



UNIVERSIDAD TECNOLÓGICA METROPOLITANA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INFORMÁTICA Y COMPUTACIÓN
ESCUELA DE INFORMÁTICA

MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL COMPORTAMIENTO DE CLIENTES EN CANAL WEB

TRABAJO DE TÍTULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN MENCIÓN
INFORMÁTICA

AUTORES:

GONZÁLEZ GÁRATE, CRISTÓBAL ANDRES
TAPIA RIQUELME, MARCELO IGNACIO
OYARCE TREJO, DIEGO ESTEBAN

PROFESORA GUÍA:

CASTRO OPAZO, PAULA

SANTIAGO - CHILE

2023

Índice general

Resumen	v
Abstract	vii
1 Presentación del proyecto	1
1.1 Descripción del trabajo de título	1
1.2 Objetivos	1
1.3 Alcances y Limitaciones	2
2 La empresa	4
2.1 Historia	4
2.2 Descripción general	6
2.3 Misión y visión	6
2.3.1 Misión	6
2.3.2 Visión	7
3 Marco teórico	8
3.1 Importancia de predecir el comportamiento del cliente en un sitio web	8
3.2 Comportamiento del cliente/afiliado en el canal web	9
3.2.1 Definición y relevancia del comportamiento del cliente para el negocio	9
3.2.2 Características del comportamiento del cliente en el canal web	11
3.2.3 Factores que afectan el comportamiento del cliente	12
3.3 Herramientas para la predicción del comportamiento del cliente en el canal web	13
3.3.1 Introducción a las herramientas de análisis de datos	13
3.3.2 Métodos, técnicas y tecnologías de análisis de datos	15

3.3.3	Modelos de predicción de comportamiento del cliente	16
3.3.4	Tabla comparativa de los modelos de predicción	35
3.3.5	Series Temporales	36
3.3.6	Redes LSTM Long Short-Term Memory	56
3.3.7	Metodología del proyecto	57
3.3.8	Metodología del sistema	59
4	Proceso ETL	63
4.1	Diseño Proceso ETL	63
4.1.1	Requisitos ETL	64
4.1.2	Identificación fuente de datos	64
4.1.3	Diseño del modelo de datos objetivo	65
4.1.4	Planificación de las transformaciones	65
4.1.5	Selección de herramientas	67
4.1.6	Construcción y prueba proceso ETL	68
4.1.7	Monitoreo proceso ETL	69
5	Exploratory Data Analysis (EDA)	71
5.1	Introducción al EDA	71
5.2	Recopilación de datos	72
5.3	Estadísticas descriptivas	73
	Referencias	77

Índice de figuras

2.1	Historia AFP Capital	5
3.1	Estructura de un árbol de decisión	28
3.2	Estructura de un random forest	33
3.3	Cantidad semanal de pasajeros que volaron en la clase económica de Ansett Airlines	37
3.4	Ventas mensuales de medicamentos antidiabéticos en Australia	39
3.5	Demanda electrica debido a las temperaturas en 2 ciudades en Australia	40
3.6	Producción trimestral de cerveza australiana	41
3.7	Coefficientes de autocorrelación de la producción de cerveza en Australia	43
3.8	ACF de la produccion trimestral de cerveza	43
3.9	Demanda mensual de electricidad de Australia	44
3.10	ACF demanda mensual de electricidad de Australia	44
3.11	Ruido blanco	45
3.12	ACF Ruido blanco	46
3.13	Aplicación de logartimos y diferenciación estacional a ventas de medicamentos antidiabéticos	49
3.14	Modelos autorregresivos con diferentes parámetros	51
3.15	Modelos de media móvil con diferentes parámetros	52
4.1	Construcción proceso ETL con Python.	69
5.1	Datos faltantes dataframes.	74

RESUMEN

El presente documento de Trabajo de Titulación tiene como objetivo mostrar la forma y el plan de trabajo que se utilizan a lo largo del proceso de desarrollo del proyecto propuesto. Además, se presentan los resultados del proceso investigativo que se ha realizado hasta la fecha, incluyendo estudios sobre el comportamiento de los clientes en canales web, modelos y algoritmos de predicción, y cómo desarrollar un proceso ETL.

El objetivo principal de este proyecto es analizar el comportamiento de los clientes de AFP Capital y sus preferencias de uso en un período de hasta 6 meses, con el fin de predecir futuras navegaciones personalizadas.

El proyecto consta de cuatro fases para su desarrollo. La primera fase abarca la planificación y el planteamiento de los antecedentes generales para la realización del proyecto. La segunda fase se centra en la investigación de la problemática en estudio, basándose en la situación actual planteada. La tercera fase abarca el modelamiento y desarrollo del proyecto, incluyendo el modelamiento de datos y el enfoque del proceso ETL. Esta fase también involucra el desarrollo del código que respaldará y ejecutará el modelo predictivo, mediante la construcción de bases de datos, APIs y la realización de pruebas para mitigar posibles errores encontrados. La cuarta y última fase concluye el desarrollo del proyecto y se enfoca en las conclusiones y recomendaciones, donde se presentarán las conclusiones obtenidas a lo largo del proceso y se elaborará un manual de usuario con las recomendaciones de uso.

Además, este proyecto se llevará a cabo bajo un marco de trabajo de desarrollo ágil, utilizando Scrum, y se utilizarán metodologías de análisis y minería de datos, como CRISP-DM y OSEMN. El entorno de desarrollo se basará en Python, junto con bibliotecas de análisis y minería de datos como Pandas y Numpy, y frameworks de desarrollo de APIs como Flask, Django y FastAPI.

La duración del proyecto será de dos semestres académicos, correspondientes a las asignaturas Título I y Título II. Los entregables incluirán un Informe Final de Trabajo de Título y el sistema (MVP) del proyecto propuesto.

Palabras clave: Afiliado, Administradora de Fondos de Pensiones, API (Application Programming Interfaces), EDA (Exploratory Data Analysis), Algoritmos de predicción, Algoritmos de clasificación, Modelos de predicción, ETL (Extract, Transform and Load), ARIMA (Modelo de Autorregresión integrada de media móvil), SARIMA (Modelo Estacional de Autorregresión integrada de media móvil), Redes Neuronales Artificiales (ANN), Redes LSTM (Long Short-Term Memory), Redes Neuronales Recurrentes (RNN).

ABSTRACT

The purpose of this thesis is to show the form and the work plan used throughout the development process of the proposed project. In addition, the results of the research process that has been carried out to date are presented, including studies on customer behavior in web channels, predictive models and algorithms, and how to develop an ETL process.

The main objective of this project is to analyze AFP Capital's customer behavior and usage preferences over a period of up to 6 months, in order to predict future personalized browsing.

The project consists of four phases for its development. The first phase covers the planning and general background for the realization of the project. The second phase focuses on the investigation of the problem under study, based on the current situation. The third phase covers the modeling and development of the project, including the data modeling and ETL process approach. This phase also involves the development of the code that will support and execute the predictive model, through the construction of databases, APIs and testing to mitigate possible errors found. The fourth and last phase concludes the development of the project and focuses on conclusions and recommendations, where the conclusions obtained throughout the process will be presented, and a ...

Keywords: Affiliate, Pension Fund Administrator, API (Application Programming Interfaces), EDA (Exploratory Data Analysis), Predictive Algorithms, Classification Algorithms, Predictive Models, ETL (Extract, Transform and Load) ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal AutoRegressive Integrated Moving Average), Artificial Neural Network (ANN), LSTM Networks (Long Short-Term Memory), Recurrent Neural Network (RNN).

CAPÍTULO 1: PRESENTACIÓN DEL PROYECTO

1.1. Descripción del trabajo de título

El trabajo de titulación se basa en un proyecto empresarial que tiene como objetivo procesar los registros de navegación del sitio web para afiliados de AFP Capital. El propósito principal es detectar comportamientos de los clientes y sus preferencias de uso, con el fin de personalizar las futuras experiencias de navegación. La lectura de los registros se realizará extrayendo la información desde Kibana, una plataforma basada en Elasticsearch, que registra la información a través de diversas APIs utilizadas en el sitio web.

Los elementos fundamentales del proyecto incluyen el análisis exploratorio de datos, extracciones, transformaciones, cargas, modelos de predicción y detección de preferencias. El objetivo final es desarrollar un modelo capaz de predecir el comportamiento de los clientes en el canal web.

1.2. Objetivos

Objetivo general

Analizar el comportamiento de los clientes y sus preferencias de uso en un período de hasta 6 meses, con el fin de predecir navegaciones futuras personalizadas.

Objetivos específicos

- Realizar una investigación sobre las herramientas utilizadas para la predicción del comportamiento de usuarios en un canal web.
- Realizar un análisis y estudio de los datos proporcionados por la empresa.

- Realizar un proceso de Extracción, Transformación y Carga (ETL) con la información de navegación web de los clientes de AFP Capital, con el objetivo de analizar su comportamiento dentro del sitio web privado.
- Desarrollar un modelo capaz de predecir el comportamiento de los clientes de AFP Capital, con el propósito de ofrecer navegaciones personalizadas en el futuro.
- Establecer recomendaciones de personalización basadas en los hallazgos del modelo de predicción, para las futuras navegaciones dentro del sitio web de AFP Capital.

1.3. Alcances y Limitaciones

Alcances

El proyecto contempla los siguientes alcances:

- Se analizará el comportamiento de los clientes de AFP Capital en su nuevo sitio web privado.
- El proyecto entregará un modelo capaz de predecir el comportamiento de los clientes de AFP Capital en el sitio web, así como una API que permita obtener recomendaciones de comportamiento personalizadas para un afiliado específico.

Limitaciones

El proyecto tiene las siguientes limitaciones:

- No se contará con acceso directo a las bases de datos de AFP Capital, por lo tanto, se trabajará con una muestra de datos.
- No se podrá acceder a información sensible de los clientes de AFP Capital, como los RUTs (Rol Único Tributario) u otra información personal identificable.

- El análisis se basará únicamente en datos cualitativos de la navegación web de los usuarios.

CAPÍTULO 2: LA EMPRESA

AFP Capital se erige como un actor destacado en su industria, con una trayectoria consolidada y un enfoque constante en la innovación y la excelencia operativa. Este capítulo del informe proporcionará una visión detallada de su historia, una descripción general, misión, visión y su papel en el mercado actual. A lo largo de estas páginas, exploraremos los aspectos clave que definen a esta empresa y su relevancia en el panorama empresarial actual.

2.1. Historia

La historia de AFP Capital se remonta a noviembre de 1980, cuando se implementó en Chile el sistema de pensiones de capitalización individual. El 16 de enero de 1981, se constituyó la sociedad Administradora de Fondos de Pensiones Santa María, la cual más tarde se transformaría en AFP Capital S.A. Desde sus inicios, la empresa se destacó por su filosofía de servicio, enfocada en satisfacer las necesidades y expectativas de sus afiliados.

En 1995, AFP Capital estableció la filial Santa María Internacional S.A., con el propósito de expandir su alcance y ofrecer servicios a personas naturales o jurídicas del extranjero, así como invertir en AFP o sociedades relacionadas con materias previsionales en otros países. Esta iniciativa consolidó la presencia de AFP Capital en el ámbito internacional y fortaleció su posición como una administradora de fondos de pensiones líder en la región.

En el año 2000, se produjo una transacción relevante en la historia de AFP Capital. ING Group adquirió Aetna Inc., incluyendo el 96,56 % de las acciones de AFP Capital S.A. Esta adquisición tuvo como objetivo reforzar la posición de liderazgo de AFP Capital en el mercado previsional chileno y contribuir a su crecimiento y desarrollo.

Posteriormente, en 2008, AFP Capital llevó a cabo una fusión con AFP Bansander, otra reconocida administradora de fondos de pensiones en Chile. Esta fusión permitió consolidar aún más las operaciones de AFP Capital y fortalecer su presencia en el país. A fines de 2011, Grupo SURA, una empresa líder en el negocio de pensiones en Latinoamérica, adquirió las operaciones de ING en la región. Esta adquisición llevó a AFP Capital a formar parte de Grupo SURA y a beneficiarse de su amplia experiencia y recursos, consolidándose como una compañía destacada en el mercado previsional latinoamericano.

En resumen, la historia de AFP Capital está marcada por su constante evolución, consolidación y liderazgo en el mercado de administración de fondos de pensiones en Chile. A lo largo de los años, ha demostrado su compromiso con la excelencia en la prestación de servicios previsionales y su capacidad de adaptación a los cambios y desafíos del entorno económico y regulatorio.

Figura 2.1: Historia AFP Capital



Fuente: AFP Capital. Recuperado de <https://www.afpcapital.cl/Quienes-Somos/Paginas/Historia.aspx>

2.2. Descripción general

AFP Capital es una destacada compañía chilena dedicada al negocio de pensiones y administración de fondos de pensiones. Forma parte de SURA, una reconocida empresa que ofrece servicios financieros y previsionales en Chile y otros países de América Latina. El enfoque principal de AFP Capital es proporcionar a sus afiliados asesoría personalizada y servicios diferenciados que les permitan alcanzar una mejor pensión al momento de su jubilación. La empresa se distingue por su compromiso con la optimización de la calidad de sus servicios, la entrega de información transparente y relevante a sus afiliados, y su solidez empresarial.

Con más de tres décadas de experiencia en el mercado, AFP Capital se ha posicionado como una de las principales administradoras de fondos de pensiones en Chile. Esto se debe en gran medida a su administración seria, responsable y eficiente en el manejo de los Fondos de Pensiones, así como a su enfoque prudente y estratégico en la inversión y gestión de los recursos.

La compañía cuenta con un equipo de colaboradores altamente capacitados y comprometidos, quienes contribuyen a la excelencia en la atención al cliente y al logro de los objetivos financieros de los afiliados. Además, AFP Capital se destaca por su constante innovación y adaptación a los cambios regulatorios y a las necesidades cambiantes de los afiliados, con el fin de ofrecer soluciones efectivas y satisfactorias en el ámbito de las pensiones.

2.3. Misión y visión

2.3.1. Misión

La misión de AFP Capital es: "Acompañamos a nuestros clientes, a través de una asesoría experta y diferenciadora en soluciones de ahorro para alcanzar su número, su Pensión, creciendo sustentablemente, desarrollando a nuestros colaboradores e integrándose responsablemente a la comunidad." (*AFP Capital*, 2023)

2.3.2. Visión

La visión de AFP Capital es: "Somos Guías, acompañamos a nuestros clientes a lograr sus sueños a través del ahorro." (*AFP Capital*, 2023)

CAPÍTULO 3: MARCO TEÓRICO

En el ámbito empresarial actual, la capacidad de predecir y comprender el comportamiento del cliente en un sitio web es esencial para el éxito. Este capítulo del informe se adentrará en la importancia estratégica de estas predicciones, analizará el comportamiento de los clientes y afiliados en el canal web y presentará las herramientas clave utilizadas para anticipar y optimizar la experiencia del usuario en línea.

3.1. Importancia de predecir el comportamiento del cliente en un sitio web

La predicción del comportamiento del cliente dentro de un entorno web implica la aplicación de técnicas y modelos analíticos para anticipar, en cierta medida, las posibles necesidades, acciones, preferencias y decisiones que un cliente pueda tomar mientras interactúa en una plataforma en línea o sitio web. En los últimos años, la predicción del comportamiento de los clientes ha sido de gran importancia para las empresas, ya que les permite anticiparse a las necesidades y preferencias de sus clientes, adaptando así sus productos y servicios para brindar una mayor satisfacción al cliente (Zheng, Thompson, Lam, Yoon, y Gnanasambandam, 2013).

La lealtad de los clientes es un valor clave para las empresas, ya que un cliente leal seguirá consumiendo los productos y servicios de la empresa. Por lo tanto, mejorar la experiencia del usuario aumenta la satisfacción del cliente, lo que a su vez genera un incremento en las ganancias de la empresa.

Según (Zheng y cols., 2013), la predicción del comportamiento del cliente ayuda a las empresas a identificar oportunidades de mejora y de mercado, además de respaldar la toma de decisiones informadas sobre estrategias de publicidad y marketing. El objetivo principal de predecir el comportamiento del cliente en un entorno web

es comprender y anticipar las acciones de los clientes con el fin de personalizar y mejorar la experiencia del usuario, y así aumentar la satisfacción y fidelidad de los clientes.

