elaboración propia).*

En el método Wss la pendiente comienza a decrecer a partir de K = 2; sin embargo, esta termina de caer en K = 5, lo que nos indica que la cantidad de grupos óptimos se debe de encontrar entre 2 o 5. El siguiente método a probar fue el de silueta el cual se puede ver en la Figura 4.16. De igual, manera, en este método entre más cercano sea el valor a 1, mejor es el agrupamiento, por lo tanto, este método nos indica que el número de grupos óptimo se encontraría entre 3 y 5. Debido a que el número de grupos

en ambos métodos coincide se utilizará una K = 3 y después K =5 para identificar cuál valor logra captar y agrupar de manera más adecuada el comportamiento de los precios por año.

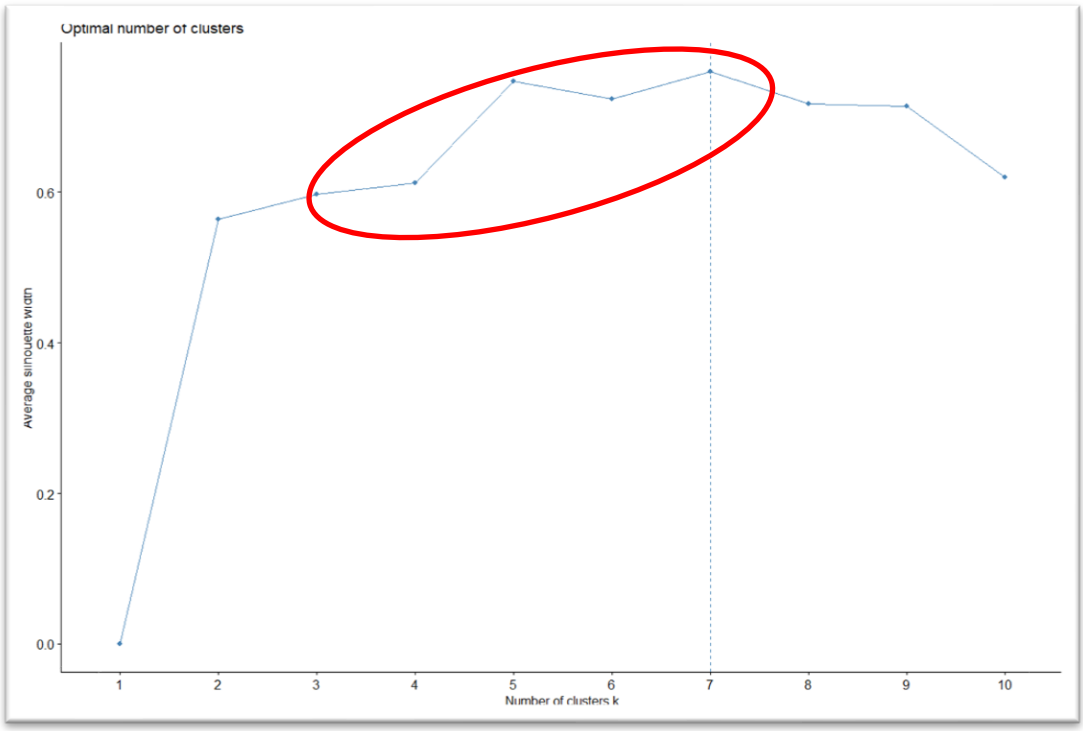

Figura 4.17. *método de silueta para identificar K (elaboración propia).*

Para la realización del agrupamiento, se utilizó el algoritmo de clúster jerárquico a través de la función “Hclust” del paquete “Stats” de R estudio, considerando el método de distancia “Ward” y la distancia euclidiana con 3 y 5 grupos; los resultados de este algoritmo se muestran en la Figura 4.17 para 3 grupos y la Figura 4.18 para cinco grupos.