Las predicciones pueden abarcar diferentes aspectos del comportamiento de un cliente dentro de un canal web. En términos generales, existen cuatro tipos de predicciones que se pueden realizar. En primer lugar, están las predicciones de compras, donde se analizan los patrones de navegación, el historial de compras, las preferencias y las características demográficas del cliente para predecir sus compras futuras. Luego, se encuentra la predicción de clics, que busca anticipar los enlaces o elementos con los cuales un cliente interactuará en un sitio web, con el objetivo de mejorar la calidad del contenido y la usabilidad del sitio. Además, se encuentra la predicción de abandono de carrito, que permite identificar a aquellos clientes que agregan productos a un carrito de compra pero no completan el proceso de compra, con el fin de tomar acciones de recuperación o retención del cliente. Por último, está la predicción de retención de clientes, que busca predecir qué clientes están más propensos a abandonar o finalizar su relación con el sitio web, para poder implementar estrategias que aumenten su fidelización y retención.

3.2. Comportamiento del cliente/afiliado en el canal web

3.2.1. Definición y relevancia del comportamiento del cliente para el negocio

Considerando los modelos de negocio establecidos por las Administradoras de Fondos de Pensiones (AFP), surge la importancia de la figura del cliente. Según la Real Academia Española, un cliente es una persona que realiza una compra o utiliza los servicios ofrecidos por un profesional o empresa (Real Academia Española, s.f). Sin embargo, en el contexto de las AFP, los clientes se denominan afiliados, ya que contribuyen o están inscritos en un plan de pensiones (Rasekhi, Fard, y Kim, 2016).

El afiliado es el centro del negocio y su importancia radica principalmente en la

rentabilidad que aporta. Cada trabajador que decide afiliarse representa una ganancia, mientras que cada afiliado que decide desafilarse genera una pérdida. Además, la experiencia del servicio que brinda la AFP hacia el afiliado es crucial, ya que puede promover la marca si es positiva. En tercer lugar, el afiliado, al ser una fuente de ganancias para el modelo, puede contribuir al crecimiento de la empresa al darle su preferencia. Además, la experiencia del cliente y su retroalimentación son valiosas, ya que pueden proporcionar conocimientos sobre los puntos débiles y las áreas de mejora del sistema (Rodríguez, 2023).

Dentro de las diferentes funciones que tiene el cliente, en primer lugar, se encuentra el cliente como consumidor. Esta es una de las funcionalidades más tradicionales, ya que el objetivo intrínseco del cliente es consumir o contratar servicios. Como consumidor, adquiere un producto o servicio y lo utiliza para satisfacer una necesidad, lo que representa la principal fuente de ingresos para la empresa.

En segundo lugar, se encuentra el cliente como "prosumidor", es decir, alguien que consume y produce al mismo tiempo (Toffler, 1980). Además de consumir, el cliente también deja reseñas o realiza comentarios en lugares especializados, lo cual es útil para generar información que mejore la experiencia del servicio.

En tercer lugar, se considera al cliente como crítico, ya que si la experiencia del cliente es negativa, los comentarios y reseñas negativas que proporcione pueden tener un impacto constructivo o destructivo.

En cuarto lugar, el cliente es una pieza fundamental en el desarrollo de productos y servicios. Los comentarios de los clientes pueden guiar el desarrollo de servicios innovadores que se ajusten a las necesidades que ellos indican. En el caso específico de las AFP, esto se refiere a los afiliados.

En quinto lugar, el cliente se desempeña como evaluador de la experiencia. Relacionado con los puntos anteriores, la mejor manera de mejorar la experiencia del cliente es tener en cuenta sus comentarios sobre este aspecto, lo que puede marcar la diferencia frente a otras empresas competidoras en el mercado.

Por último, el cliente puede convertirse en un embajador eventual de la marca, es decir, puede promover el negocio mediante recomendaciones, comentarios y reseñas positivas.

3.2.2. Características del comportamiento del cliente en el canal web

Para comprender la experiencia y el comportamiento del cliente en un canal web, es importante reconocer la existencia del customer journey, el cual describe las distintas etapas por las que un cliente pasa al consumir un producto o servicio. Según (Lemon y Verhoef, 2016), estas etapas incluyen la conciencia, investigación, consideración, compra, uso y evaluación. La etapa de conciencia refiere a la identificación de una necesidad o problema que debe ser resuelto, mientras que la investigación implica la búsqueda de información por parte del cliente para encontrar posibles soluciones y comparar entre diferentes opciones disponibles. Luego, en la etapa de consideración, el cliente evalúa las alternativas y elige la que mejor se adapte a sus necesidades, lo que lleva a la etapa de compra, donde se realiza la contratación o adquisición del servicio seleccionado. Posteriormente, viene la etapa de uso, en la cual el cliente experimenta y evalúa la calidad, funcionalidad y experiencia del servicio. Por último, se encuentra la etapa de evaluación, en la cual el cliente emite un feedback voluntario, tanto positivo como negativo, sobre su experiencia satisfactoria o insatisfactoria. En resumen, las opciones disponibles en el canal web buscan hacer del customer journey una experiencia eficiente y agradable.

Para acceder al canal web de AFP Capital, es necesario ser afiliado y contar con una cuenta privada personal que incluya el RUT y contraseña. Una vez ingresado al canal web privado, los afiliados tienen a su disposición diversas opciones para satisfacer sus necesidades. Estas incluyen revisión del pago o no de la cotización mensual, la obtención de certificados de cotizaciones, afiliación, antecedentes previsionales y traspaso de fondos, así como certificados tributarios. Además, se pueden obtener certificados generales, como de residencia, suscripción de ahorro previsional voluntario (APV), cuenta 2, remuneraciones imponibles, periodos no cotizados y trabajo pesado. En el caso de afiliados pensionados, también se pueden obtener certificados de asignación familiar, calidad de pensionado, pensiones pagadas, pensión en trámite,

ingreso base y comprobante de pago de pensión. Además, es posible acceder a la cartola en línea. El canal web privado permite realizar el ahorro obligatorio y voluntario, inversiones, depósitos directos, consultar planillas de pagos y ver las comisiones cobradas como afiliado. También ofrece la opción de verificar el fondo de pensiones, los tipos de fondos disponibles (A, B, C, D, E) y sus porcentajes de rentabilidad, así como realizar cambios de fondo de pensiones y acceder a educación previsional. Además, se brinda la posibilidad de realizar giros en cuentas personales, acceder a rescates financieros y tramitar la pensión.

3.2.3. Factores que afectan el comportamiento del cliente

(Lemon y Verhoef, 2016) proponen que los principales factores que influyen en el comportamiento del usuario y su experiencia son los sensoriales, afectivos, cognitivos, puntos de contacto y externos. La experiencia sensorial se refiere a los aspectos perceptibles por los sentidos del cuerpo, como la vista, el olfato y el tacto. En cuanto a la experiencia afectiva, se debe considerar la emocionalidad del cliente como resultado de la experiencia con el producto o servicio. En el aspecto cognitivo, se refiere a los pensamientos, creencias y actitudes que el cliente puede tener hacia la compañía, el producto o el servicio (Lemon y Verhoef, 2016). Los puntos de contacto hacen referencia a las diferentes formas en que el cliente y la compañía interactúan, como la publicidad, el servicio al cliente, las redes sociales o las interacciones transaccionales. Por último, el factor externo se refiere al contexto actual, las condiciones socioeconómicas y otros factores que pueden afectar la experiencia del usuario y que están fuera del control de la compañía.

Dentro de los factores que pueden influir en el comportamiento de un cliente en el canal web, se encuentran principalmente la usabilidad y el diseño. En cuanto a la usabilidad, depende de siete características que garantizan una buena experiencia para el usuario. Según Sánchez (Sánchez, 2011), la accesibilidad, legibilidad, navegabilidad, facilidad de aprendizaje, velocidad de utilización, eficiencia del usuario y tasas de error del canal web influyen en la experiencia del usuario y en el feedback que este pueda brindar sobre el uso de los servicios.

Por otro lado, el diseño del sitio web depende de cinco características para

garantizar un buen contenido y estética, y lograr que el usuario encuentre lo que busca en el menor tiempo posible, es decir, eficiencia. El autor Walter Sánchez (Sánchez, 2011) indica que el diseño debe ser entendible, novedoso, comprensible, inteligente y atractivo, lo que permite acercar los contenidos de mejor manera al usuario y lograr una navegación más intuitiva. Estos factores son de gran importancia para que el usuario pueda encontrar el contenido que busca en el menor tiempo posible y tener una experiencia positiva al interactuar con la interfaz del sitio web.

3.3. Herramientas para la predicción del comportamiento del cliente en el canal web

3.3.1. Introducción a las herramientas de análisis de datos

En el entorno empresarial actual, la capacidad de tomar decisiones informadas y basadas en datos se ha vuelto fundamental para el éxito y la competitividad de las organizaciones. El análisis de datos desempeña un papel crucial en este proceso, permitiendo a las empresas obtener información valiosa a partir de grandes volúmenes de datos y utilizarla para comprender el comportamiento del cliente de manera más profunda y precisa. Esto resulta de suma importancia, ya que la calidad de las decisiones tomadas marca la diferencia entre el éxito y el fracaso (Contreras Arteaga y Sánchez Cotrina, 2019).

Dentro de las herramientas de análisis de datos, se destacan cuatro conceptos clave que han revolucionado la forma en que se procesan y se obtiene información de los datos: Business Intelligence, Big Data, Machine Learning y Data Mining. Estas herramientas proporcionan a las empresas la capacidad de extraer conocimientos y patrones significativos de los datos, lo que a su vez les permite tomar decisiones estratégicas más acertadas y personalizar sus estrategias de marketing y atención al cliente.

El Business Intelligence (BI) se refiere a la recopilación, análisis y presentación de datos empresariales para facilitar la toma de decisiones. Mediante el uso de diversas técnicas y herramientas, el BI permite a las empresas visualizar y comprender

mejor los datos de sus operaciones y clientes. Esto incluye la generación de informes, el análisis de tendencias, la monitorización de indicadores clave de rendimiento (KPI) y la creación de tableros de control interactivos. El BI ayuda a las organizaciones a identificar oportunidades, detectar áreas de mejora y optimizar su rendimiento en función de datos históricos y en tiempo real. Sobre la inteligencia de negocios, se ha determinado que cada implementación es única para cada proceso empresarial (Garcia-Estrella y Barón Ramírez, 2021).

El Big Data se refiere a la gestión y análisis de grandes volúmenes de datos, tanto estructurados como no estructurados, que superan la capacidad de las herramientas tradicionales de almacenamiento y procesamiento. El Big Data se caracteriza por las tres V's: Volumen (gran cantidad de datos), Velocidad (alta velocidad de generación y procesamiento de datos) y Variedad (diversidad de fuentes y formatos de datos). Para aprovechar el potencial del Big Data, las empresas emplean técnicas de procesamiento distribuido y herramientas específicas para el almacenamiento, procesamiento y análisis de estos datos masivos. El análisis de Big Data permite identificar patrones, tendencias y correlaciones ocultas en los datos, lo que brinda información valiosa para entender y anticipar el comportamiento del cliente.

El Machine Learning (aprendizaje automático) es una rama de la inteligencia artificial que permite a los sistemas informáticos aprender y mejorar automáticamente a partir de la experiencia sin ser programados explícitamente. En lugar de basarse en una analítica descriptiva, el Machine Learning ofrece una analítica predictiva (Garcia-Estrella y Barón Ramírez, 2021). Mediante algoritmos y modelos, el Machine Learning permite a las empresas analizar grandes conjuntos de datos y detectar patrones complejos en el comportamiento del cliente. Esto permite realizar predicciones y recomendaciones personalizadas, así como automatizar tareas y procesos, lo que mejora la eficiencia operativa y la experiencia del cliente.

El Data Mining (minería de datos) se refiere al proceso de descubrir información valiosa, patrones y relaciones desconocidas en grandes conjuntos de datos. Utilizando técnicas estadísticas y algoritmos avanzados, el Data Mining permite identificar correlaciones y tendencias ocultas en los datos, lo que ayuda a las empresas a comprender mejor el comportamiento del cliente y tomar decisiones más acertadas. Esta

herramienta es especialmente útil para la segmentación de clientes, la detección de fraudes, la recomendación de productos y la personalización de ofertas.

3.3.2. Métodos, técnicas y tecnologías de análisis de datos

En el análisis de datos para predecir el comportamiento del cliente, se utilizan una variedad de métodos, técnicas y tecnologías que permiten procesar y analizar grandes volúmenes de información con el fin de obtener información valiosa. Estas herramientas proporcionan a las empresas y organizaciones la capacidad de comprender mejor a sus clientes, identificar patrones y tendencias, y tomar decisiones estratégicas más acertadas.

Entre los métodos y modelos más utilizados se encuentran la regresión logística, que permite predecir la probabilidad de que un cliente realice una determinada acción o tome una decisión; el clustering, que agrupa a los clientes en segmentos o categorías similares con características y comportamientos comunes; los árboles de decisión, que representan un conjunto de reglas lógicas para clasificar a los clientes en diferentes grupos; el Random Forest, que combina múltiples árboles de decisión para mejorar la precisión de las predicciones; y el Gradient Boosting Machine, que utiliza múltiples modelos de aprendizaje débiles para construir un modelo más robusto y preciso.

Además de los métodos y modelos, existen diversas técnicas que se aplican en el análisis de datos para predecir el comportamiento del cliente. Entre ellas se encuentran las redes neuronales artificiales (ANN), que son modelos inspirados en el funcionamiento del cerebro humano y se utilizan para reconocer patrones y realizar predicciones complejas; y el Support Vector Machine (SVM), que es un algoritmo de aprendizaje automático utilizado para clasificar y predecir datos.

En cuanto a las tecnologías utilizadas en el análisis de datos, se destacan diversas herramientas y lenguajes de programación. Algunas de las más populares son Tableau, que permite visualizar y explorar los datos de manera interactiva; Python, con bibliotecas como Pandas, NumPy y Scikit-learn, que ofrecen una amplia gama de funciones y algoritmos para el análisis de datos; R, con paquetes como dplyr, caret y randomForest, que brindan herramientas estadísticas y de aprendizaje automáti-

co; Apache Spark, que permite procesar y analizar grandes volúmenes de datos de manera distribuida; KNIME y RapidMiner, que son plataformas de análisis de datos visuales; y QlikView y Power BI, que son herramientas de visualización de datos y creación de tableros de control.

3.3.3. Modelos de predicción de comportamiento del cliente

En la era digital, los modelos de predicción de comportamiento del cliente son un recurso fundamental para las empresas que buscan tomar decisiones informadas y personalizar sus estrategias. Este subcapítulo del informe se adentrará en los diversos modelos utilizados para anticipar las acciones y preferencias de los clientes, destacando su relevancia en la toma de decisiones estratégicas y la mejora de la experiencia del usuario en el entorno online.

- **Modelos de regresión logística**

La regresión logística corresponde a un algoritmo de aprendizaje automático supervisado que es empleado para resolver problemas de clasificación. Si bien, su nombre contiene “regresión”, en realidad corresponde a un método de clasificación.

Se da uso a la regresión logística cuando la variable de respuesta o variable objetivo es categórica. En lugar de predecir un valor numérico como en la regresión lineal, la regresión logística estima la probabilidad de que una observación pertenezca a una categoría específica.

Los modelos de regresión logística se basan en la función logística, también conocida como función sigmoide, que mapea cualquier valor real a un rango entre 0 y 1. La función sigmoide tiene la siguiente forma matemática:

$$f(z) = \frac{1}{(1 + e^{-z})}$$

En la regresión logística, se ajusta un modelo lineal a los datos de entrada y se aplica la función sigmoide al resultado para obtener la probabilidad de

pertenencia a una clase. La ecuación del modelo se expresa como:

$$p(y = 1|x) = \frac{1}{(1 + e^{-(b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n)})}$$

Donde:

$p(y=1|x)$ es la probabilidad condicional de que la variable de respuesta sea igual a 1 dada la entrada x .

$b_0, b_1, b_2, \dots, b_n$ son los coeficientes del modelo que se ajustan durante el proceso de entrenamiento.

x_1, x_2, \dots, x_n son los valores de las variables de entrada.

El proceso de ajuste de la regresión logística implica encontrar los mejores valores para los coeficientes del modelo con la finalidad de maximizar la verosimilitud de los datos observados. Esto se puede hacer mediante métodos numéricos como la maximización de la función de verosimilitud o mediante algoritmos de optimización como el gradiente descendente.

Una vez entrenado el modelo, se puede utilizar para hacer predicciones clasificando nuevas observaciones según la probabilidad estimada. Por ejemplo, si la probabilidad estimada de pertenencia a una clase es superior a un umbral (generalmente 0.5), se clasificará como perteneciente a esa clase.

Para nuestro caso en particular, puede ser utilizado el modelo de regresión logística para predecir el comportamiento de usuarios en un canal web, para ello se necesitaría tener datos históricos que contengan información relevante sobre el comportamiento pasado de los usuarios y las variables predictoras asociadas. Estas variables predictoras pueden incluir características demográficas, patrones de uso del sitio web o aplicación, historial de compras, interacciones anteriores, entre otros.

Una vez que se tienen los datos y las variables predictoras, se puede entrenar un modelo de regresión logística utilizando técnicas de ajuste como la maximización de la verosimilitud o el gradiente descendente. Una vez entrenado el modelo, puede ser utilizado para predecir el comportamiento futuro de los usuarios en función de nuevas observaciones o datos entrantes.

Es importante tener en consideración que la calidad de las predicciones dependerá de la calidad de los datos utilizados para entrenar el modelo y de la selección adecuada de las variables predictoras. Además, es fundamental realizar una validación adecuada del modelo utilizando técnicas como la validación cruzada o la separación de conjuntos de entrenamiento y prueba para evaluar su rendimiento y generalización en datos no vistos.

- **Ventajas de los modelos de regresión logística**

- **Interpretación de resultados:** La regresión logística proporciona coeficientes que indican la dirección y la magnitud de la relación entre las variables predictoras y la variable de respuesta. Esto permite interpretar el efecto relativo de cada variable en la probabilidad de pertenecer a una clase específica.
- **Manejo de variables independientes categóricas:** La regresión logística puede manejar tanto variables independientes continuas como categóricas. Incluso puede manejar variables categóricas con más de dos categorías mediante técnicas como la codificación de variables ficticias.
- **Estimación de probabilidades:** La regresión logística estima la probabilidad de pertenencia a una clase específica en lugar de simplemente clasificar observaciones en categorías. Esto es útil cuando se necesita una medida de certeza o riesgo asociado con la clasificación.
- **Buena capacidad de generalización:** La regresión logística puede funcionar bien con conjuntos de datos pequeños o moderados, y es

menos propensa al sobreajuste en comparación con otros algoritmos más complejos. Esto la hace adecuada para aplicaciones con muestras limitadas.

- **Desventajas de los modelos de regresión logística**

- **Linealidad de la relación:** La regresión logística asume una relación lineal entre las variables predictoras y la probabilidad logarítmica de la variable de respuesta. Si existe una relación no lineal, la regresión logística puede no ajustarse adecuadamente o requerir transformaciones adicionales de las variables.
- **Sensible a valores atípicos y datos faltantes:** Los valores atípicos o datos faltantes pueden afectar negativamente el rendimiento de la regresión logística. Es necesario manejarlos adecuadamente para evitar sesgos o imprecisiones en los resultados.
- **Suposición de independencia:** La regresión logística asume que las observaciones son independientes entre sí. Si hay dependencias o correlaciones entre las observaciones, la precisión de los resultados puede verse comprometida.
- **No apto para problemas no lineales:** Si existe una relación compleja y no lineal entre las variables predictoras y la variable de respuesta, la regresión logística puede no ser el modelo más adecuado. En tales casos, se pueden requerir técnicas más avanzadas, como modelos no lineales o de aprendizaje profundo.

- **Modelos de recomendación**

Los modelos de recomendación son algoritmos y técnicas utilizados en sistemas de recomendación para ofrecer sugerencias personalizadas a los usuarios. Estos modelos se utilizan en una amplia gama de aplicaciones, como plataformas de comercio electrónico, servicios de streaming de música y video, redes sociales y más (Elizabeth, 2023).

El objetivo de un modelo de recomendación es predecir o sugerir elementos que sean relevantes o interesantes para un usuario en particular, basándose en su historial de preferencias, comportamiento pasado o en información de usuarios similares. Estos modelos aprovechan el poder del aprendizaje automático y la minería de datos para analizar patrones y relaciones en grandes conjuntos de datos.

Existen varios tipos de modelos de recomendación, entre los más comunes se encuentran (Vatsal, 2021):

- Filtrado colaborativo: Este enfoque se basa en la idea de que si a un grupo de usuarios con preferencias similares les gusta un conjunto de elementos, entonces a un usuario nuevo con características similares también le podrían gustar esos elementos. El filtrado colaborativo utiliza la información de las interacciones pasadas de los usuarios (por ejemplo, clasificaciones o historial de compras) para generar recomendaciones.
- Filtrado basado en contenido: Este enfoque utiliza información sobre las características y atributos de los elementos para recomendar otros elementos similares. Por ejemplo, en un servicio de streaming de música, se pueden recomendar canciones o artistas similares a los que un usuario ha escuchado anteriormente en función de género, estilo o letras.
- Modelos híbridos: Estos modelos combinan múltiples enfoques, como filtrado colaborativo y basado en contenido, para aprovechar sus fortalezas y proporcionar recomendaciones más precisas y personalizadas.

Los modelos de recomendación se construyen utilizando técnicas de aprendizaje automático, como regresión logística, árboles de decisión, redes neuronales o algoritmos de factorización matricial. Estos modelos se entrenan utilizando conjuntos de datos históricos que contienen información sobre las preferencias y elecciones de los usuarios, y luego se aplican en tiempo real para generar recomendaciones en función de nuevos datos.

- **Ventajas de los modelos de recomendación**

- **Personalización:** Los modelos de recomendación ofrecen sugerencias personalizadas a los usuarios, lo que mejora la experiencia del usuario y facilita la búsqueda de productos o contenido relevante.
- **Descubrimiento de nuevos elementos:** Los modelos de recomendación pueden ayudar a los usuarios a descubrir nuevos elementos que podrían ser de su interés, ampliando así sus opciones y experiencias.
- **Mejora de la retención y fidelidad de los usuarios:** Al proporcionar recomendaciones precisas y relevantes, los modelos de recomendación pueden aumentar la satisfacción del usuario, mejorar la retención y fomentar la fidelidad a la plataforma o servicio.
- **Eficiencia en la toma de decisiones:** Los usuarios pueden ahorrar tiempo y esfuerzo al recibir sugerencias personalizadas, lo que les ayuda a tomar decisiones más rápidas y eficientes.

- **Desventajas de los modelos de recomendación**

- **Sesgo y burbujas de filtro:** Los modelos de recomendación pueden verse afectados por el sesgo inherente en los datos de entrenamiento y pueden crear burbujas de filtro, limitando la diversidad y la exposición a nuevas ideas o perspectivas.
- **Fracaso en captar preferencias cambiantes:** Los modelos de recomendación pueden tener dificultades para captar las preferencias cambiantes de los usuarios a medida que sus gustos y necesidades evolucionan con el tiempo.
- **Problemas de inicio en frío:** Los modelos de recomendación pueden tener dificultades para ofrecer recomendaciones precisas para nuevos usuarios o elementos que tienen una falta de información histórica.

- **Privacidad y preocupaciones éticas:** Los modelos de recomendación recopilan y utilizan datos de los usuarios, lo que puede plantear preocupaciones de privacidad y cuestiones éticas relacionadas con el manejo de la información personal.

■ Modelos de series temporales

Los modelos de series temporales son técnicas utilizadas para analizar y predecir datos secuenciales que están organizados en función del tiempo. En una serie temporal, los datos se registran en intervalos regulares (como horas, días, meses, etc.) y cada punto de datos está asociado con una marca de tiempo (Ajitesh, 2023).

El objetivo principal de los modelos de series temporales es comprender y capturar los patrones, tendencias y estacionalidad en los datos a lo largo del tiempo, y utilizar esta información para hacer predicciones futuras. Estos modelos son ampliamente utilizados en diversos campos, como la economía, las finanzas, la meteorología, la demanda de productos, la planificación de inventario y más.

Los modelos de series temporales se basan en la suposición de que los datos pasados pueden proporcionar información útil para predecir el futuro. Algunos de los modelos más comunes utilizados en el análisis de series temporales son (Ajitesh, 2023):

- **Media móvil (MA):** Este modelo estima el valor futuro de la serie temporal en función de un promedio de los errores pasados. Se utiliza para capturar patrones aleatorios o no sistemáticos en los datos.
- **Autoregresión (AR):** Este modelo estima el valor futuro de la serie temporal en función de valores pasados de la propia serie. Se utiliza para capturar la dependencia de la serie en sí misma a lo largo del tiempo.
- **Autoregresión de media móvil (ARMA):** Este modelo combina los enfoques AR y MA para capturar tanto la dependencia de la serie en sí

misma como los patrones aleatorios.

- **Autoregresión integrada de media móvil (ARIMA):** Este modelo amplía el modelo ARMA al considerar también las diferencias entre los valores de la serie temporal. Se utiliza para capturar tendencias y estacionalidad en los datos.

Además de estos modelos clásicos, también se utilizan enfoques más avanzados, como los modelos de espacio de estados, los modelos de suavizado exponencial y los modelos de redes neuronales recurrentes (RNN), que pueden capturar relaciones más complejas y no lineales en los datos de series temporales.

Es importante destacar que el análisis de series temporales requiere un enfoque cuidadoso para la selección del modelo, la identificación de patrones y la evaluación de la precisión de las predicciones. Además, se deben tener en cuenta factores como la estacionalidad, la estacionariedad de la serie y la presencia de datos faltantes o valores atípicos para obtener resultados confiables.

- **Ventajas de los modelos de series temporales**

- **Captura de patrones temporales:** Los modelos de series temporales pueden capturar patrones, tendencias y estacionalidad en los datos a lo largo del tiempo. Esto permite comprender mejor la dinámica de los datos y hacer predicciones más precisas.
- **Predicciones a corto plazo:** Los modelos de series temporales son adecuados para hacer predicciones a corto plazo, ya que utilizan la información histórica para predecir los valores futuros. Esto es especialmente útil en aplicaciones donde se necesita anticipar eventos próximos, como demanda de productos o pronóstico del clima.
- **Utilización de datos secuenciales:** Los modelos de series temporales aprovechan la estructura secuencial de los datos y utilizan la información de los puntos anteriores para hacer predicciones en el si-

guiente punto. Esto permite tener en cuenta la dependencia temporal en los datos y obtener resultados más precisos.

- **Flexibilidad en la elección del modelo:** Existen diferentes tipos de modelos de series temporales que se pueden utilizar según la naturaleza de los datos y los patrones presentes. Esto proporciona flexibilidad para seleccionar el modelo más adecuado para el problema específico.

- **Desventajas de los modelos de series temporales**

- **Sensibilidad a datos faltantes o valores atípicos:** Los modelos de series temporales pueden verse afectados negativamente por la presencia de datos faltantes o valores atípicos. Estos pueden distorsionar los patrones y afectar la precisión de las predicciones.
- **Dificultad con tendencias no lineales:** Los modelos de series temporales asumen a menudo que las relaciones son lineales o pueden ser capturadas por modelos lineales. Si hay tendencias no lineales en los datos, los modelos lineales pueden no ajustarse adecuadamente y se pueden requerir enfoques más avanzados.
- **Necesidad de datos históricos adecuados:** Los modelos de series temporales requieren una cantidad suficiente de datos históricos para hacer predicciones precisas. En ausencia de datos suficientes, los modelos pueden tener dificultades para capturar patrones y generar resultados confiables.
- **Problemas con cambios estructurales:** Si hay cambios estructurales significativos en los datos de series temporales (por ejemplo, cambios en la estacionalidad o en los patrones), los modelos de series temporales pueden tener dificultades para adaptarse y pueden requerir ajustes manuales.

■ Modelos de atribución

Los modelos de atribución permiten predecir el recorrido que los clientes seguirán al momento de concretar una compra. Este recorrido puede contener las redes sociales, el uso del sitio web del vendedor, el correo electrónico, entre otros. Los modelos de atribución permiten determinar el impacto que tiene el uso de las acciones para el sistema de marketing (Baker, 2023). Este tipo de modelo permite darle mayor importancia a los canales de marketing y a los puntos de contacto que existen entre el cliente y el vendedor, que llevaron al cliente a realizar una compra.

Al asignar crédito a sus canales de marketing y puntos de contacto, se puede aumentar la posibilidad de que los clientes logren concretar una compra, esto a través de la identificación de las áreas del recorrido del comprador que se puedan mejorar, la determinación del retorno de la inversión para cada canal o punto de contacto, el descubrimiento de las áreas más efectivas para gastar el presupuesto de marketing y la adaptación de las campañas de marketing y muestra de contenido totalmente personalizado por clientes (Baker, 2023)

Existen variados tipos de modelos de atribución, todos tienen el mismo procedimiento de asignar crédito a los canales y punto de contacto, cada uno de estos tipos de modelo le atribuyen un peso distinto a cada canal y punto de contacto (Baker, 2023). Los modelos a continuación son los más aptos para lograr la predicción del comportamiento de un cliente:

- **Modelo de atribución Multi-Touch:** Este modelo demuestra ser poderoso ya que tiene en cuenta todos los canales y puntos de contacto con los que los clientes interactúan a lo largo de su camino al concretar una compra. Deja en evidencia cuáles de los canales y punto de contacto fueron más influyentes y de cómo estas trabajaron en conjunto para influenciar al cliente.
- **Modelo de atribución Lineal:** Corresponde a un tipo de modelo de atribución Multi-Touch que le entrega el mismo peso a cada uno de los canales y puntos de contacto con los que el cliente interactúa en su camino

al concretar una compra.

- **Modelo de atribución Time-Decay:** También llamado modelo de atribución de declive en el tiempo, además de considerar todos los puntos de contacto, también considera el tiempo que cada uno de estos puntos de contacto ocurrió, por lo que, los puntos de contacto o interacciones que sucedieron más cercano al momento en que se concretó la compra reciben mayor peso.
- **Ventajas de los modelos de atribución**
 - **Facilita el rastrear de mejor manera el paso a paso del cliente:** Esto gracias a la atención que se le entrega a cada canal y punto de contacto con el cual el cliente interactúa a la hora de concretar una compra.
 - **Permiten mayor personalización de rastreo de los clientes:** Al saber que canales y punto de contacto tiene cada uno de los clientes, se puede llegar a entregar una experiencia personalizada a cada uno de los clientes.
 - **Comprender la contribución de cada canal y punto de contacto:** Permite comprender como cada canal y punto de contacto contribuye a lograr los objetivos comerciales. Siendo de gran ayuda para identificar como asignar los recursos de manera mas efectiva y lograr optimizar las estrategias.
 - **Identificar canales y puntos de contacto de alto rendimiento:** Un modelo de atribución puede revelar qué canales o puntos de contacto tienen un mayor interacción con los clientes en términos de generación de resultados. Esto permite a las empresas enfocar sus recursos en los canales más efectivos y maximizar su retorno de inversión.

- **Desventajas de los modelos de atribución**

- **Poseen una mayor complejidad que los otros modelos:** La implementación de un modelo de atribución puede ser compleja y requerir un enfoque personalizado según las necesidades y características de cada empresa. Además, no hay un modelo de atribución único que sea universalmente aceptado, lo que puede generar falta de consenso y confusión en la industria.
- **La interpretación de los resultados puede ser subjetiva:** La interpretación de los resultados de un modelo de atribución puede estar sujeta a la interpretación y suposiciones del analista. Diferentes personas pueden llegar a conclusiones diferentes basadas en los mismos resultados, lo que puede generar cierta subjetividad en la interpretación de los datos.
- **Poseen limitaciones en la medición del seguimiento:** El modelo de atribución depende de la disponibilidad y calidad de los datos. Si los datos son limitados o imprecisos, los resultados del modelo pueden no ser confiables o representativos de la realidad.

- **Modelos de árboles de decisión**

Los árboles de decisión son modelos de aprendizaje supervisado que se utilizan para predecir a qué clase o categoría pertenece un caso conocido mediante uno o más atributos. Estos modelos se construyen utilizando un algoritmo llamado *partición binaria recursiva*. Durante el entrenamiento, el algoritmo realiza divisiones en un subconjunto de los datos basadas en decisiones asociadas a variables conocidas, generando así dos nuevos subconjuntos. Este proceso se repite de manera recursiva hasta alcanzar un punto de terminación predefinido, lo que resulta en la creación del clasificador basado en árbol de decisión. Luego, cada nuevo dato, que posee atributos conocidos, sigue las ramificaciones del árbol siguiendo las reglas y decisiones generadas durante el proceso de entrenamiento.

Como se mencionó anteriormente, un árbol de decisión realiza tareas de clasificación. Un clasificador es un algoritmo que nos permite asignar sistemáticamente una clase a cada uno de los casos presentados.

[illegible]

En la figura anterior se puede visualizar la estructura que posee un árbol de decisión, en este se aprecia como actua el algoritmo de partición binaria mencio-

nado al comienzo, tomando un conjunto y separandolo en subconjuntos hasta llegar a un final previamente establecido.

Para estimar la precisión de un clasificador, se calcula la tasa de error de clasificación verdadera. Esta tasa se obtiene evaluando un conjunto de valores X a los que el clasificador asigna una clase incorrecta, y se divide por el total de valores en X . Idealmente, se debería conocer la clase de todos los casos en el universo antes del entrenamiento, o en su defecto, de una muestra de tamaño similar al universo. Sin embargo, en la mayoría de los casos reales, no se dispone de todos los datos del universo, por lo que se trabaja con una muestra y se estima la tasa de error mencionada anteriormente utilizando *estimadores internos*.