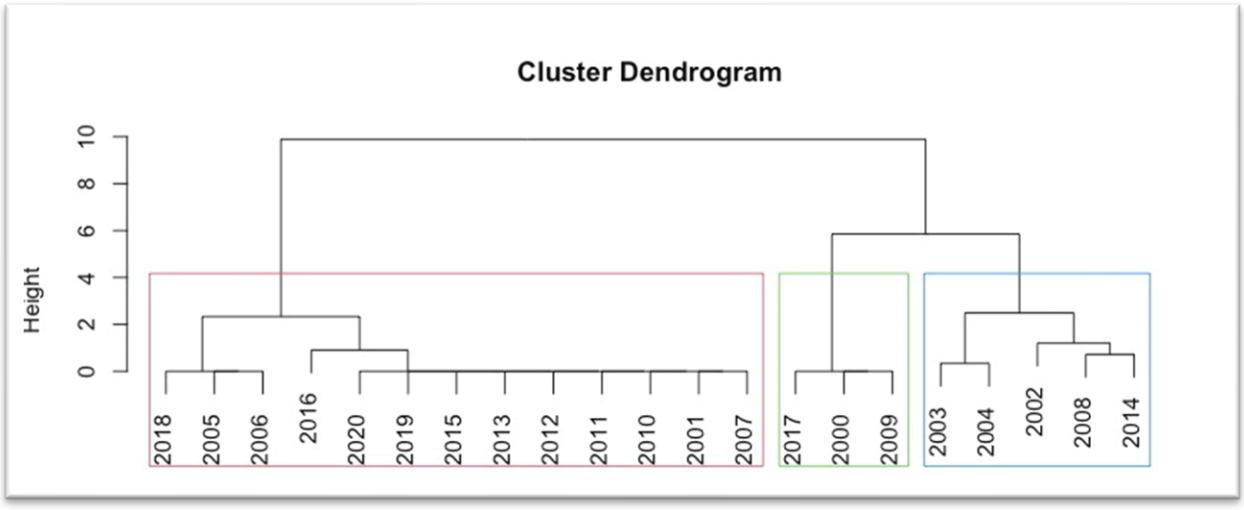

Figura 4.18. *Dendrograma cuando K = 3 (elaboración propia).*

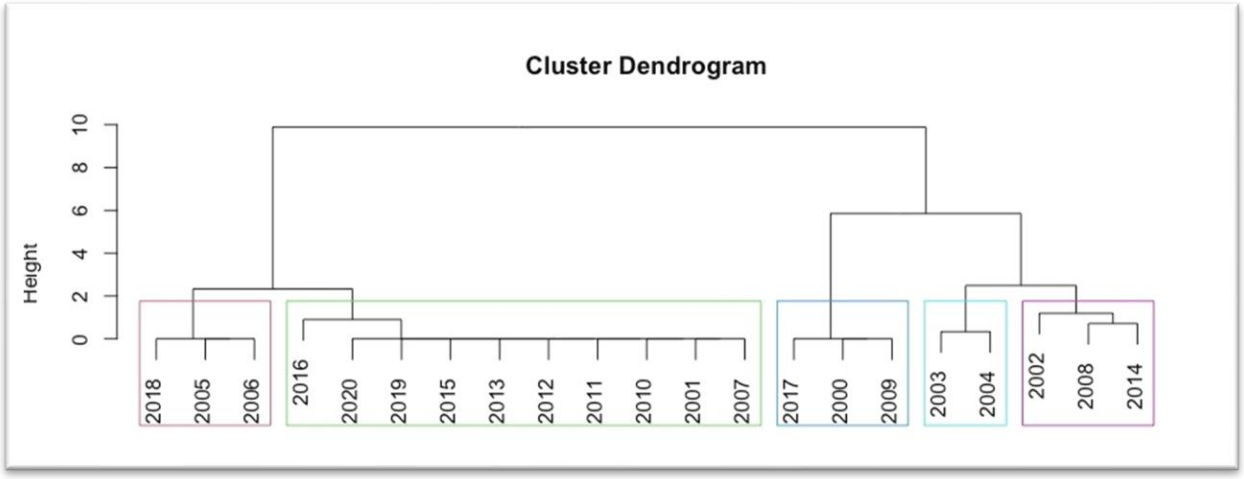

Figura 4.19. *Dendrograma cuando K =5 (elaboración propia).*

Como se puede ver en los dendrograma, los grupos formados no cuentan con una gran diferencia referente a los años que lo conforman; sin embargo, estos grupos no brindan mucha información a primera vista, por lo que se debe de analizar cada uno mediante estadísticos descriptivos y el comportamiento de los precios por grupo, con el fin de ver claras diferencias entre ellos.

Precio (K =3)											
Precio	AR	I	MA	Modelo	K = 3	Min	Max	Promed	Duracion	Inicio	Fin
2001	0	2	1	(0,2,1)	1	0.422	2.282	1.045	73	130	202
2005	0	1	1	(0,1,1)	1	0.621	1.438	0.834	84	127	210
2006	0	1	1	(0,1,1)	1	0.692	1.814	1.327	86	115	200
2007	0	2	1	(0,2,1)	1	0.484	1.727	0.862	81	125	205
2010	0	2	1	(0,2,1)	1	0.491	1.478	0.740	77	128	204
2011	0	2	1	(0,2,1)	1	0.924	1.955	1.390	80	123	202
2012	0	2	1	(0,2,1)	1	0.726	2.192	1.146	61	122	182
2013	0	2	1	(0,2,1)	1	0.682	1.932	0.902	97	126	222
2015	0	2	1	(0,2,1)	1	0.664	1.774	0.946	78	119	196
2016	0	2	2	(0,2,2)	1	0.839	1.977	1.316	97	120	216
2018	0	1	1	(0,1,1)	1	0.737	1.941	1.317	94	107	200
2019	0	2	1	(0,2,1)	1	0.525	1.869	1.018	114	105	218
2020	0	2	1	(0,2,1)	1	0.687	1.860	1.161	119	87	205
2000	0	1	0	(0,1,0)	2	0.472	0.659	0.565	78	117	194
2009	0	1	0	(0,1,0)	2	0.587	1.983	1.052	92	120	211
2017	0	1	0	(0,1,0)	2	0.644	1.733	1.044	93	103	195
2002	1	2	0	(1,2,0)	3	0.565	1.829	0.844	84	124	207
2003	2	2	1	(2,2,1)	3	0.624	1.998	0.789	94	118	211
2004	1	2	1	(1,2,1)	3	0.532	1.662	0.917	75	127	201
2008	4	2	0	(4,2,0)	3	0.700	1.579	0.876	97	124	220
2014	3	1	0	(3,1,0)	3	0.587	1.887	1.060	86	114	199

Tabla 4.12. *Información de la ventana de mercado por año cuando K = 3 (Elaboración propia).*

K = 3							
Grupo	Mínimo	Máximo	Promedio	Duración	Inicio	Fin	Modelos
1	0.6533	1.8645	1.0773	88	118	205	(0,1,1), (0,2,1), (0,2,2)
2	0.5676	1.4582	0.8870	88	113	200	(0,1,0)
3	0.6016	1.7908	0.8971	87	121	208	(1,2,0), (1,2,1), (2,2,1), (3,1,0), (4,2,0)

Tabla 4.13. *Ventana de mercado, sus precios y modelos cuando K=3 (elaboración propia).*

Analizando las Tablas 4.12 y 4.13, no se logra apreciar una clara diferencia entre los grupos referente a la duración, inicio y final de la ventana de mercado; sin embargo, al momento de analizarlos con sus parámetros del modelo ARIMA, se puede ver que el grupo 1 contiene aquellos modelos que se centran en su parte de medias móviles, mientras que el grupo 3 son combinaciones de los modelos AR, I y MA, siendo series más complicadas de modelar que los otros grupos. En lo que se refiere a los precios por grupo, el 1 es el que alcanza los precios más altos en mínimo, máximo y promedio, mientras que los otros 2 solo presentan diferencias algo notables en sus valores máximos alcanzados. Para poder expandir el análisis de estos grupos, se graficarán

sus valores del precio por día en el año (Figura 4.20) y ver su similitud en el comportamiento.