- **Ventajas de los árboles de decisión**

- **Interpretabilidad:** Los árboles de decisión son fácilmente interpretables y comprensibles para los humanos. La estructura del árbol se puede visualizar de manera intuitiva, lo que permite comprender cómo se toman las decisiones y qué atributos son más relevantes para la clasificación.
- **Facilidad de uso:** La construcción y el uso de un árbol de decisión son relativamente sencillos en comparación con otros algoritmos de aprendizaje automático más complejos. No requieren una preparación exhaustiva de los datos ni un procesamiento previo complicado. Además, los árboles de decisión pueden manejar datos numéricos y categóricos sin requerir transformaciones adicionales, lo que simplifica el flujo de trabajo de modelado.
- **Capacidad para manejar datos faltantes y variables irrelevantes:** Los árboles de decisión tienen la capacidad de manejar datos faltantes en los atributos de forma natural. Durante la construcción del árbol, si un atributo tiene valores faltantes, el modelo puede utilizar otros atributos para tomar decisiones sin requerir imputación de

datos. Además, los árboles de decisión son resistentes a variables irrelevantes, lo que significa que pueden ignorar atributos que no aportan información útil para la clasificación.

- **Flexibilidad y robustez:** Los árboles de decisión pueden manejar tanto problemas de clasificación como de regresión. Además, son capaces de capturar relaciones no lineales entre los atributos y la variable objetivo. Aunque cada árbol individual puede ser susceptible al sobreajuste, se pueden aplicar técnicas de regularización, como la poda, para mejorar la generalización y evitar el sobreajuste.
- **Eficiencia en tiempo de entrenamiento y predicción:** Los árboles de decisión tienen tiempos de entrenamiento y predicción rápidos, ya que solo implican la evaluación de una serie de reglas de decisión. Aunque el tiempo de construcción puede ser mayor para conjuntos de datos grandes, una vez construido, el árbol puede ser utilizado eficientemente para hacer predicciones en tiempo real.

- **Desventajas de los árboles de decisión**

- **Sensibilidad a cambios pequeños en los datos:** Los árboles de decisión son muy sensibles a cambios pequeños en los datos de entrenamiento. Una modificación mínima en los datos de entrada puede dar lugar a un árbol de decisión completamente diferente. Esto puede hacer que el modelo sea inestable y su rendimiento pueda variar significativamente.
- **Tendencia al sobreajuste:** Los árboles de decisión tienen la capacidad de adaptarse demasiado a los datos de entrenamiento. Si no se controla adecuadamente, el árbol puede memorizar el ruido o las fluctuaciones aleatorias en los datos de entrenamiento, lo que puede resultar en un mal rendimiento en datos nuevos y no vistos. La poda y otras técnicas de regularización se utilizan para mitigar este problema.

- **Limitaciones en la representación de relaciones complejas:** Aunque los árboles de decisión pueden capturar relaciones no lineales entre atributos y la variable objetivo, pueden tener dificultades para representar relaciones complejas que requieren una combinación de múltiples atributos. Las decisiones tomadas en cada nodo se basan en un solo atributo, lo que puede limitar su capacidad para modelar interacciones más sofisticadas.
- **Propensión a sesgos en los datos de entrenamiento:** Los árboles de decisión pueden verse afectados por sesgos en los datos de entrenamiento, especialmente cuando hay desequilibrios en las clases o falta representación de ciertas categorías. Esto puede resultar en una clasificación desigual o inexacta en casos minoritarios o poco representados.

■ Modelo Random Forest

El algoritmo random forest corresponde a un algoritmo empleado en machine learning registrado por Leo Breiman y Adele Cutler (IBM., 2023), este combina la salida de múltiples árboles de decisión para llegar a un resultado. El uso de random forest se ha hecho popular a causa de su facilidad de uso y flexibilidad, ya que puede ser empleado para problemas de clasificación y regresión.

El random forest se encuentra formado por varios árboles de decisión, los cuales son propensos a tener problemas como sesgos o sobreajuste, pero cuando se trata con una gran cantidad de árboles se logra llegar a resultados más precisos. "Mientras que los árboles de decisión consideran todas las posibles divisiones de características, los bosques aleatorios solo seleccionan un subconjunto de esas características." (IBM., 2023)

Random forest cuenta con tres hiperparámetros principales que se deben de configurar antes de iniciar el entrenamiento (IBM., 2023):

- Tamaño del nodo.

- Cantidad de árboles de decisión.
- Cantidad de características muestreadas.

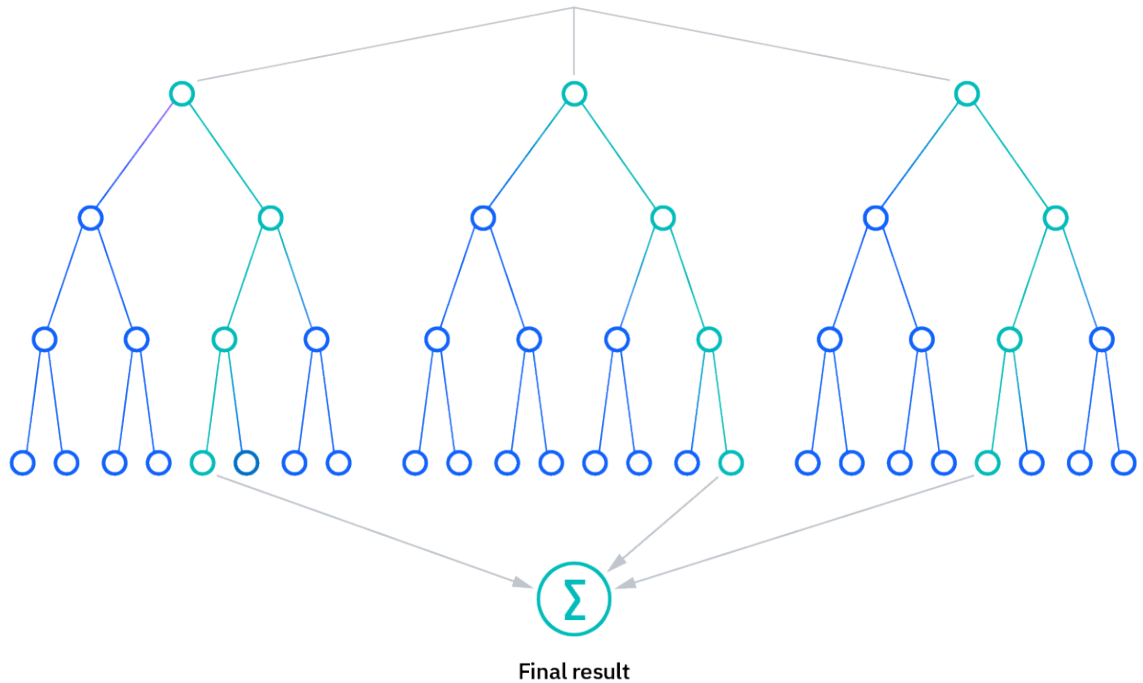
El algoritmo se encuentra compuesto de un conjunto de árboles de decisión, cada árbol del conjunto se encuentra compuesto de una muestra de datos, la cual proviene de un conjunto de entrenamiento con reemplazo, llamada muestra de arranque (IBM., 2023).

A partir de la muestra de entrenamiento, se extrae un porcentaje para reservarlos como datos de prueba, los cuales se conocen como muestra fuera de la bolsa (oob). Luego, se inyecta otra instancia de aleatoriedad mediante el agrupamiento de características, lo que agrega más diversidad al conjunto de datos y reduce la correlación entre los árboles de decisión (IBM., 2023).

En un Random Forest, el proceso de predicción puede variar según el tipo de problema que se esté abordando (IBM., 2023). En el caso de tareas de regresión, se utiliza un enfoque de promediado, donde las predicciones de los árboles de decisión individuales se promedian para obtener el valor final de la predicción. Esto proporciona una estimación más precisa y estable del resultado deseado. Por otro lado, en tareas de clasificación, se utiliza un enfoque de votación mayoritaria. Cada árbol de decisión emite su propia predicción y la clase que obtiene la mayoría de votos se selecciona como la clase predicha. Esto permite tomar una decisión conjunta basada en las opiniones de múltiples árboles, lo que puede mejorar la precisión en la clasificación de las muestras.

Finalmente, la muestra extraída en un comienzo, la muestra fuera de la bolsa (oob) será utilizada para realizar una validación cruzada, finalizando la predicción.

Figura 3.2: Estructura de un random forest



Fuente: IBM. Recuperado de <https://www.ibm.com/mx-es/topics/random-forest>

- **Ventajas de random forest**

- **Riesgo reducido de sobreajuste:** Los árboles de decisión corren el riesgo de sobre ajustarse, ya que tienden a ajustar todas las muestras que se encuentran dentro de los datos de entrenamiento. Sin embargo, cuando hay una gran cantidad de árboles de decisión dentro del random forest, el clasificador no será capaz de ajustarse demasiado al modelo, ya que el promedio de los árboles no correlacionados logra reducir la varianza general y el error de predicción.
- **Aporta Flexibilidad:** Debido a su capacidad para abordar con gran precisión tanto tareas de regresión como de clasificación, el método conocido como random forest es ampliamente utilizado por los científicos de datos. Además, su capacidad de agrupar características

lo convierte en una herramienta eficaz para estimar valores faltantes, manteniendo la precisión incluso cuando falta parte de los datos.

- **Importancia de la característica fácil de determinar:** El random forest ofrece una forma conveniente de evaluar la importancia o contribución de las variables en un modelo. Existen varias formas de medir la importancia de las características. Por lo general, se utilizan el índice de Gini y la disminución media de impurezas (MDI) para evaluar cuánto afecta la exclusión de una variable específica a la precisión del modelo. Sin embargo, otra medida de importancia es la importancia de permutación, también conocida como precisión de disminución media (MDA). La MDA determina la disminución promedio en la precisión al permutar de forma aleatoria los valores de las características en las muestras out-of-bag (muestras que no se utilizan en el proceso de entrenamiento).

- **Desventajas de random forest**

- **Proceso que requiere mucho tiempo:** Debido a que los algoritmos de random forest son capaces de manejar conjuntos de datos extensos, suelen ofrecer predicciones más precisas. Sin embargo, es importante tener en cuenta que el procesamiento de datos puede volverse lento, ya que se deben calcular los datos para cada árbol de decisión de forma individual.
- **Requiere más recursos:** Debido a que los random forest procesan conjuntos de datos más grandes, es cierto que se requieren más recursos para almacenar dichos datos. El aumento en el tamaño del conjunto de datos implica una mayor necesidad de memoria y capacidad de almacenamiento para garantizar un funcionamiento eficiente del algoritmo.
- **Más Complejo:** La interpretación de la predicción de un solo árbol de decisiones resulta más sencilla en comparación con la interpreta-

ción de un conjunto de árboles de decisión.

3.3.4. Tabla comparativa de los modelos de predicción

La tabla comparativa proporciona información sobre los modelos planteados anteriormente, presentando de forma resumida sus Ventajas, Desventajas y Aplicaciones comunes de los modelos. Permitiendo tener una visión general de las características y consideraciones clave de cada modelo.

Modelo	Ventajas	Desventajas	Aplicaciones
Modelos de Regresión	Proporciona una relación cuantitativa entre variables independientes y la variable de respuesta.	Supone una relación lineal entre variables, lo que puede no ser válido en todos los casos.	Predicción de valores numéricos continuos.
Modelos de Recomendación	Personalización de sugerencias para los usuarios.	Puede requerir una gran cantidad de datos y tener problemas con datos faltantes o sesgos inherentes.	Recomendaciones de productos en comercio electrónico.
Modelos de Series Temporales	Captura patrones temporales y estacionales en los datos a lo largo del tiempo.	Sensibilidad a valores atípicos y datos faltantes, y dificultad para capturar tendencias no lineales.	Predicción de la demanda de productos. Pronóstico del clima.
Modelos de Atribución	Permite cuantificar la contribución relativa de diferentes variables a un resultado o impacto.	Puede ser difícil determinar la verdadera relación causal entre las variables.	Evaluación del retorno de inversión (ROI) de una campaña publicitaria.
Modelos de Árboles de Decisión	Proporciona una estructura de decisiones fácilmente interpretable.	Pueden ser propensos al sobreajuste si no se controla adecuadamente.	Clasificación y predicción en diversos campos, como medicina, marketing y finanzas.
Modelo Random Forest	Combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste.	Puede ser computacionalmente costoso y más difícil de interpretar que un solo árbol de decisión.	Clasificación y predicción en una amplia gama de aplicaciones, como análisis de datos médicos y detección de fraudes.

3.3.5. Series Temporales

Luego de realizar distintas investigaciones, hemos llegado a la certeza de que serán necesarios los modelos basados en series temporales. Se llego a esta decisión debido que se busca predecir el comportamiento futuro de los clientes, en base a su ruta de navegación pasada. A continuación se abarcaran las distintas propiedades de series temporales y sus graficos, además, se presentaran 3 modelos ampliamente aplicados en series temporales, los cuales son, el modelo ARIMA, SARIMA y redes LSTM:

Graficos de series temporales

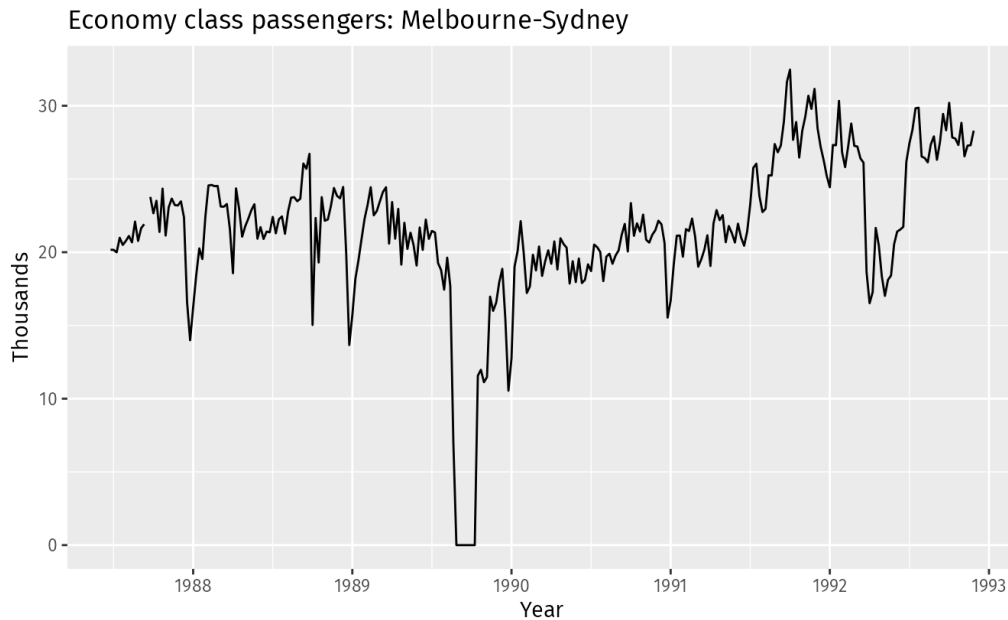
Uno de los primeros pasos para poder realizar cualquier tipo de análisis de datos es necesario graficar los datos, los gráficos permiten ver las distintas propiedades que los datos puedan tener, como lo son patrones, observaciones inusuales (outliers), cambios producidos por el tiempo y las relaciones que se encuentran entre las distintas variables que los datos puedan tener (Hyndman y Athanasopoulos, 2023). Por lo que, la selección de un buen grafico para representar los datos resulta ser de gran importancia para poder seleccionar el modelo predictivo más adecuado.

Dentro de esta sección revisaremos los gráficos más relevantes para nuestro estudio:

- **Trama de tiempo (Time plot):** Una de las maneras más practicas y usada para poder entender las series de tiempo es graficar los datos y una de las primeras opciones es una trama de tiempo. Esto quiere decir graficar la datos contra el tiempo de obtención de los datos.

La siguiente figura muestra la cantidad semanal de pasajeros que volaron en la clase económica de Ansett Airlines entre las dos ciudades más grandes de Australia (Hyndman y Athanasopoulos, 2023):

Figura 3.3: Cantidad semanal de pasajeros que volaron en la clase económica de Ansett Airlines



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

- **Patrones:** Una serie de tiempo puede presentar distintos tipos de patrones a lo largo del tiempo de la observación, y esta puede llegar a presentar este tipo de patrones (Hyndman y Athanasopoulos, 2023):
 - **Tendencia (Trend):** Una tendencia existe cuando se presenta un incremento o disminución en los datos dentro de un largo periodo de tiempo, esta tendencia no tiene por qué ser lineal, pudiendo tener tendencias positivas o negativas.
 - **Estacional (Seasonal):** Un patrón estacional existe cuando la serie de tiempo se ve afectada por factores como el tiempo del año o el día de semana. Gracias a esto se puede identificar que los patrones estacionales tienen una frecuencia conocida que por lo general no varía.