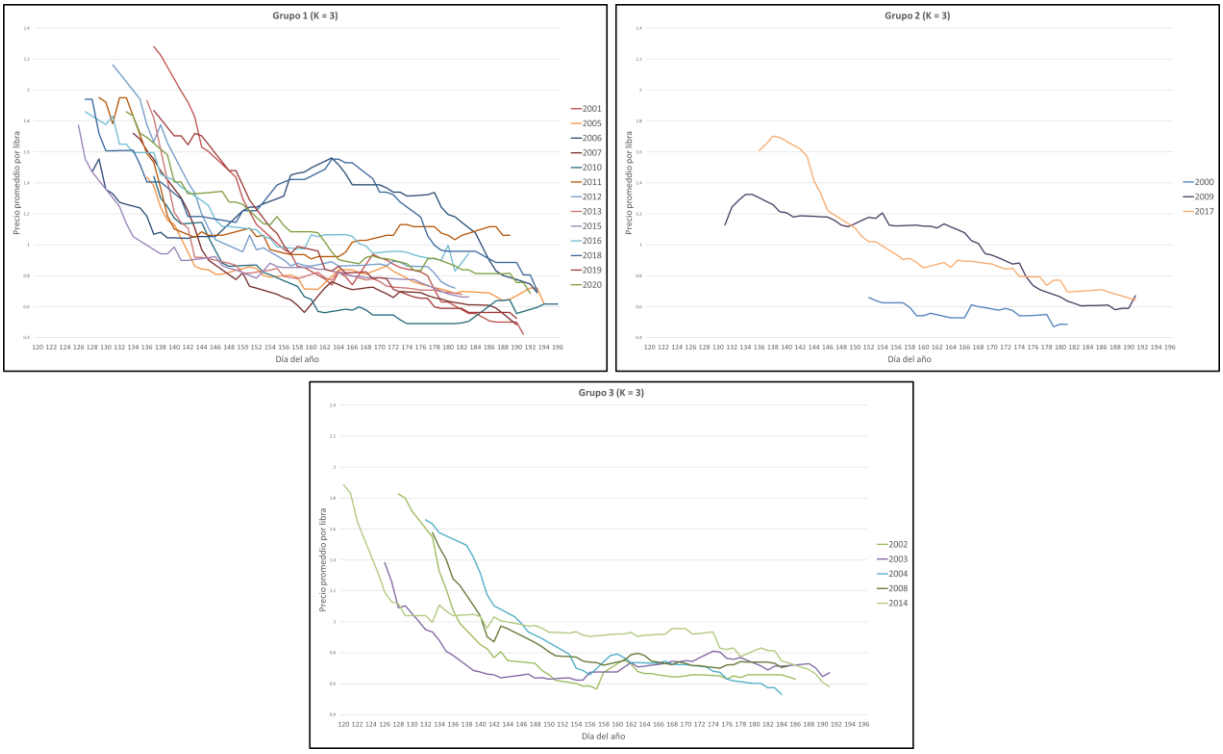

Figura 4.20. *Gráfico de línea del precio por libra promedio del 2000 al 2020 (elaboración propia).*

Analizando el comportamiento de las gráficas de la Figura 4.20, se observa que los grupos logran comparar la dinámica de las series de tiempo que lo conforman; en el caso del grupo 3, se presentan precios muy altos al inicio de la temporada y sufren una caída considerable y posteriormente se estabilizan, el grupo 2 comienza teniendo precios competitivos, pero sufre una rápida caída en sus precios y posteriormente constantemente van perdiendo más valor. En el caso del grupo 1, se puede ver que la mayoría de los casos tienen un comportamiento similar; sin embargo, en los años 2018 y 2006 se presenta un comportamiento notablemente diferente. En general todas las series de tiempo comienzan teniendo precios altos, sufren una rápida caída de los precios y posteriormente se estabilizan con una ligera tendencia a la baja; sin embargo, en los años mencionados no se sigue la leve tendencia a la baja, si no que tienen un comportamiento en forma de campana antes de terminar su ciclo, lo que pudiera

significar que el número de grupo igual a 3 no es el más adecuado para este conjunto de datos.

Precio (K = 5)

Precio	AR	I	MA	Modelo	K = 5	Min	Max	Promed	Duracion	Inicio	Fin
2005	0	1	1	(0,1,1)	1	0.621	1.438	0.834	84	127	210
2006	0	1	1	(0,1,1)	1	0.692	1.814	1.327	86	115	200
2018	0	1	1	(0,1,1)	1	0.737	1.941	1.317	94	107	200
2001	0	2	1	(0,2,1)	2	0.422	2.282	1.045	73	130	202
2007	0	2	1	(0,2,1)	2	0.484	1.727	0.862	81	125	205
2010	0	2	1	(0,2,1)	2	0.491	1.478	0.740	77	128	204
2011	0	2	1	(0,2,1)	2	0.924	1.955	1.390	80	123	202
2012	0	2	1	(0,2,1)	2	0.726	2.192	1.146	61	122	182
2013	0	2	1	(0,2,1)	2	0.682	1.932	0.902	97	126	222
2015	0	2	1	(0,2,1)	2	0.664	1.774	0.946	78	119	196
2016	0	2	2	(0,2,2)	2	0.839	1.977	1.316	97	120	216
2019	0	2	1	(0,2,1)	2	0.525	1.869	1.018	114	105	218
2020	0	2	1	(0,2,1)	2	0.687	1.860	1.161	119	87	205
2000	0	1	0	(0,1,0)	3	0.472	0.659	0.565	78	117	194
2009	0	1	0	(0,1,0)	3	0.587	1.983	1.052	92	120	211
2017	0	1	0	(0,1,0)	3	0.644	1.733	1.044	93	103	195
2003	2	2	1	(2,2,1)	4	0.624	1.998	0.789	94	118	211
2004	1	2	1	(1,2,1)	4	0.532	1.662	0.917	75	127	201
2002	1	2	0	(1,2,0)	5	0.565	1.829	0.844	84	124	207
2008	4	2	0	(4,2,0)	5	0.700	1.579	0.876	97	124	220
2014	3	1	0	(3,1,0)	5	0.587	1.887	1.060	86	114	199

Tabla 4.14. *Información de la ventana de mercado por año cuando K = 5 (elaboración propia).*

K = 5

Grupo	Mínimo	Máximo	Promedio	Duración	Inicio	Fin	Modelos
1	0.6835	1.7311	1.1591	88	116	203	(0,1,1)
2	0.6443	1.9045	1.0528	88	119	205	(0,2,1), (0,2,2)
3	0.5676	1.4582	0.8870	88	113	200	(0,1,0)
4	0.5778	1.8299	0.8527	85	123	206	(1,2,1), (2,2,1)
5	0.6174	1.7648	0.9267	89	121	209	(1,2,0), (3,1,0), (4,2,0)

Tabla 4.15. *Ventana de mercado, sus precios y modelos cuando K=5 (elaboración propia).*

En el caso de 5 grupos, se observa que, de igual manera como en el caso de las duraciones (inicio y fin), no existe una diferencia notable entre sus valores; sin embargo, en base a los parámetros de los modelos, se ve como en el grupo 5, son solo series con componente AR, en el caso del grupo 4, es una combinación de modelos, para el grupo 3, es un modelo sencillo con solo una diferencia, para el grupo

2, son modelos MA sin componentes AR y en el grupo 1 son modelos de igual manera MA pero solo con una diferencia y no 0, siendo series de tiempo más sencillas de modelar. Referente al precio promedio que alcanzan los grupos los que mejores cotizaciones tienen son los grupos 1 y 2, mientras que los más bajos son los 3 y 4.

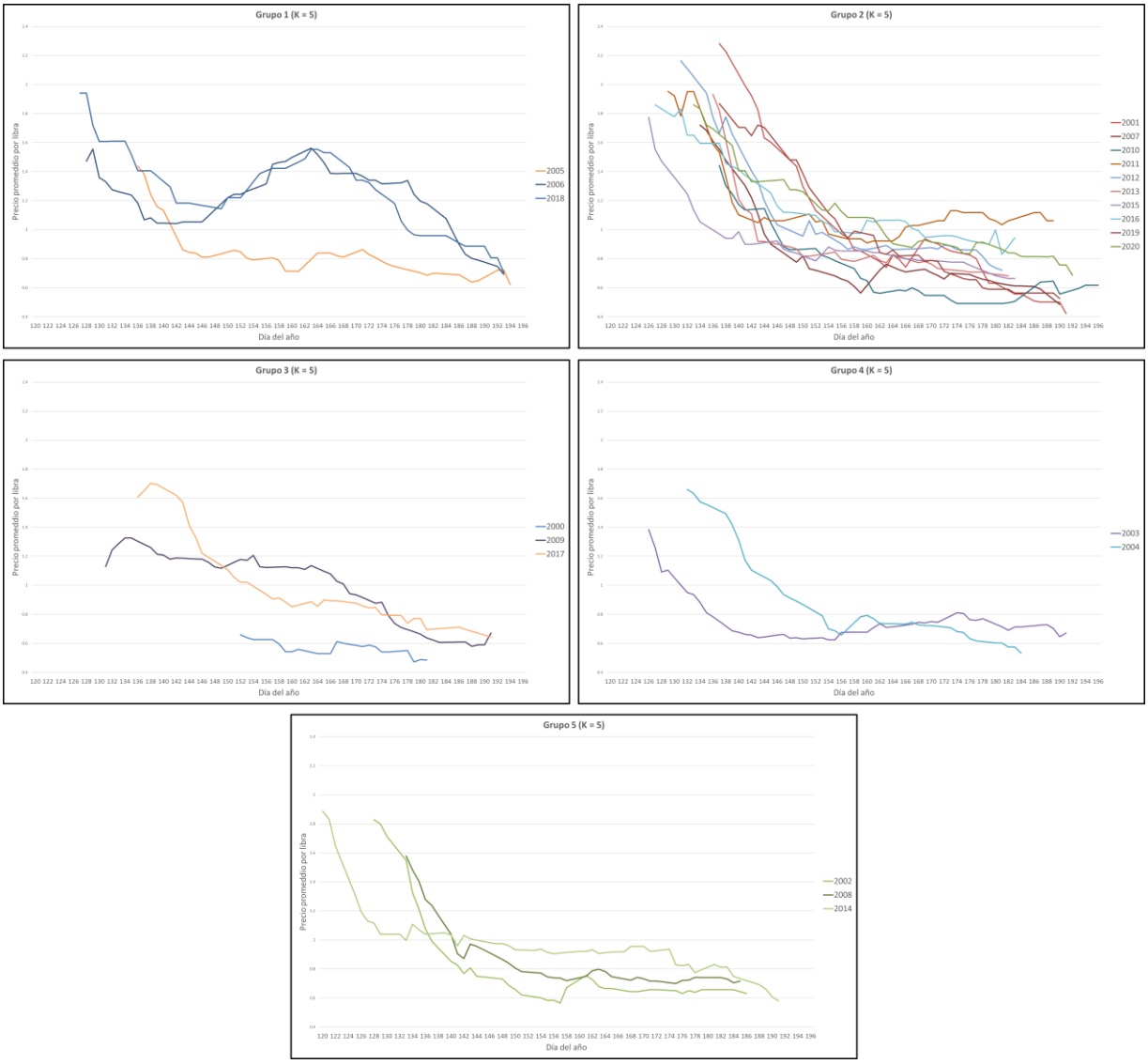

Figura 4.21. *Gráfico de línea del precio por libra promedio del 2000 al 2020 K =5 (elaboración propia).*

Analizando el comportamiento de los diferentes grupos, se puede ver como el caso del año 2006 y 2018, que distan mucho del grupo en el que fueron asignados cuando k era igual 3, ahora forman su propio grupo incluyendo al año 2005; el clúster 1 se

caracteriza por precios altos, después sufren una leve caída en sus cotizaciones y vuelven a incrementar su precios y caer de manera lenta, en el caso del grupo 2 estos tienen un comportamiento lineal, alcanzan los precios de cotización más altos de todos los grupos, pero con una tendencia marcada a la baja, el grupo 3 cuenta con un ligero gancho al inicio de su temporada teniendo precios crecientes y después siguen el mismo comportamiento del grupo 2, en el caso del grupo 4, comienza con precios altos y posteriormente sufre una caída marcada, para después estabilizarse, mientras en el grupo 5 estos cuentan con una caída marcada de su precios seguido de una ligera recuperación y estabilización.

4.4.4 Evaluación del desempeño

Al momento no se han implementado algoritmos de clasificación o predicción, por lo que la parte de evaluación del desempeño del modelo no es necesario hasta esta parte del estudio.

5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

Esta sección comprende las conclusiones obtenidas a partir del trabajo de investigación y su aplicación, se presentan una serie de recomendaciones, así como trabajos futuros referente a esta investigación.

5.1 Conclusiones

Al término de este proyecto, una de las principales conclusiones a las que se llegó es que la utilización de las metodologías para la extracción del conocimiento, como lo son CRISP -DM y KDD, permitieron desarrollar el proyecto de manera ordenada y lógica, teniendo un enfoque de trabajo estándar que lleva a traducir de manera eficiente y correcta el problema a tareas específicas por realizar, permitiendo que personas sin mucha experiencia en el desarrollo de este tipo de proyecto o trabajos puedan llevarlos a cabo y que consideren los pasos esenciales para que pueda ser llevado a cabo con éxito.

Referente a la implementación y utilización de los modelos ARIMA, éstos fueron de gran utilidad para entender el comportamiento de las series de tiempo, lo que a su vez permitiría la utilización del algoritmo de agrupamiento jerárquico para agrupar aquellas ventanas de mercado por año con características similares, y estudiar sus comportamientos por grupo, teniendo como conclusión que la utilización de estos algoritmos de forma conjunta permitió obtener conocimiento muy valioso para investigaciones futuras, demostrando que mediante el agrupamiento de los parámetros de los modelos ARIMA se puede caracterizar de manera correcta el comportamiento de los precios de cotización.

La identificación de las ventanas de mercado, mediante el algoritmo desarrollado para este proyecto, permitió conocer el comportamiento de estas mismas, analizar si se han ido reduciendo o incrementando la ventana de oportunidad y contrastarlo con la participación de México, con la finalidad de saber si ha sido aprovechada por los

productores; también el conocimiento del inicio, fin y duración de la ventana permitirá a los productores plantear mejores estrategias de producción y distribución del producto, así como también prever con base a otros factores, como los climatológicos, políticos o administrativos. El entender cómo se comportaron los precios de la uva de mesa mexicana en dichas ventanas de mercado refuerza el tema de las estrategias y la prevención en años futuros.