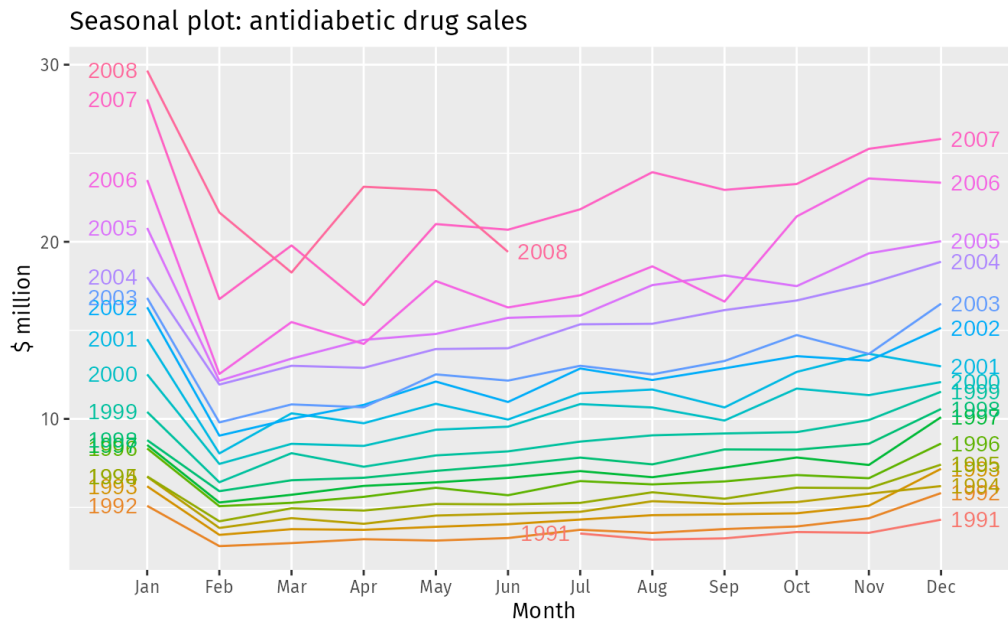
- **Cíclico (Cyclic):** Un patrón cíclico en una serie de tiempo existe cuando los datos presentan incrementos y disminuciones como lo haría un patrón estacional, pero sin una frecuencia conocida o fija, usualmente estos ciclos suelen tener una duración mayor a los de los patrones estacionales.

Muchas veces nos encontraremos con series de tiempo que contienen uno o más de estos patrones, por lo que para poder elegir un método predictivo resulta imperativo identificar los patrones presentes en los datos (Hyndman y Athanasopoulos, 2023).

- **Tramas de tiempo estacionales (Seasonal plots):** Esta gráfica es similar a la trama de tiempo antes mencionada, la diferencia reside en el hecho de que los datos u observaciones están graficadas en contra de cada “estación individual” de tiempo donde se realizó la observación (Hyndman y Athanasopoulos, 2023).

En la siguiente figura se representa el valor monetario de las ventas mensuales de medicamentos antidiabéticos en Australia:

Figura 3.4: Ventas mensuales de medicamentos antidiabéticos en Australia



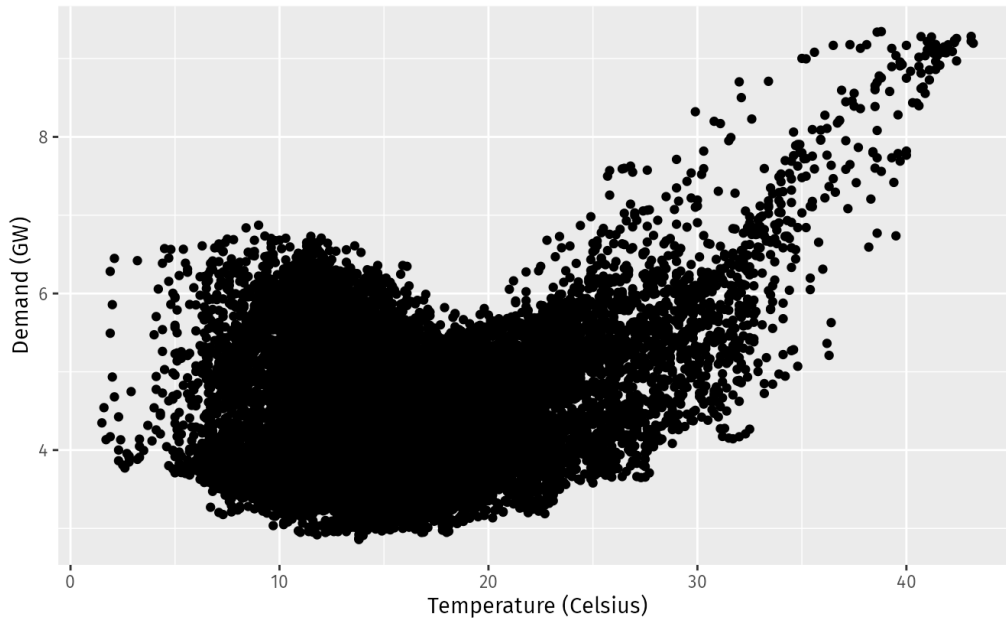
Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Este tipo de gráficos permite identificar de mejor manera los patrones que no se pudieron apreciar anteriormente en la trama de tiempo, ayuda especialmente para poder observar en que años el patrón de los datos cambia.

- **Gráficos de dispersión (Scatterplots):** Este tipo de grafico permite explorar las relaciones que existen entre series de tiempo, sus distintas variables y como esto afecta a la hora de predecir una serie de tiempo.

La siguiente figura muestra dos series de tiempo: demanda de electricidad (en GW y temperatura (Celsius)), para 2014 en Victoria, Australia. Las temperaturas son para Melbourne, la ciudad más grande de Victoria, mientras que los valores de demanda son para todo el estado (Hyndman y Athanasopoulos, 2023).

Figura 3.5: Demanda electrica debido a las temperaturas en 2 ciudades en Australia

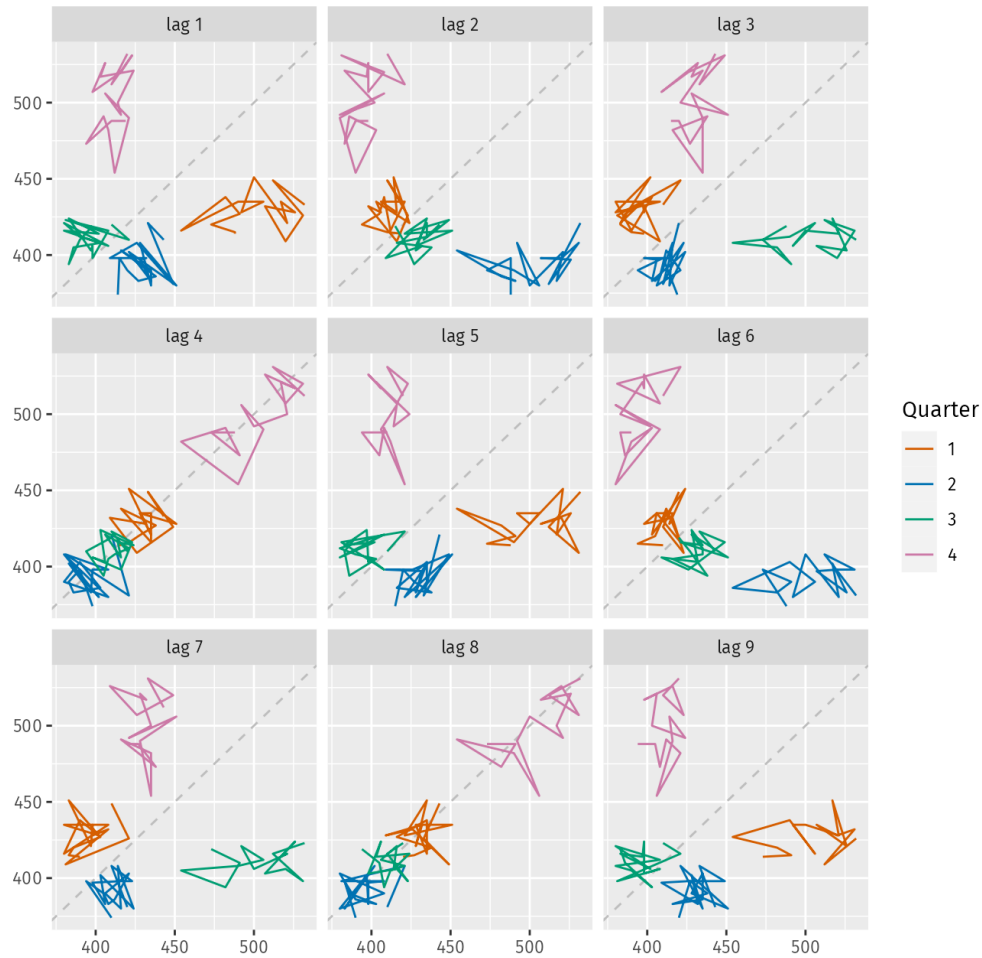


Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Este grafico de dispersión es de gran ayuda para visualizar la relación que tienen las variables, y poder entender los datos de mejor manera.

- **Gráficos de desfase (Lag plots):** Este tipo de gráfico representa la observación $Y(t)$ graficada contra la observación $Y(t-k)$ para cada valor de k . Donde en el eje horizontal se muestran valores desfasados de la serie de tiempo (Hyndman y Athanasopoulos, 2023). La siguiente figura muestra diagramas de dispersión de la producción trimestral de cerveza australiana:

Figura 3.6: Producción trimestral de cerveza australiana



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Los colores representan el cuatrimestre de la variable en el eje vertical. Se puede apreciar que en los desfases 4 y 8 se presenta una fuerte relación de las variables, lo que muestra un patrón estacional fuerte.

- **Correlación:** El estudio de la correlación explora la relación lineal entre dos variables, resulta de importancia calcularla, esto debido a que entrega que tan intrínsecamente relacionadas se encuentran las variables a analizar (Hyndman

y Athanasopoulos, 2023).

A continuación, se presenta la fórmula para conocer el coeficiente de correlación entre dos variables, x e y:

$$r = \frac{\sum (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum (x_t - \bar{x})^2} \sqrt{\sum (y_t - \bar{y})^2}}$$

El valor de r variara entre -1 y 1, dependiendo de que tan fuerte sea la correlación de las variables, mientras más cercano al -1, r representa una correlación negativa, por lo que, si r se encuentra más cercano a 1, esto quiere decir que las variables tienen una correlación positiva (Hyndman y Athanasopoulos, 2023).

- **Autocorrelación:** La autocorrelación mide la relación lineal entre valores rezagados de una serie de tiempo. Hay variados coeficientes de autocorrelación, dependiendo de cada panel del Lag plot, por ejemplo, r1 mide la relación entre las variables Y(t) y Y(t-1) y r2 mide la relación entre las variables Y(t) y Y(t-2) y así hasta considerar todos los datos (Hyndman y Athanasopoulos, 2023).

Para calcular el valor de r(k), donde T corresponde al largo de la serie de tiempo, se ocupa la siguiente fórmula:

$$r_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

De ejemplo, se calcularon los primeros 9 coeficientes de autocorrelación de la producción de cerveza en Australia estudiado en la sección pasada, obteniendo los siguientes valores:

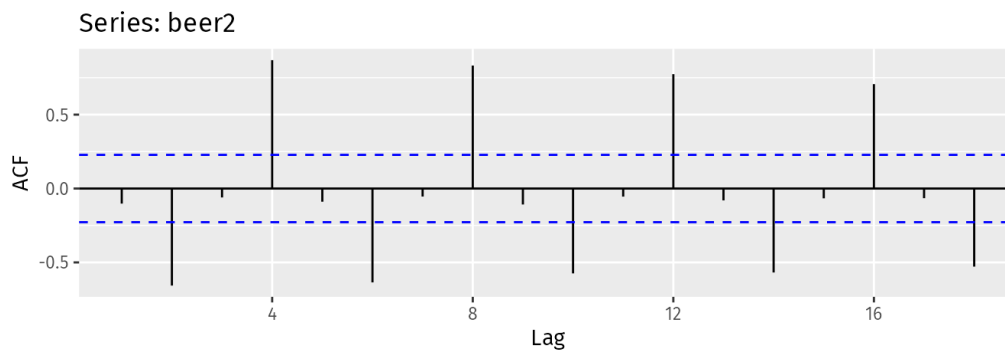
Figura 3.7: Coeficientes de autocorrelación de la producción de cerveza en Australia

r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9
-0.102	-0.657	-0.060	0.869	-0.089	-0.635	-0.054	0.832	-0.108

Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Estos coeficientes de autocorrelación son graficados para representar la **función de autocorrelación (ACF)**, que se muestra a continuación:

Figura 3.8: ACF de la producción trimestral de cerveza



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

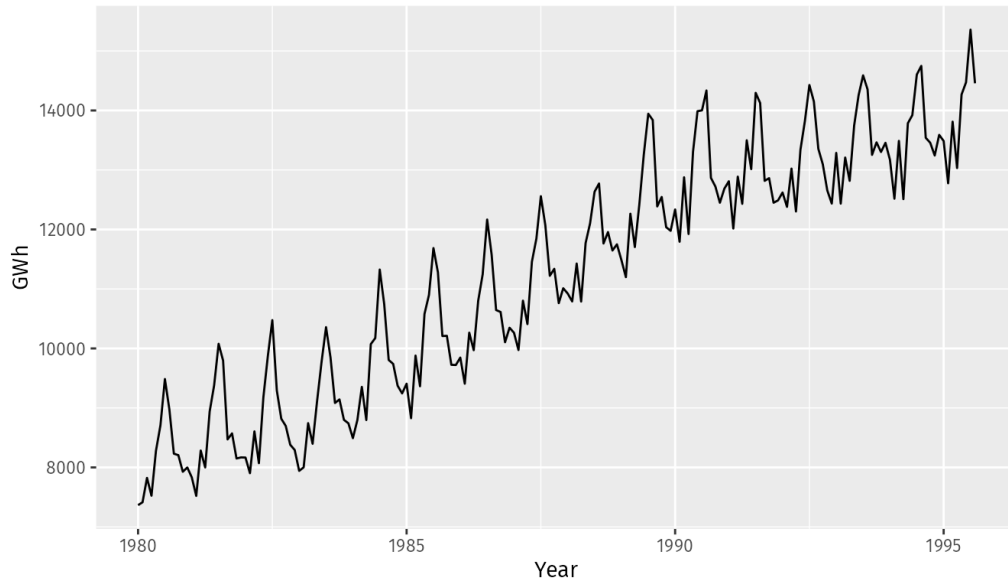
Las líneas azules indican que tan lejos de 0 pueden estar las correlaciones para que estas sean significativas.

Cuando los datos tienen alguna tendencia, el ACF tiene valores positivos que lentamente van disminuyendo, si los datos presentan un patrón estacional, el ACF mostrara valores más grandes para los desfases estacionales, por lo general siguiendo alguna frecuencia estacional.