Este trabajo pretende que el CIAD, con sus proyectos de investigación de la CPUMS, tenga a la mano una metodología de trabajo estándar para aplicar técnicas de minería de datos de manera ordenada, y así delimitar ventanas de mercado y analizar el comportamiento de los precios de la uva de mesa en estos periodos de comercialización, haciendo posible que trabajos futuros de investigación sean llevados a cabo con herramientas básicas de análisis y que sus investigadores se enfoquen en estudios en los que logren mayor precisión considerando variables que no fueron tomadas en cuenta en este trabajo.

5.2 Recomendaciones

Como parte de las recomendaciones, se enlistan las siguientes:

1. Se recomienda contactar directamente al Departamento de Agricultura de los Estados Unidos y solicitar la información necesaria con el formato adecuado, dejando de lado el sistema de reportes que probó ser deficiente en la información que brinda, y así obtener una mayor cantidad de información confiable.
2. La parametrización de los modelos ARIMA deberá de hacerse a través de algún método de optimización como lo es AUTO.ARIMA para que el algoritmo de agrupamiento brinde mejores resultados.
3. Se recomienda siempre realizar la prueba “Agnes” para identificar el tipo de conexión más adecuada para el conjunto de datos.

4. Una vez realizada la agrupación de los datos, es recomendable utilizar el método “Wss” o de “Silueta” para identificar la cantidad óptima de grupos naturales que se deberán de considerar.
5. Se recomienda realizar el modelo de agrupamiento según distintos parámetros con la finalidad de identificar cual es el que brinda mejores resultados.
6. Tomar en consideración factores como el clima, petróleo, etc. para el estudio de los precios de venta.
7. Se recomienda compartir los resultados obtenidos de la investigación de manera resumida a los productores, con la finalidad de que validen los resultados obtenidos y sea utilizada para la generación de mejores estrategias de comercialización y producción.

5.3 Trabajos futuros

La delimitación de las ventanas de mercado y el análisis del comportamiento de los precios dentro de las ventanas de mercado, es una herramienta útil para los productores de la uva de mesa; sin embargo, para trabajos futuros se recomienda tener en consideración una base de datos que cuente con la demanda de dicho producto para poder contrastarlo con el comportamiento de los precios y de las duraciones de las ventanas de mercado, así como utilizar otros factores relevantes como lo son el clima, temperatura, sobre todo tener en consideración los tiempos y condiciones necesarias para poder producir el producto, la logística de exportación para que se tome ventaja de la ventana de comercialización y de los precios más competitivos.

Otro trabajo futuro consistirá en generar un modelo que funcione en base a modelos de regresión, y con base a los parámetros estimados, clasifique dentro de uno de los grupos actuales ya identificados, y así poder apoyar a los productores a que estime el comportamiento del precio en esa ventana.

Por último, este trabajo forma parte de un proyecto de mayor alcance en el estudio de la cadena productiva de la uva de mesa para el grupo de investigación del CIAD, en el cual se buscará mejorar el rendimiento del modelado y agrupamiento de los datos y la delimitación de las ventanas de mercado.

6. REFERENCIAS

Ciad.mx. (2020). Misión, Visión y Objetivos. [online] Disponible en: https://www.ciad.mx/acerca-ciad/mision-vision [consultado 4 Mar. 2020].

México Table Grapes | AALPUM. 2020. Historia. [online] Disponible en: <https://aalpum.org/historia/> [consultado 3 abril 2020].

Ciad.mx. (2020). Desarrollo Regional. [online] Ciad.mx. Disponible en: <https://www.ciad.mx/coordinaciones/hermosillo/desarrollo-regional> [consultado 3 abril 2020].

Aranda Figueroa, A. (2016). Liderazgo y Organizaciones Sustentables en el Sistema de Vid de Mesa Sonorense. Maestría en Desarrollo Regional. Centro de Investigación en Alimentación y Desarrollo, A.C.

Karen Montaño, S. (2010). Estrategia Dual de Producción Fincada en el Mercado: Una Alternativa para Mejorar la Competitividad de la Cadena Productiva de la Uva de Mesa Sonorense. Maestría en Desarrollo Regional. Centro de Investigación en Alimentación y Desarrollo, A.C.

Aghabozorgi, S., Seyed Shirkhorshidi, A. y Ying Wah, T. (2015) “Time-series clustering – A decade review”, *Information Systems*, 53, pp. 16–38. doi: 10.1016/j.is.2015.04.007.

Anuario Estadístico de la Producción Agrícola (2020) *Servicio de Información Agroalimentaria y Pesquera*. Disponible en: https://nube.siap.gob.mx/cierreagricola/ (Consultado: el 31 de diciembre de 2021).

Aranda Figueroa, A. N. (2016) “Liderazgo Y Organizaciones Sustentables En El Sistema Vid De Mesa Sonorense”.

De arce, R. y Mahía, R. (2003) “Modelos Arima”, *Programa Citius.- Técnicas de Previsión de variables financieras*, (1), p. 32.

auto.arima function - RDocumentation (2020).

Becerra-Fernandez, I., Zanakis, S. H. y Walczak, S. (2002) “Knowledge discovery techniques for predicting country investment risk”, *Computers and Industrial Engineering*, 43(4), pp. 787–800. doi: 10.1016/S0360-8352(02)00140-7.

Bergmeir, C. y Benítez, J. M. (2012) “On the use of cross-validation for time series predictor evaluation”, *Information Sciences*, 191, pp. 192–213. doi: 10.1016/j.ins.2011.12.028.

Berrar, D. (2018) “Cross-validation”, 1, pp. 542–545.

Brockwell, P. J. y Davis, R. A. (2002) *Introduction to Time Series and Forecasting, Second Edition*, *Springer*.

Buczak, A. L. y Guven, E. (2016) “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”, *IEEE Communications Surveys and Tutorials*, 18(2), pp. 1153–1176. doi: 10.1109/COMST.2015.2494502.

Chatfield, C. (2003) *The Analysis of Time Series*, *The Analysis of Time Series*. Chapman and Hall/CRC. doi: 10.4324/9780203491683.

Chatfield, C., Little, R. J. A. y Rubin, D. B. (2002) “Statistical Analysis with Missing Data.”, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 151(2), p. 375. doi: 10.2307/2982783.

Chen, F. *et al.* (2015) “Data mining for the internet of things: Literature review and challenges”, *International Journal of Distributed Sensor Networks*, 2015(i). doi: 10.1155/2015/431047.

CIAD., A.C. (2021).

Cierre de producción agricola 2020 a nivel estatal - Sonora (2020) *sagarhpa*. Disponible en: http://oiapes.sagarhpa.sonora.gob.mx (Consultado: el 31 de diciembre de 2021).

Craig K., E. (2010) “Applied missing data analysis”, en *THE GUILFORD PRESS*. SAGE

Publications, Inc.

Dieterle, D. F. (2019) *2.4.1. Crossvalidation*. Disponible en: http://www.frank-dieterle.de/phd/2_4_1.html (Consultado: el 24 de noviembre de 2021).

Dietterich, T. (1995) “Overfitting and Undercomputing in Machine Learning”, *ACM Computing Surveys (CSUR)*, 27(3), pp. 326–327. doi: 10.1145/212094.212114.

E. O. Box, G. *et al.* (2008) *Time Series Analysis Forecasting and Control*.

Esling, P. y Agon, C. (2012) “Time-series data mining”, *ACM Computing Surveys*, 45(1). doi: 10.1145/2379776.2379788.

“Estadística uva de mesa sonorense” (2020).

Estimación del volumen y valor de los principales productos agrícolas exportados 2020 - Sonora (2020) *sagarhpa*. Disponible en: http://oiapes.sagarhpa.sonora.gob.mx (Consultado: el 31 de diciembre de 2021).

Fattah, J. *et al.* (2018) “Forecasting of demand using ARIMA model”, *International Journal of Engineering Business Management*, 10, p. 184797901880867. doi: 10.1177/1847979018808673.

Fayyad, U., Haussler, D. y Stolorz, P. (1996) “KDD for Science Data Analysis: Issues and Examples”, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 50–56.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996a) “From data mining to knowledge discovery in databases”, *AI Magazine*, 17(3), pp. 37–53.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996b) “Knowledge Discovery and Data Mining: Towards a Unifying Frame Work”, *AAAI Press*, 96, pp. 82–88.

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996c) “The KDD Process for Extracting Useful Knowledge from Volumes of Data”, *Communications of the ACM*, 39(11), pp. 27–34. doi: 10.1145/240455.240464.

Fayyad, U. y Stolorz, P. (1997) “Data mining and KDD: Promise and challenges”, *Future Generation Computer Systems*, 13(2–3), pp. 99–115. doi: 10.1016/s0167-739x(97)00015-0.

Fourier, A. (1999) *Handbook of Time Series Analysis, Signal Processing, and Dynamics*. Elsevier. doi: 10.1016/B978-0-12-560990-6.X5000-3.

Fulcher, B. D. y Jones, N. S. (2014) “Highly comparative feature-based time-series classification”, *IEEE Transactions on Knowledge and Data Engineering*, 26(12), pp. 3026–3037. doi: 10.1109/TKDE.2014.2316504.

Gamarra, C., Guerrero, J. M. y Montero, E. (2016) “A knowledge discovery in databases approach for industrial microgrid planning”, *Renewable and Sustainable Energy Reviews*, 60, pp. 615–630. doi: 10.1016/j.rser.2016.01.091.

Gamboa, J. C. B. (2017) “Deep Learning for Time-Series Analysis”. Disponible en: http://arxiv.org/abs/1701.01887.

Geurts, M., Box, G. E. P. y Jenkins, G. M. (1977) “Time Series Analysis: Forecasting and Control”, *Journal of Marketing Research*, 14(2), p. 269. doi: 10.2307/3150485.

Ghahramani, Z. (2004) “Unsupervised Learning”, en *Machine Learning*, pp. 72–112. doi: 10.1007/978-3-540-28650-9_5.

Graham, J. W. (2009) “Missing data analysis: Making it work in the real world”, *Annual Review of Psychology*, 60, pp. 549–576. doi: 10.1146/annurev.psych.58.110405.085530.

Gupta, A. *et al.* (2019) “Stock Market Prediction Using Data Mining Techniques”, *SSRN Electronic Journal*, 2(2). doi: 10.2139/ssrn.3370789.

Guyon, I. (1997) “A scaling law for the validation-set training-set size ratio”, *AT&T Bell Laboratories*, pp. 1–11. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1337&rep=rep1&type=pdf.

Habib, U., Hayat, K. y Zucker, G. (2016) “Complex building’s energy system operation patterns analysis using bag of words representation with hierarchical clustering”, *Complex Adaptive Systems Modeling*, 4(1). doi: 10.1186/s40294-016-0020-0.

Han, J. y Kamber, M. (2006) *Data Mining: Concepts and Techniques*, *Morgan Kaufmann*. Editado por Intergovernmental Panel on Climate Change. Cambridge.

Haykin, S. (1999) *Neural Networks, A Comprehensive Foundation*.

Hierarchical Cluster Analysis (2020).

Hillmer, S. C. y Wei, W. W. S. (1991) “Time Series Analysis: Univariate and Multivariate Methods.”, *Journal of the American Statistical Association*, 86(413), p. 245. doi: 10.2307/2289741.

Ho, A. (2017) “Beyond the Dataset: Understanding Sociotechnical Aspects of the Knowledge Discovery Process Among Modern Data Professionals”, *UWSpace*. Disponible en: https://uwspace.uwaterloo.ca/handle/10012/11835.

Hyndman J, R. (2011) *Cyclic and seasonal time series*.

Hyndman, R. J. (2009) “Moving averages”, p. 5.

Hyndman, Rob J y Athanasopoulos, G. (2018a) *2.9 White noise | Forecasting: Principles and Practice (2nd ed)*. Disponible en: https://otexts.com/fpp2/wn.html# (Consultado: el 7 de enero de 2022).

Hyndman, Rob J y Athanasopoulos, G. (2018b) *8.1 Stationarity and differencing | Forecasting: Principles and Practice (2nd ed)*. Disponible en: https://otexts.com/fpp2/stationarity.html (Consultado: el 8 de enero de 2022).

Hyndman, Rob J y Athanasopoulos, G. (2018c) *8.3 Autoregressive models | Forecasting: Principles and Practice (2nd ed)*. Disponible en: https://otexts.com/fpp2/AR.html (Consultado: el 7 de enero de 2022).

Hyndman, Rob J. y Athanasopoulos, G. (2018) *Business Forecasting: Principles and*

Practice, *Monash University*. doi: 10.2307/1054108.

Hyndman, Rob J y Athanasopoulos, G. (2018d) *Chapter 8 ARIMA models | Forecasting: Principles and Practice (2nd ed)*. Disponible en: https://otexts.com/fpp2/arima.html (Consultado: el 8 de enero de 2022).

Hyndman, R. J. y Koehler, A. B. (2006) “Another look at measures of forecast accuracy”, *International Journal of Forecasting*, 22(4), pp. 679–688. doi: 10.1016/j.ijforecast.2006.03.001.

J. Frawley, W., Piatetsky-Shapiro, G. y J. Matheus, C. (1992) “Knowledge Discovery in Databases: An Overview”, en *AI Magazine*. Berlin, Heidelberg: AI Magazine. doi: https://doi.org/10.1609/aimag.v13i3.1011.

javatpoint (2019) *Supervised Machine learning - Javatpoint*. Disponible en: https://www.javatpoint.com/supervised-machine-learning (Consultado: el 24 de noviembre de 2021).