Cuando los datos presentan una tendencia y además un patrón estacional, se pueden apreciar ambos efectos (Hyndman y Athanasopoulos, 2023). En las

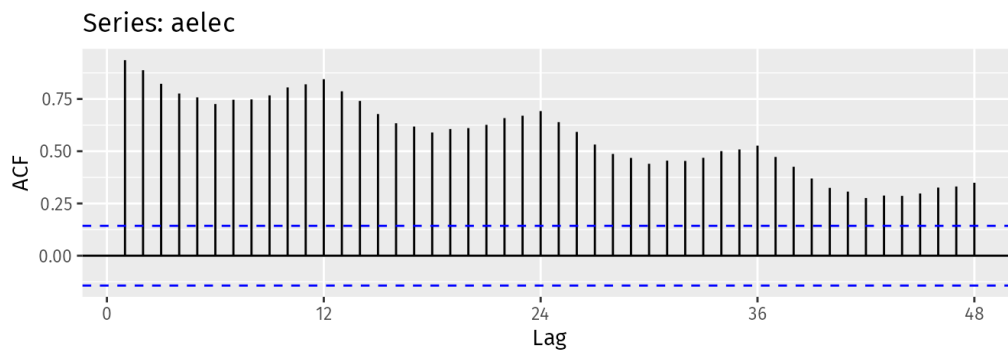
siguientes figuras se muestra la demanda mensual de electricidad de Australia, la cual presenta una tendencia y patrón estacional:

Figura 3.9: Demanda mensual de electricidad de Australia



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Figura 3.10: ACF demanda mensual de electricidad de Australia

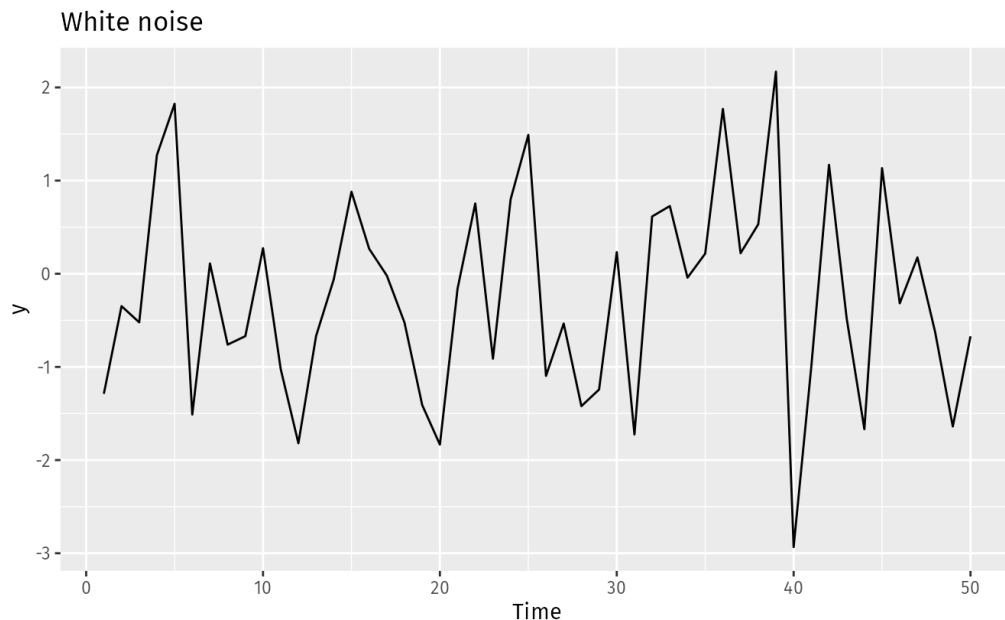


Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

En el ACF se puede apreciar una tendencia, esto debido a que los valores van disminuyendo lentamente, mientras que la forma de ondas es debido a la estacionalidad que se presenta cada año.

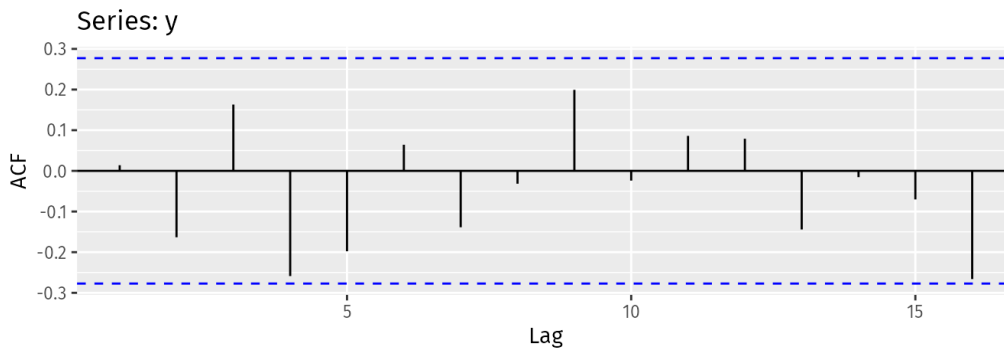
- **Ruido Blanco (White Noise):** Las series de tiempo que no tienen autocorrelación son llamadas ruido blanco, para esto se espera que cada autocorrelación sea lo más cercana a 0. Debido a la variación que estos valores pueden tener, por lo que para que una serie de tiempo sea considerada ruido blanco, se espera que el 95 % de sus valores representados en el ACF estén unos límites demarcados por $\pm 2/\sqrt{T}$, donde T corresponde al largo de la serie de tiempo, estos límites se encuentran representados comúnmente por líneas azules en el ACF. Por consiguiente, si un valor o más se encuentra fuera de los límites o el más del 5 % de los valores estén fuera de los límites, la serie de tiempo en cuestión probablemente no sea ruido blanco.

Figura 3.11: Ruido blanco



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Figura 3.12: ACF Ruido blanco



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Teniendo en cuenta que $T = 50$, por lo tanto, los límites están calculados como $\pm 2/\sqrt{50} = \pm 0,28$. En el ACF se puede apreciar que todos los coeficientes de autocorrelación se encuentran dentro estos límites, concluyendo que los datos son ruido blanco.

Modelo ARIMA

El modelo ARIMA o modelo autorregresivo de media móvil integrado, por sobre otros modelos de series temporales, se centra en describir las autocorrelaciones que existen entre los datos (Hyndman y Athanasopoulos, 2023).

- **Estacionariedad:** Una serie de tiempo estacionaria, se presenta cuando sus propiedades no dependen del momento en el que fue registrada la observación. Por lo que, las series de tiempo que presentan tendencias o patrones estacionales son series de tiempo no estacionarias, ya que las tendencias y las distintas estaciones de tiempo pueden afectar la serie de tiempo en varias ocasiones (Hyndman y Athanasopoulos, 2023).

Las series de tiempo estacionarias por lo general son series de ruido blanco, ya que no muestran autocorrelación entre los datos y tampoco presentan patrones predecibles a lo largo del tiempo.

- **Diferenciación:** Una manera de poder cambiar una serie de tiempo no estacionaria en una estacionaria, es aplicar la diferenciación, esto se refiere a calcular la diferencia entre observaciones consecutivas (Hyndman y Athanasopoulos, 2023).

Una de las transformaciones más ocupada para estabilizar la varianza de los datos son los logaritmos. Por otro lado, la diferenciación estabiliza el promedio de la serie de tiempo debido a que es capaz de reducir o remover los distintos cambios que se puedan presentar en la serie de tiempo, tales como las tendencias y patrones estacionales.

- **Modelo Random walk:** La serie diferenciada es el cambio que se presenta entre las observaciones consecutivas de la serie original, esta se denota por:

$$y'_t = y_t - y_{t-1}$$

Ya que es imposible conseguir la diferenciada de la primera observación y'_1 , la serie diferenciada solo tendrá valores T-1. Si la serie diferenciada es un ruido blanco, la fórmula se puede escribir de la siguiente manera, donde ε_t representa el ruido blanco (Hyndman y Athanasopoulos, 2023):

$$y_t - y_{t-1} = \varepsilon_t$$

El **Modelo Random walk** se obtiene luego de despejar y_t :

$$y_t = y_{t-1} + \varepsilon_t$$

Debido a que los movimientos futuros en los datos son impredecibles, las predicciones del modelo son iguales a la última observación, por lo que este tipo de modelos es ampliamente ocupado por datos no estacionarios. Pudiendo concluir que el modelo random walk sustenta los pronósticos naïve (Hyndman y Athanasopoulos, 2023).

$$y_t - y_{t-1} = C + \varepsilon_t \leftrightarrow y_t = C + y_{t-1} + \varepsilon_t$$

El valor de C representa el promedio de los cambios entre observaciones consecutivas (Hyndman y Athanasopoulos, 2023), dependiendo del signo de C la serie se verá derivada positivamente si es positivo o negativamente en el caso contrario.

- **Diferenciación estacional:** Esta diferenciación se realiza para calcular el cambio entre una observación y la observación previa de la misma estación de tiempo (Hyndman y Athanasopoulos, 2023), obteniendo la siguiente fórmula:

$$y'_t = y_t - y_{t-m}$$

Donde m = cantidad de estaciones, también son llamadas "diferencias m -desfazadas", ya que realizamos la resta con una observación luego de m periodos (Hyndman y Athanasopoulos, 2023).

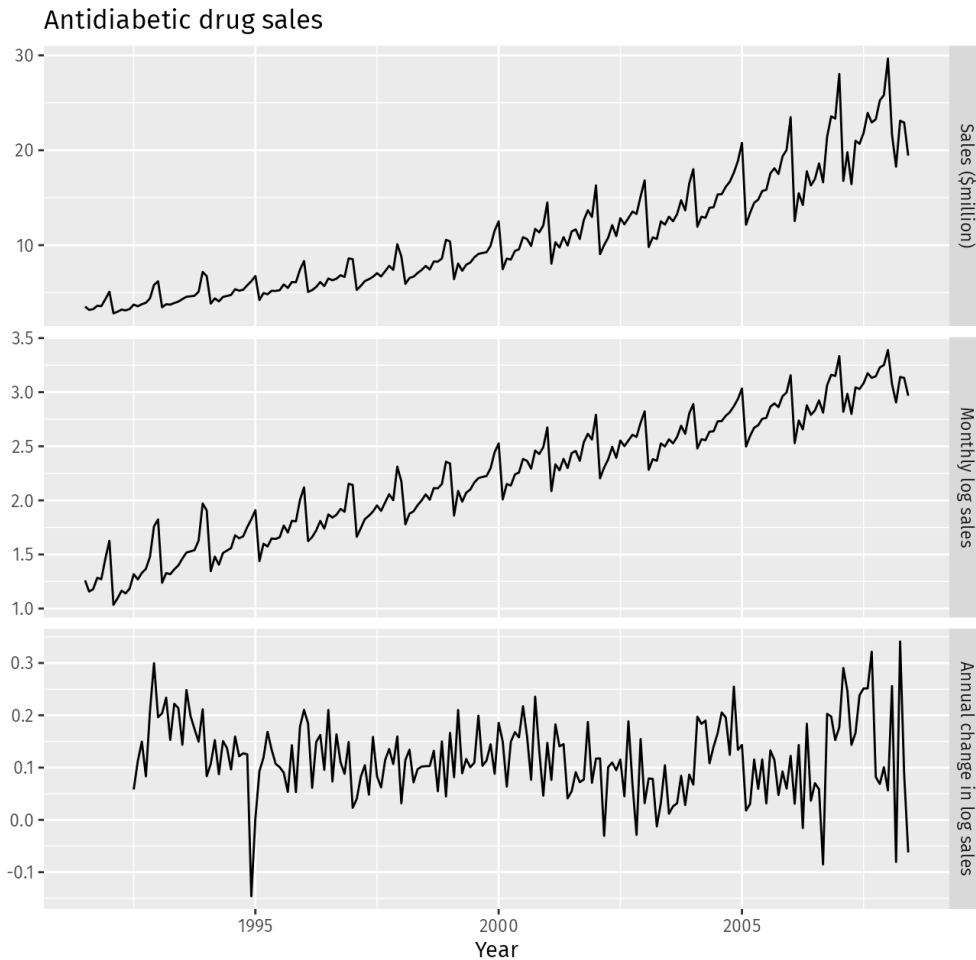
Si luego de realizar una diferenciación estacional a una serie de tiempo, esta parece haberse transformado a ruido blanco, se modelaran los datos originales de la siguiente manera:

$$y_t = y'_t + m + \varepsilon_t$$

Como este tipo de modelo entrega predicciones iguales a la última observación de la estación seleccionada, en otras palabras, este modelo entrega pronósticos naïve (Hyndman y Athanasopoulos, 2023).

En la siguiente figura se muestra como las diferenciaciones estacionales y la aplicación de logaritmos a una serie de tiempo, en este caso la venta de medicamentos antidiabéticos, lograr cambiar la forma de la serie a una de ruido (Hyndman y Athanasopoulos, 2023):

Figura 3.13: Aplicación de logartimos y diferenciación estacional a ventas de medicamentos antidiabéticos



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

Por lo general, a la diferenciación ordinaria se le llama "primera diferenciación", refiriéndose a la diferenciación en el desfase 1.

Para poder obtener una serie de tiempo que parezca ruido blanco, muchas veces se tendrá que aplicar ambas diferenciaciones, la primera diferenciación y diferenciación estacional, el orden de aplicación no afecta el resultado (Hyndman y Athanasopoulos, 2023).

Tambien existe la segunda diferenciada estacional, por lo que el modelo de la serie doblemente diferenciado se escribe de la siguiente manera:

$$y_t'' = y_t' - y_{t-1}' \leftrightarrow (y_t - y_{t-m}) - (y_{t-1} - y_{t-m-1}) \leftrightarrow y_t - y_{t-1} - y_{t-m} + y_{t-m-1}$$

Si la serie de tiempo, presenta un patrón estacional, es recomendable realizar la diferenciación estacional como primer paso para obtener una serie de tiempo estacionaria, pero si se ocupa la primera diferenciación puede que todavia hayan patrones estacionales en la serie de tiempo, teniendo que aplicar más diferenciaciones para obtener la estacionariedad (Hyndman y Athanasopoulos, 2023).

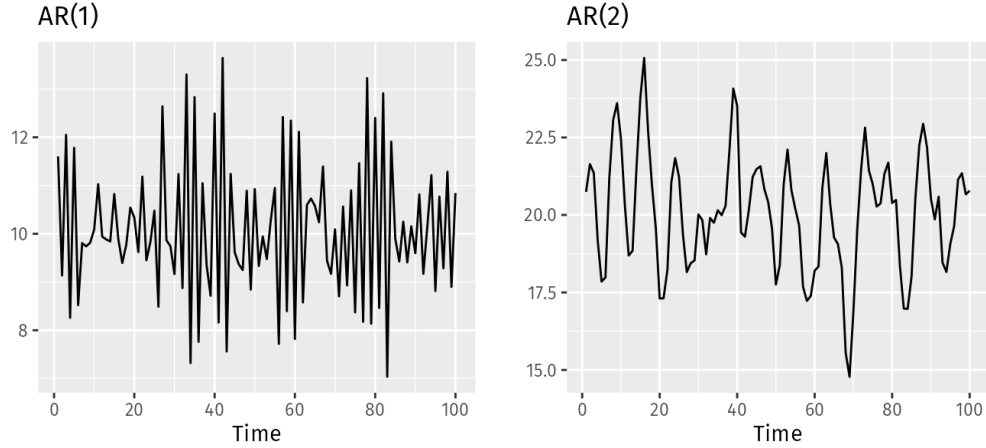
- **Modelo de autorregresión:** Como se menciona anteriormente, los modelos autorregresivos predicen la variable de interes utilizando una combinación lineal de valores pasados de la variable. El término autorregresión indica que se trata de una regresión de la variable contra sí misma.

Por lo que un modelo autorregresivo de orden p , donde ε_t corresponde a ruido blanco, se escribe de la siguiente manera:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

Este tipo de modelo es llamado **AR**(p), gracias a la flexibilidad de los modelos autorregresivos, estos pueden ser aplicados en series de tiempo que presenten distintos patrones (Hyndman y Athanasopoulos, 2023). A continuación, se presenta una figura que muestran dos modelos AR, AR(1) y AR(2).

Figura 3.14: Modelos autorregresivos con diferentes parámetros



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

El modelo AR(1) tiene como fórmula: $y_t = 18 - 0,8y_{t-1} + \varepsilon_t$ y el modelo AR(2) tiene como fórmula: $y_t = 8 + 1,3y_{t-1} - 0,7y_{t-2} + \varepsilon_t$, en ambos casos ε_t (ruido blanco) tiene una distribución normal, con promedio igual a 0 y varianza igual a 1 (Hyndman y Athanasopoulos, 2023).

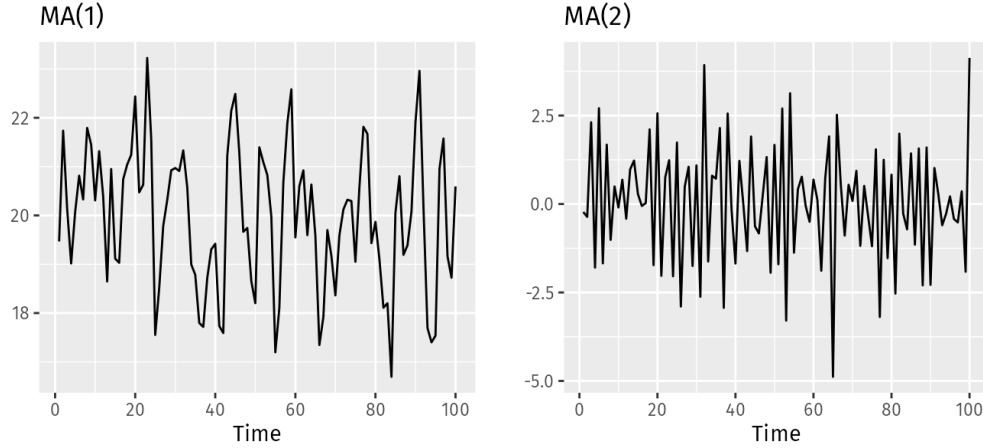
- **Modelo de media móvil:** Este tipo de modelo ocupa el error de los pronósticos como si fuera un modelo de regresión, con la diferencia en una regresión se ocupan los valores pasados.