Jawadi, F. y Ftiti, Z. (2019) “Oil price collapse and challenges to economic transformation of Saudi Arabia: A time-series analysis”, *Energy Economics*, 80, pp. 12–19. doi: 10.1016/j.eneco.2018.12.003.

Jerez, J. M. *et al.* (2010) “Missing data imputation using statistical and machine learning methods in a real breast cancer problem”, *Artificial Intelligence in Medicine*, 50(2), pp. 105–115. doi: 10.1016/j.artmed.2010.05.002.

Jothi, N., Rashid, N. A. y Husain, W. (2015) “Data Mining in Healthcare – A Review”, *Procedia Computer Science*, 72, pp. 306–313. doi: 10.1016/j.procs.2015.12.145.

Junninen, H. *et al.* (2004) “Methods for imputation of missing values in air quality data sets”, *Atmospheric Environment*, 38(18), pp. 2895–2907. doi: 10.1016/j.atmosenv.2004.02.026.

Kalpakis, K., Gada, D. y Puttagunta, V. (2001) “Distance measures for effective clustering of ARIMA time-series”, *Proceedings - IEEE International Conference on Data*

Mining, ICDM, pp. 273–280. doi: 10.1109/icdm.2001.989529.

Kang, H. (2013) “The prevention and handling of the missing data”, *Korean Journal of Anesthesiology*, 64(5), pp. 402–406. doi: 10.4097/kjae.2013.64.5.402.

Kantardzic, M. (2011) *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition*, *Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition*. doi: 10.1002/9781118029145.

Khawaja, S. G. *et al.* (2017) “A novel multiprocessor architecture for k-means clustering algorithm based on network-on-chip”, *Proceedings of the 2016 19th International Multi-Topic Conference, INMIC 2016*, pp. 1–5. doi: 10.1109/INMIC.2016.7840112.

Kotsiantis, S. B. (2007) “Supervised Machine Learning: A Review of Classification Techniques”.

Kotu, V. y Deshpande, B. (2015) “Data Mining Process”, en *Predictive Analytics and Data Mining*. Elsevier, pp. 17–36. doi: 10.1016/B978-0-12-801460-8.00002-1.

Kutbay, U. (2018) “Partitional Clustering”, *Recent Applications in Data Clustering*. doi: 10.5772/intechopen.75836.

Kwiatkowski, D. *et al.* (1992) “Testing the null hypothesis of stationarity against the alternative of a unit root”, *Journal of Econometrics*, 54(1–3), pp. 159–178. doi: 10.1016/0304-4076(92)90104-Y.

Larrañaga, P., Inza, I. y Moujahid, A. (2012) “Tema 14. Clustering”, pp. 1–11. Disponible en: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=4&ved=2ahUKEwig3-ne88_nAhXOrFkKHbisDRUQFjADegQIBxAB&url=http%3A%2F%2Fwww.sc.ehu.es%2Fccwbayes%2Fdocencia%2Fmmcc%2Fdocs%2Ft14clustering.pdf&usg=AOvVaw2IBAoHOrXTyVbnd9Ld6M-n.

Little, T. D. *et al.* (2016) “Missing Data”, en *Developmental Psychopathology*. Hoboken,

NJ, USA: John Wiley & Sons, Inc., pp. 1–37. doi: 10.1002/9781119125556.devpsy117.

M.Z., N. y Amir H., A. (2020) “Insights into Performance Fitness and Error Metrics for Machine Learning”, pp. 1–25.

Madin Rivera, A. (2021) *Serie de Tiempo (Introducción)*. Disponible en: https://rpubs.com/AlbertoMadinRivera/710202 (Consultado: el 25 de febrero de 2022).

Marsh, H. W. (1998) “Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes”, *Structural Equation Modeling*, 5(1), pp. 22–36. doi: 10.1080/10705519809540087.

Martínez, B. y Gilabert, M. A. (2009) “Vegetation dynamics from NDVI time series analysis using the wavelet transform”, *Remote Sensing of Environment*, 113(9), pp. 1823–1842. doi: 10.1016/j.rse.2009.04.016.

Matheus, C. J., Chan, P. K. y Piatetsky-Shapiro, G. (1993) “Systems for Knowledge Discovery in Databases”, *Data Mining for Bioinformatics*, (December), pp. 81–112. doi: 10.1201/b13091-3.

ML | Hierarchical clustering (Agglomerative and Divisive clustering) (2019).

Montaño, K. y Preciado, J. M. (2017) “La productividad del trabajo en la producción de uva de mesa sonorense.”, *Revista de turismo, economía y negocios*, 3(2), pp. 58–82.

Moreno, M. V. *et al.* (2017) “Applicability of Big Data Techniques to Smart Cities Deployments”, *IEEE Transactions on Industrial Informatics*, 13(2), pp. 800–809. doi: 10.1109/TII.2016.2605581.

Mota López, A. (2016) *PRONÓSTICO DEL PRECIO DEL CRUDO DE EXPORTACIÓN MEXICANO CON LA METODOLOGÍA DE BOX-JENKINS PARA SERIES DE TIEMPO*. BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA.

Murtagh, F. (1983) “A survey of recent advances in hierarchical clustering algorithms”,

Computer Journal, 26(4), pp. 354–359. doi: 10.1093/comjnl/26.4.354.

N. O. Sadiku, M. *et al.* (2016) “Data visualization”, *International Journal of Engineering Research And Advanced Technology*, 02(12).

Nanda, S. J. y Panda, G. (2014) “A survey on nature inspired metaheuristic algorithms for partitional clustering”, *Swarm and Evolutionary Computation*, 16, pp. 1–18. doi: 10.1016/j.swevo.2013.11.003.

El Naqa, I. y Murphy, M. J. (2015) “What Is Machine Learning?”, en *Machine Learning in Radiation Oncology*. Cham: Springer International Publishing, pp. 3–11. doi: 10.1007/978-3-319-18305-3_1.

Nau, R. (2020a) *Stationarity and differencing of time series data*.

Nau, R. (2020b) *Statistical forecasting: notes on regression and time series analysis*, *Fuqua School of Business*. Disponible en: https://people.duke.edu/~rnau/411home.htm (Consultado: el 9 de enero de 2022).

Nitish, S. *et al.* (2014) “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, *Journal of Machine Learning Research*, 15, pp. 1929–1958.

Noor, M. N. *et al.* (2013) “Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution”, *Key Engineering Materials*, 594–595(January), pp. 889–895. doi: 10.4028/www.scientific.net/KEM.594-595.889.

Noor, N. M. *et al.* (2015) “Comparison of linear interpolation method and mean method to replace the missing values in environmental data set”, *Materials Science Forum*, 803, pp. 278–281. doi: 10.4028/www.scientific.net/MSF.803.278.

Ogbuabor, G. y F. N, U. (2018) “Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value”, *International Journal of Computer Science and Information Technology*, 10(2), pp. 27–37. doi: 10.5121/ijcsit.2018.10203.

P., J. M. (2011) “Agrupamiento de patrones correlacionados y con incertidumbre: caso

patrones climáticos en la producción de uva de mesa en un viñedo de Sonora".

Palachy, S. (2019) *Stationarity in time series analysis*.

Pete, C. *et al.* (2000) "Crisp-Dm 1.0", *CRISP-DM Consortium*, p. 76.

Potter, K. (2006) "Methods for Presenting Statistical Information: The Box Plot", *Visualization of Large and Unstructured Data Sets*, 4, pp. 97–106.

Preciado Rodriguez, J. M., Romero Dessens, L. F. y Ojeda Benítez, S. (2011) *Agrupamiento de patrones correlacionados y con incertidumbre: caso patrones climáticos en la producción de uva de mesa en un viñedo de Sonora*. Universidad Autónoma de Baja California.

Puchalsky, W. *et al.* (2018) "Agribusiness time series forecasting using Wavelet neural networks and metaheuristic optimization: An analysis of the soybean sack price and perishable products demand", *International Journal of Production Economics*, 203(June), pp. 174–189. doi: 10.1016/j.ijpe.2018.06.010.

Riyadi, M. A. A. *et al.* (2017) "Clustering stationary and non-stationary time series based on autocorrelation distance of hierarchical and K-means algorithms", *International Journal of Advances in Intelligent Informatics*, 3(3), pp. 154–160. doi: 10.26555/ijain.v3i3.98.

Roberts, D. R. *et al.* (2017) "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure", *Ecography*, 40(8), pp. 913–929. doi: 10.1111/ecog.02881.

Robinson, A. P. y Hamann, J. D. (2011) *Forest Analytics with R*, *Forest Analytics with R*. doi: 10.1007/978-1-4419-7762-5.

RPubs (2016) *Gaussian White Noise Time Series*, *RStudio*. Disponible en: https://rpubs.com/ks190995/118845 (Consultado: el 7 de enero de 2022).

Samuel, A. L. (1988) "Some Studies in Machine Learning Using the Game of Checkers.

II—Recent Progress", en *Computer Games I*. New York, NY: Springer New York, pp. 366–400. doi: 10.1007/978-1-4613-8716-9_15.

Savalei, V. y Rhemtulla, M. (2012) "On Obtaining Estimates of the Fraction of Missing Information From Full Information Maximum Likelihood", *Structural Equation Modeling*, 19(3), pp. 477–494. doi: 10.1080/10705511.2012.687669.

Saxena, A. *et al.* (2017) "A review of clustering techniques and developments", *Neurocomputing*, 267, pp. 664–681. doi: 10.1016/j.neucom.2017.06.053.

Shahabi, C., Tian, X. y Zhao, W. (2000) "TSA-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries on time-series data", *Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM*, pp. 55–68. doi: 10.1109/ssdm.2000.869778.

Shcherbakov, M. V. *et al.* (2013) "A survey of forecast error measures", *World Applied Sciences Journal*, 24(24), pp. 171–176. doi: 10.5829/idosi.wasj.2013.24.itmies.80032.

Shearer, C. *et al.* (2000) "The CRISP-DM model: The New Blueprint for Data Mining", *Journal of Data Warehousing*, 5(4), pp. 13–22.

Shmueli, G., C. Bruce, P. y R. Patel, N. (2017) *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMINER.*

Siami-Namini, S., Tavakoli, N. y Siami Namin, A. (2018) "A Comparison of ARIMA and LSTM in Forecasting Time Series", en *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 1394–1401. doi: 10.1109/ICMLA.2018.00227.

Simon, R. (2003) "Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n)", *ACM SIGKDD Explorations Newsletter*, 5(2), pp. 31–36. doi: 10.1145/980972.980978.

Speier, C., Valacich, J. S. y Vessey, I. (1999) "The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective", *Decision Sciences*,

30(2), pp. 337–360. doi: 10.1111/j.1540-5915.1999.tb01613.x.

Statista (2021) *Total data volume worldwide*.

Stumme, G., Wille, R. y Wille, U. (1998) “Conceptual knowledge discovery in databases using formal concept analysis methods”, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1510, pp. 450–458. doi: 10.1007/bfb0094849.

Sun, W. y Huang, C. (2020) “A carbon price prediction model based on secondary decomposition algorithm and optimized back propagation neural network”, *Journal of Cleaner Production*, 243(3), p. 118671. doi: 10.1016/j.jclepro.2019.118671.

The Pennsylvania State University (2019) *AIC vs. BIC – The Methodology Center*.

Tong, H. y Lim, K. S. (2009) “Threshold autoregression, limit cycles and cyclical data”, *Exploration of a Nonlinear World: An Appreciation of Howell Tong’s Contributions to Statistics*, 42(3), pp. 9–56. doi: 10.1142/9789812836281_0002.

Tsai, C. *et al.* (2014) “Data Mining for Internet of Things : A Survey”, *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, 16(1), pp. 77–97. doi: 10.1109/SURV.2013.103013.00206.

Wei, W. W. S. (2013) *Time Series Analysis*. Oxford University Press. doi: 10.1093/oxfordhb/9780199934898.013.0022.

Williams, G. J. y Huang, Z. (1996) “Modelling the KDD Process A Four Stage Process and Four Element Model”, *CSIRO*, pp. 1–8.

Wirth, R. (2000) “CRISP-DM : Towards a Standard Process Model for Data Mining”, *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, (24959), pp. 29–39. doi: 10.1.1.198.5133.

Woods, D. D. *et al.* (1999) “Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis”, *Proceedings of the Human Factors and Ergonomics Society*

Annual Meeting, 43(3), pp. 174–178. doi: 10.1177/154193129904300310.

Wright, R. (2008) “Data Visualization”, en *Software Studies*. The MIT Press, pp. 79–86. doi: 10.7551/mitpress/9780262062749.003.0011.

Yang, X. y Liu, B. (2019) “Uncertain time series analysis with imprecise observations”, *Fuzzy Optimization and Decision Making*, 18(3), pp. 263–278. doi: 10.1007/s10700-018-9298-z.

Yenidogan, I. *et al.* (2018) “Bitcoin Forecasting Using ARIMA and PROPHET”, en *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. IEEE, pp. 621–624. doi: 10.1109/UBMK.2018.8566476.

Zhang, P. G. (2003) “Time series forecasting using a hybrid ARIMA and neural network model”, *Neurocomputing*, 50, pp. 159–175. doi: 10.1016/S0925-2312(01)00702-0.

Zhang, Y. y Yang, Y. (2015) “Cross-validation for selecting a model selection procedure”, *Journal of Econometrics*, 187(1), pp. 95–112. doi: 10.1016/j.jeconom.2015.02.006.