Teniendo ε_t como ruido blanco, el modelo de media móvil de orden q o **MA**(q) se escribe de la siguiente manera (Hyndman y Athanasopoulos, 2023):

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \dots + \theta_q\varepsilon_{t-q} + \varepsilon_t$$

A continuación, se presenta una figura que muestran dos modelos MA, MA(1) y MA(2).

Figura 3.15: Modelos de media móvil con diferentes parámetros



Fuente: Forecasting: Principles and Practice (Hyndman y Athanasopoulos, 2023). Recuperado de <https://otexts.com/fpp2/time-plots.html>

El modelo MA(1) tiene como fórmula: $y_t = 20 + \varepsilon_t + 0,8\varepsilon_{t-1}$ y el modelo MA(2) tiene como fórmula: $y_t = \varepsilon_t - \varepsilon_{t-1} + 0,8\varepsilon_{t-2}$, en ambos casos ε_t (ruido blanco) tiene una distribución normal, con promedio igual a 0 y varianza igual a 1 (Hyndman y Athanasopoulos, 2023).

Un modelo AR(p) estacionario es capaz de ser escrito como un modelo MA(∞), esto se puede demostrar para un modelo AR(1) luego de realizar repetidas sustituciones:

$$\begin{aligned}
 y_t &= \phi_1 y_{t-1} + \varepsilon_t \\
 &= \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\
 &= \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\
 &= \phi_1^3 y_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t
 \end{aligned}$$

Teniendo $-1 < \phi_1 < 1$, el valor de ϕ_1^k irá disminuyendo a medida que k siga creciendo, obteniendo un proceso MA(∞) (Hyndman y Athanasopoulos, 2023),

este tipo de modelo se escribe de la siguiente manera:

$$y_t = \varepsilon_t + \phi_1 \varepsilon_t - 1 + \phi_1^2 \varepsilon_t - 2 + \phi_1^3 \varepsilon_t - 3 + \dots$$

También se puede realizar el proceso inverso, para obtener un modelo $AR(\infty)$ desde un modelo MA, si es que se definen ciertos límites. Este tipo de modelos que se pueden transcribir de MA a $AR(\infty)$ y AR a $MA(\infty)$, son llamados modelos **invertibles** (Hyndman y Athanasopoulos, 2023).

Combinando los modelos de diferenciación, autorregresivo y de media móvil obtenemos el modelo **ARIMA** (Hyndman y Athanasopoulos, 2023), este queda modelado de la siguiente manera:

$$y'_t = c + \phi_1 y'_t - 1 + \dots + \phi_p y'_t - p + \theta_1 \varepsilon_t - 1 + \dots + \theta_q \varepsilon_t - q + \varepsilon_t$$

Teniendo en cuenta que y'_t corresponde a la serie diferenciada, a la derecha de la igualdad se encuentran las variables que permiten realizar las predicciones, entre estas se encuentran valores desfasados de y_t y errores desfasados. El modelo **ARIMA**(p, d, q) tiene 3 variables, donde cada una de estas variables está relacionada a (Hyndman y Athanasopoulos, 2023):

- p = orden de la parte autorregresiva del modelo.
- d = grado de la primera diferenciación.
- q = orden de la parte de media móvil.

De la misma manera que la estacionariedad e invertibilidad se aplican a los modelos de autorregresión y de media móvil, también se aplican para los modelos ARIMA. Este modelo presenta casos especiales, los cuales representan modelos anteriormente mencionados, los cuales son:

- White noise: ARIMA(0,0,0)
- Random walk: ARIMA(0,1,0) sin constante
- Random walk con desfase: ARIMA(0,1,0) con constante
- Autorregresión: ARIMA(p ,0,0)
- Media móvil: ARIMA(0,0, q)

Modelo SARIMA

El modelo SARIMA es una extensión del ARIMA, diseñado específicamente para series temporales que presentan estacionalidad. Mientras que ARIMA es eficaz para modelar series no estacionarias sin componentes estacionales, SARIMA incorpora parámetros adicionales para capturar y predecir patrones que se repiten a intervalos regulares en el tiempo.

La estacionalidad en una serie temporal refleja un patrón regular que se repite cada S períodos. Aquí, S indica el número de períodos requeridos para que el patrón complete un ciclo y comience a repetirse (Morales Oñate, 2022).

- **Diferenciación no estacional:** Si en los datos se detecta una tendencia, es probable que sea necesario aplicar una diferenciación no estacional. A menudo, al realizar una primera diferenciación no estacional, se logra eliminar la tendencia de la serie temporal. Para este propósito, utilizamos la fórmula:

$$(1 - B)x_t = x_t - x_{t-1}$$

- **Diferenciación para tendencia:** Cuando tanto la tendencia como la estacionalidad están presentes en una serie temporal, es posible que necesitemos aplicar tanto una diferenciación no estacional como una estacional.

Por lo tanto, es esencial examinar tanto el ACF (Función de Autocorrelación)

como el PACF (Función de Autocorrelación Parcial) de la expresión:

$$(1 - B^{12})(1 - B)x_t = (x_t - x_{t-1}) - (x_{t-12} - x_{t-13})$$

Es importante destacar que eliminar la tendencia no necesariamente implica que se haya eliminado toda la dependencia en la serie. Mientras que la tendencia (o el componente de la media) puede haber sido eliminada, aún puede persistir un comportamiento periódico. Al aplicar la diferenciación, descomponemos la dependencia en eventos recientes y eventos a largo plazo (Morales Oñate, 2022).

- **Diferenciación para estacionalidad:** Cuando se tienen datos con componentes estacionales, es posible que los componentes no estacionales a corto plazo sigan influyendo en el modelo. Por ejemplo, en el caso de las ventas de componentes electrónicos, las ventas de los últimos uno o dos meses, junto con las ventas del mismo mes del año anterior, pueden ser relevantes para predecir las ventas del mes actual.

Por lo tanto, es crucial examinar el comportamiento del ACF (Función de Autocorrelación) y PACF (Función de Autocorrelación Parcial) en los primeros rezagos para determinar qué términos no estacionales pueden ser significativos en el modelo.

Siendo expresado de manera abreviada:

$$(p, d, q) \times (P, D, Q, S)$$

- P : Es el número de términos autorregresivos estacionales.
- D : Es el grado de diferenciación estacional.
- Q : Es el número de términos de promedio móvil estacional.

- S : Es el número de periodos en una temporada (por ejemplo, 12 para datos mensuales si se observa una periodicidad anual).

3.3.6. Redes LSTM Long Short-Term Memory

¿Qué es una red neuronal?

Una red neuronal es un modelo computacional inspirado en la estructura y funcionamiento del cerebro humano, utilizado en el campo de la inteligencia artificial. Forma parte del aprendizaje profundo dentro del machine learning y está compuesta por unidades, llamadas neuronas, organizadas en capas. Estas neuronas están interconectadas y procesan información adaptándose y aprendiendo de los datos. Las redes neuronales se emplean para abordar tareas complejas, como el reconocimiento de imágenes o la traducción automática, mejorando su rendimiento a medida que se les expone a más datos (Amazon, 2023).

¿Qué son las redes Long Short-Term Memory?

La LSTM es un tipo de red neuronal recurrente (RNN) diseñada para resolver el problema del gradiente evanescente mediante la introducción de una célula de memoria que puede almacenar información durante periodos de tiempo más largos (Filzinger, 2023).

Arquitectura de una Long Short-Term Memory

- **Puerta de entrada:** La puerta de entrada en una LSTM utiliza una función de activación sigmoidea para decidir qué valores pasar a la célula de memoria. La función sigmoidea genera valores entre 0 y 1, lo que permite que la puerta de entrada *decida* en qué grado un valor debe ser almacenado en la célula de memoria. Un valor cercano a 1 indica que casi toda la información debe ser conservada, mientras que un valor cercano a 0 sugiere que la información debe ser mayormente descartada (Filzinger, 2023).
- **Puerta del Olvido:** Controla qué información de la célula de memoria del paso temporal anterior debe ser conservada y cuál debe ser descartada. Utiliza una función de activación sigmoidea para determinar esto. Un valor cercano a

1 de la sigmoidea indica que la información debe ser conservada, mientras que un valor cercano a 0 indica que debe ser olvidada. (Filzinger, 2023).

- **Puerta de Salida:** Controla qué información de la célula de memoria debe ser enviada como salida. Utiliza una función de activación sigmoidea para decidir qué partes de la memoria se deben considerar y luego aplica una función de tangente hiperbólica para escalar los valores entre -1 y 1, determinando así la salida final de la célula LSTM (Filzinger, 2023).
- **Célula de memoria:** Es el componente central de la arquitectura LSTM. Esta célula almacena información a través del tiempo, pudiendo olvidar selectivamente información no relevante y añadir nueva información a su estado interno.

En cada paso temporal, el modelo LSTM recibe un vector de entrada junto con un vector de estado oculto proveniente del paso temporal anterior. A partir de estos, la puerta de entrada decide qué nueva información se debe almacenar en la célula de memoria, mientras que la puerta de olvido decide qué información previamente almacenada debe descartarse.

Después, se genera un estado candidato utilizando la función de activación tangente hiperbólica. Este estado candidato se combina con el estado actual de la célula de memoria mediante una operación de suma elemento a elemento, actualizando así la información almacenada.

Finalmente, la puerta de salida determina cuál de esta información actualizada se emitirá como salida de la LSTM. Este output, junto con el estado actualizado de la célula de memoria, se pasa al siguiente paso temporal (Filzinger, 2023).

3.3.7. Metodología del proyecto

Para llevar a cabo el desarrollo del proyecto, se definieron cuatro fases que corresponden a la totalidad del proyecto, las cuales corresponden a:

Fase 1: Planteamiento y planificación

Para la primera fase del proyecto, se llevará a cabo una planificación de la manera en la que será abordada la problemática, para desarrollar un anteproyecto que será utilizado para evaluar y planificar las actividades correspondientes al desarrollo del proyecto. Entre ellas se encuentran:

- Planteamiento del proyecto y sus objetivos.
- Definición de alcances y limitaciones.
- Creación de un cronograma de actividades.

Fase 2: Investigación

Para la segunda fase, se realizará una investigación de herramientas y recursos necesarios para llevar a cabo un diseño de la solución para la problemática del proyecto planteado, sumado a un análisis de las bases de datos brindadas por la empresa AFP Capital. Una vez realizado lo anterior, se llevará a cabo una propuesta de diseño para la problemática, siendo entregada y analizada por la empresa, con la finalidad de pasar a desarrollo. Algunas de las actividades de esta fase corresponden a:

- Investigación del problema.
- Toma de requerimientos.
- Investigación de tecnologías de análisis de datos.

Fase 3: Modelamiento y desarrollo

Para la tercera fase, se llevará a cabo el diseño y desarrollo del sistema propuesto, además de realizar pruebas para verificar el correcto funcionamiento. Algunas de las actividades de esta fase corresponden a:

- Modelado del sistema ETL.

- Modelado de la API.
- Implementación del modelo propuesto.
- Pruebas y validaciones.
- Correcciones de errores.

Fase 4: Conclusiones y recomendaciones

Para la última fase, se dará fin al desarrollo del proyecto, elaborando un manual de usuario el cual indicaría algunas funcionalidades del sistema. Algunas de las actividades de esta fase corresponden a:

- Desarrollo de manual de usuario.
- Redacción de conclusiones y recomendaciones.
- Cierre del proyecto.

3.3.8. Metodología del sistema

CRISP-DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un proceso estándar utilizado para realizar proyectos de minería de datos. La metodología CRISP-DM se divide en seis fases distintas que se describen a continuación (Eric, 2023):

1. **Comprensión del problema:** En esta fase se define el problema a resolver y se establecen los objetivos del proyecto. También se recopilan los datos necesarios para el proyecto.
2. **Comprensión de los datos:** En esta fase se realiza una exploración de los datos para comprender su calidad, estructura y relevancia para el problema en

cuestión.

3. **Preparación de los datos:** En esta fase se limpian y procesan los datos para que puedan ser utilizados en la etapa de modelado.
4. **Modelado:** En esta fase se aplican técnicas de modelado para desarrollar un modelo predictivo. Se prueban diferentes modelos y se selecciona el que mejor se ajuste a los datos.
5. **Evaluación:** En esta fase se evalúa el modelo desarrollado en la fase anterior. Se verifica que el modelo funcione correctamente y se ajuste adecuadamente a los datos.
6. **Implementación:** En esta fase se implementa el modelo desarrollado en la fase de modelado en un entorno de producción. También se establecen planes para monitorear el rendimiento del modelo y actualizarlo según sea necesario.

Las fases de la metodología CRISP-DM son iterativas, lo que significa que es posible volver a una fase anterior si es necesario.

OSEMN

La metodología OSEMN (acrónimo de las palabras en inglés: Obtain, Scrub, Explore, Model, Interpret) es un proceso utilizado en la minería de datos y el análisis de datos para trabajar con grandes conjuntos de datos de manera efectiva (Eric, 2023).

1. **Obtener (Obtain):** En esta etapa, se recopilan los datos necesarios para el análisis. Los datos pueden provenir de diferentes fuentes, como bases de datos, archivos en línea o registros de sensores. La calidad y la cantidad de los datos obtenidos son cruciales para el éxito del análisis.
2. **Limpieza (Scrub):** Una vez que se han obtenido los datos, es necesario realizar una limpieza para eliminar datos innecesarios o incorrectos. Esta etapa

puede implicar la eliminación de duplicados, la corrección de errores y la eliminación de valores atípicos. El objetivo de esta etapa es obtener datos limpios y coherentes para el análisis.

3. **Exploración (Explore):** En esta etapa, se utilizan técnicas de visualización y estadísticas para explorar los datos y obtener información sobre ellos. Se pueden identificar patrones, tendencias y relaciones entre diferentes variables. El objetivo es obtener una comprensión más profunda de los datos y de cómo se relacionan entre sí.
4. **Modelado (Model):** En esta etapa, se utilizan técnicas de modelado estadístico o de aprendizaje automático para crear modelos que puedan predecir resultados futuros o identificar patrones en los datos. El objetivo es utilizar los datos para crear un modelo que pueda utilizarse para tomar decisiones informadas.
5. **Interpretación (Interpret):** En esta etapa, se interpretan los resultados obtenidos en la etapa de modelado. Los resultados pueden ser utilizados para tomar decisiones o para generar nuevas hipótesis que puedan ser exploradas en futuros análisis.

Se propone el uso de la metodología OSEMN, ya que se enfoca en el análisis de datos y la creación de modelos predictivos. OSEMN también es una metodología más flexible que CRISP-DM, lo que puede ser útil en un proyecto de SCRUM donde se busca una mayor adaptabilidad.

Por otro lado, también se propone el uso de la metodología CRISP-DM, ya que el proyecto incluye una etapa de exploración y análisis de datos, seguida por una fase de construcción de modelos. CRISP-DM se enfoca en el proceso completo de minería de datos, desde la comprensión del problema hasta la implementación del modelo, lo que puede servir para realizar un trabajo más estructurado.

Ya que este proyecto se encuentra bajo el marco de trabajo SCRUM, ambas metodologías pueden ser utilizadas de manera complementaria, utilizando OSEMN para las fases de creación de modelos y CRISP-DM para la etapa de exploración y

análisis de datos.

CAPÍTULO 4: PROCESO ETL

El proceso ETL (Extract, Transform, Load) representa el núcleo de la gestión de datos en numerosas organizaciones, desempeñando un papel esencial en la recopilación, transformación y carga de información crítica. En este capítulo del informe, exploraremos en detalle el funcionamiento de un proceso ETL, así como su diseño estratégico. Analizaremos cómo esta metodología se convierte en una piedra angular para la toma de decisiones basadas en datos y la mejora de la eficiencia operativa.

4.1. Diseño Proceso ETL

El diseño de un proceso ETL (Extracción, Transformación y Carga) implica seguir distintos pasos para asegurar que este proceso y el flujo de datos sean eficientes, precisos y cumplan con los requisitos del proyecto (Kimball y Caserta, 2004). Los pasos que se acordaron seguir son los siguientes:

- Requisitos ETL
- Identificación fuente de datos
- Diseño modelo de datos objetivo
- Planificación de las transformaciones
- Selección herramientas
- Construcción y prueba proceso ETL
- Monitoreo proceso ETL

4.1.1. Requisitos ETL

En esta etapa se definen los requisitos del proyecto, las fuentes de datos, los objetivos comerciales y del proceso ETL, las necesidades de análisis y los plazos para realizar el proceso. Estableciendo una base sólida para el diseño y buen funcionamiento del proceso ETL.

- **Fuente de datos:** La fuente de datos corresponde a varios archivos .CSV o de texto plano que contienen información de la navegación web de los clientes en forma de web logs. Estos archivos los proporciona la empresa luego de realizar un proceso de recopilación, ordenamiento y anonimización de los datos.
- **Objetivos comerciales:** Analizar el comportamiento de los clientes y sus preferencias de uso en un período igual o inferior a 1 mes, para poder predecir navegaciones futuras, y a partir de esto proporcionar atenciones personalizadas.
- **Objetivos proceso ETL:** Realizar las transformaciones necesarias para asegurar que el flujo de datos sea eficiente y preciso, a través de la limpieza de los datos, la normalización, la agregación, el filtrado, el enriquecimiento de datos, así como los cálculos y derivaciones necesarios.
- **Necesidades de análisis:** Realizar un análisis exploratorio de los datos entregados.

4.1.2. Identificación fuente de datos

En esta etapa se determinan las fuentes de datos a ser usadas para el proyecto, incluyendo bases de datos, archivos .CSV y APIs. Esto además comprende la definición de la estructura, el formato y ubicación de cada fuente de datos dentro del proyecto.

La fuente de datos corresponde a varios archivos .CSV que contiene información de la navegación web de los clientes en forma de web logs. Los web logs registrados vienen con 4 campos, especificados a continuación:

- **rut cliente:** Este atributo representa un identificador único por cliente.
- **fecha evento:** Representa la fecha y hora de la interacción del cliente con el sitio web.
- **metodo:** Este atributo representa cuál fue el método al cual el cliente llamó al interactuar con el sitio web.
- **canal:** Corresponde al canal web con el cual el cliente realizó la interacción en el sitio web.

Para almacenar la fuente de datos se utiliza una estructura de carpetas, siendo las siguientes:

- **Input:** Dentro de esta carpeta se encontrarán los archivos.
- **Intermediate:** Aquí se almacenará el archivo con la información preprocesada.
- **Output:** Dentro de esta última carpeta se almacenará la información ya procesada y lista para ser usada.

4.1.3. Diseño del modelo de datos objetivo

En esta etapa, se lleva a cabo el diseño del modelo de datos objetivo. Dado el tipo de proyecto, no será necesario crear un modelo multidimensional u algo similar para su implementación. Esto se debe a que los datos de entrada para los algoritmos predictivos consistirán en datos planos, en forma de Dataframes de pandas.

4.1.4. Planificación de las transformaciones

Dentro de esta etapa se realiza la planificación detallada de las transformaciones necesarias para construir una base sólida y consistente para el desarrollo del proyecto. Estas transformaciones implican una serie de pasos que permiten limpiar, filtrar, combinar y enriquecer los datos de manera adecuada.

La planificación de las transformaciones es fundamental para garantizar la calidad y la integridad de los datos que serán utilizados en el proyecto. Durante esta etapa, se identifican las tareas específicas que deben llevarse a cabo para lograr los objetivos establecidos, teniendo en cuenta los requisitos del proyecto y las necesidades del negocio.

Algunas de las transformaciones comunes incluyen (Kimball y Caserta, 2004):

- **Limpieza de datos:** Se realizan tareas de limpieza para corregir errores, eliminar valores duplicados o inconsistentes, y garantizar la coherencia de los datos. Esto puede incluir la corrección de formatos incorrectos, la normalización de datos, el manejo de valores faltantes o la estandarización de la información. La limpieza se realizó para solo tomar en cuenta los valores no nulos y se estableció el formato 'YYYY;MM;DD;HH;MM;SS.ss' de la columna 'fecha evento'.
- **Filtrado de datos:** Se aplican filtros para seleccionar y extraer los datos relevantes para el proyecto, descartando aquellos que no cumplen con ciertos criterios o condiciones específicas, **no permitiendo registros sin métodos asociados**. Esto ayuda a reducir el volumen de datos y a enfocarse en la información más relevante y útil.
- **Combinación de datos:** Se integran datos provenientes de diferentes fuentes o fuentes de datos diversas. Esto implica fusionar conjuntos de datos relacionados, realizar uniones o cruces de tablas, y establecer relaciones entre los datos para generar una visión global y coherente.
- **Enriquecimiento de datos:** Se agregan atributos o información adicional a los datos existentes para enriquecer su contexto y mejorar su valor. Esto puede implicar la incorporación de datos externos, la realización de cálculos derivados, la normalización de datos o la aplicación de reglas específicas.

Es importante tener en cuenta que la planificación de las transformaciones considera el orden y la secuencia adecuada de ejecución, así como la documentación de cada paso y los criterios de validación y verificación para garantizar la calidad de los

datos transformados.

4.1.5. Selección de herramientas

En esta etapa, se realiza la selección de herramientas de software que se ajusten a las necesidades y requisitos del proyecto para llevar a cabo el proceso ETL de manera eficiente. Se evalúan diferentes opciones disponibles en función de su capacidad, compatibilidad y facilidad de uso, para garantizar una elección adecuada, es por esto que se seleccionaron las siguientes herramientas:

- **Visual Studio Code:** Es un editor de código fuente desarrollado por Microsoft. Es conocido por su enfoque en la simplicidad, la personalización y la eficiencia.
- **Python:** Es un lenguaje de programación interpretado, de alto nivel y de propósito general, conocido por su sintaxis clara y legible. Es utilizado en una amplia gama de aplicaciones, desde desarrollo web hasta ciencia de datos y aprendizaje automático. Python se destaca por su facilidad de aprendizaje y su amplia biblioteca estándar, que ofrece numerosas funcionalidades predefinidas para diversas tareas (van Rossum y Drake, 2003).
 - **Numpy:** Es una biblioteca fundamental para la computación científica en Python. Proporciona estructuras de datos eficientes y funciones para realizar operaciones numéricas y de manipulación de arrays (van der Walt, Colbert, y Varoquaux, 2011).
 - **Pandas:** Es una biblioteca poderosa para el análisis de datos basada en Numpy, que proporciona estructuras de datos flexibles y eficientes, como DataFrames, y un conjunto completo de funciones para la manipulación y transformación de datos (Mckinney, 2011).
 - **Dask:** Es una biblioteca de paralelización flexible que permite escalar el procesamiento de datos en Python. Proporciona estructuras de datos paralelas y operaciones distribuidas que facilitan el procesamiento de grandes

volúmenes de datos (*dask: Scale the Python tools you love*, 2023).

4.1.6. Construcción y prueba proceso ETL

En esta etapa, se lleva a cabo la implementación del diseño del proceso ETL previamente definido, utilizando las herramientas seleccionadas. Se desarrollan los flujos de extracción, transformación y carga de los datos según lo establecido en el diseño.

Una vez implementado, se procede a realizar pruebas exhaustivas para garantizar el correcto funcionamiento del proceso. Estas pruebas incluyen la verificación de la extracción de datos de las fuentes, la correcta aplicación de las transformaciones definidas y la carga exitosa de los datos en el destino final.

El objetivo de las pruebas es asegurar que el proceso ETL cumpla con los requisitos establecidos y que los resultados obtenidos sean los esperados. Esto implica validar la integridad y coherencia de los datos transformados, así como verificar el rendimiento y la escalabilidad del proceso.

En caso de encontrar inconvenientes o desviaciones durante las pruebas, se realizan los ajustes necesarios en el diseño o en la configuración de las herramientas utilizadas. Es fundamental realizar iteraciones y pruebas adicionales hasta obtener resultados consistentes y satisfactorios.

Como primer paso en la construcción del proceso ETL, se importan las bibliotecas *pandas*, *numpy* y *date* proveniente de la biblioteca integrada *datetime*. Luego se establece la *mes* para ser usado como parametro de fecha y se realiza una lectura de los datos creando un *dataframe* de *pandas* para poder ser visualizados y procesados.

Como segundo paso se estableció el formato de fecha `'YYYY;MM;DD;HH;MM;SS.ss'` para la columna `'fecha evento'` y se eliminaron del *dataframe* todas las filas que tenían valores nulos.

Figura 4.1: Construcción proceso ETL con Python.

```
1  import pandas as pd
2  from datetime import date
3
4  ##Parametros
5  today = date.today()
6  mes = str(today.strftime("%Y%m"))
7
8  df_weblogs = pd.read_csv(r'files\input\Navegacion_nuevaweb_1_5jun_ofuscado.csv', sep=';', encoding='utf-8')
9
10 df_weblogs['fecha_evento'] = pd.to_datetime(df_weblogs['fecha_evento'], format="%b %d, %Y @ %H:%M:%S.%f")
11
12 df_weblogs = df_weblogs.dropna()
13
14 df_weblogs.to_csv(r'files\intermediate\web_logs_preproc_'+mes+'.csv', sep=';', index=False)
```

Fuente: Elaboración propia.

Para finalizar el proceso ETL se guarda el dataframe procesado en un archivo .csv dentro de los archivos del proyecto.

4.1.7. Monitoreo proceso ETL

Se establece un sistema de monitoreo para supervisar de manera continua el rendimiento del proceso ETL, permitiendo identificar y abordar posibles problemas a tiempo y garantizar la calidad de los datos. En esta etapa, también se realiza el mantenimiento del proceso, lo cual implica actualizaciones de las transformaciones, resolución de problemas y optimización del proceso.

El sistema de monitoreo juega un papel fundamental en la detección temprana de cualquier anomalía o disrupción en el flujo de datos. Mediante la implementación de métricas y alertas, se pueden supervisar aspectos clave como el tiempo de ejecución, el uso de recursos, los volúmenes de datos y la integridad de los resultados.

Además, el mantenimiento del proceso ETL implica la capacidad de adaptación a medida que evolucionan las necesidades del proyecto. Esto puede implicar la actualización de las transformaciones para reflejar cambios en las fuentes de datos o requerimientos del negocio, así como la solución de problemas que puedan surgir durante la ejecución del proceso.

Asimismo, se busca optimizar el proceso ETL a través de la identificación de posibles cuellos de botella o ineficiencias. Esto puede implicar ajustes en el diseño de las transformaciones, mejoras en la selección de herramientas o la optimización de los recursos utilizados.

CAPÍTULO 5: EXPLORATORY DATA ANALYSIS (EDA)

El análisis exploratorio de datos (EDA) es una etapa crucial en el proyecto de predicción del comportamiento en un entorno web. En este capítulo, adentraremos en el proceso de EDA, que nos permitirá desvelar patrones, tendencias y relaciones ocultas en los datos recopilados. A través de este análisis, obtendremos una visión profunda y enriquecedora que sentará las bases para un mejor entendimiento del comportamiento del usuario en el entorno web y facilitará la toma de decisiones estratégicas.

5.1. Introducción al EDA

El análisis exploratorio de datos (EDA, por sus siglas en inglés, Exploratory Data Analysis) es una fase fundamental en la investigación y comprensión de un conjunto de datos. Como su nombre lo indica, el EDA tiene como objetivo explorar y examinar los datos de manera detallada, utilizando resúmenes numéricos y visuales, con el fin de descubrir patrones, tendencias y características no anticipadas. Es considerado uno de los primeros pasos en el proceso de análisis, ya que proporciona una visión general de los datos antes de realizar un análisis más profundo (Ruiz, 2022).

El enfoque principal del EDA radica en el uso de herramientas y técnicas visuales y gráficas para revelar información clave sobre los datos en estudio (Parra, 2002). Estas técnicas incluyen el diagrama de tallo y hoja, el diagrama de caja y bigotes, y el diagrama de dispersión, entre otros. Al aplicar estas técnicas de análisis gráfico, podemos obtener una comprensión más profunda de la distribución y estructura de los datos, así como identificar relaciones entre las variables de interés. Además, el EDA nos brinda la capacidad de detectar posibles errores o puntos extremos, como anomalías, que podrían afectar la calidad de los resultados del análisis.

Los beneficios clave del análisis exploratorio de datos son los siguientes (Ruiz, 2022):

- **Conocer la distribución y estructura de los datos:** El EDA nos permite examinar la distribución de las variables y comprender cómo se organizan y dispersan los datos en el conjunto. Esto es fundamental para seleccionar las técnicas adecuadas de análisis estadístico y modelado.
- **Estudiar la relación entre variables:** Mediante el análisis de correlación y la visualización de patrones en los diagramas de dispersión, podemos explorar las relaciones entre las variables y comprender cómo interactúan entre sí. Esto nos brinda información valiosa para identificar posibles dependencias y tendencias en los datos.
- **Encontrar posibles errores y anomalías:** El EDA nos ayuda a identificar valores atípicos, datos faltantes u otros errores en los datos. Estas anomalías pueden tener un impacto significativo en los resultados del análisis, por lo que es importante detectarlas y tratarlas de manera adecuada.

5.2. Recopilación de datos

Los datos recopilados de los registros de navegación de los afiliados de AFP Capital constituyen una valiosa fuente de información para comprender el comportamiento y las preferencias de los usuarios en la plataforma web. Estos registros nos permiten analizar cómo interactúan los afiliados con los diferentes canales y métodos disponibles, así como realizar un seguimiento detallado de las fechas y horarios en que se llevan a cabo estas interacciones.

El dataset inicial entregado para este proyecto cuenta con 58,252 registros de navegación de usuarios, lo cual proporciona una cantidad significativa de información para su análisis. Antes de utilizar estos datos, se realizó un proceso de anonimización para proteger la privacidad de los usuarios, específicamente modificando el campo del rut para no mostrar el dato original. De esta manera, se garantiza que los registros sean tratados de forma confidencial y segura.

Los cuatro campos principales que conforman el conjunto de datos son el rut, la fecha del evento, el método y el canal. El rut, que ha sido modificado, actúa como un identificador único para cada usuario y permite realizar análisis individuales sin revelar su identidad. La fecha del evento registra el momento exacto en que se llevó a cabo cada navegación, lo cual es crucial para identificar patrones y tendencias a lo largo del tiempo. El campo del método describe la interacción específica realizada por el usuario en el canal correspondiente, proporcionando información detallada sobre las acciones que realizan. Por último, el campo del canal indica el sitio o ambiente particular en el cual tuvo lugar cada interacción, lo que puede ser útil para comprender las preferencias de los usuarios en relación con los diferentes entornos disponibles.

Con respecto al preprocesamiento de datos realizado hasta la fecha, se ha seguido el proceso ETL (Extracción, Transformación y Carga) que se describe en detalle en el capítulo anterior. Este proceso implica extraer los datos de las fuentes de origen, transformarlos en un formato adecuado y cargarlos en un sistema de almacenamiento para su posterior análisis. Se utilizaron diversas herramientas especializadas para llevar a cabo estas tareas, asegurando la calidad y coherencia de los datos procesados.

Es importante destacar que, si bien los datos recopilados ofrecen una valiosa perspectiva sobre el comportamiento de los usuarios en la plataforma web, es necesario tener en cuenta que existe un sesgo en la muestra de datos. En particular, los registros de navegación corresponden principalmente a afiliados con rentas altas. Esto implica que los usuarios con ingresos más altos, aquellos que cotizan por un valor elevado o el valor máximo, están sobrerrepresentados en la muestra. Por lo tanto, al interpretar y generalizar los resultados obtenidos, es fundamental tener en cuenta esta limitación y considerar posibles variaciones en el comportamiento de otros segmentos de usuarios.

5.3. Estadísticas descriptivas

Resumen de las principales estadísticas descriptivas de las variables relevantes en los logs de navegación. Análisis de la distribución de los datos, incluyendo medidas de centralidad y dispersión. Haciendo uso de la librería pandas podemos obtener

informacion respecto al dataframe, cantidad de valores únicos y calcular estadísticas descriptivas:

Podemos obtener informacion respecto a la cantidad de valores faltantes de cada dataframe que se está utilizando, los cuales son **36853**:

Figura 5.1: Datos faltantes dataframes.

Reporte de Valores Faltantes en df_weblogs1:

	Columna	Valores Faltantes
0	rut_cliente	0
1	fecha_evento	0
2	metodo	0
3	canal	0

Reporte de Valores Faltantes en df_weblogs2:

	Columna	Valores Faltantes
0	rut_cliente	0
1	fecha_evento	0
2	metodo	0
3	canal	0

Reporte de Valores Faltantes en df_weblogs3:

	Columna	Valores Faltantes
0	rut_cliente	0
1	fecha_evento	0
2	metodo	1
3	canal	36852

Fuente: Elaboración propia.

- **rut cliente:** Contiene datos de tipo int64.
- **fecha evento:** Contiene datos de tipo object.

- **metodo:** Contiene datos de tipo object.
- **canal:** Contiene datos de tipo object.

Además se puede conocer la cantidad de valores únicos que posee cada una de las columnas de los 2 datasets:

Dataset 1:

- **rut cliente:** Contiene 40125 valores únicos.
- **fecha evento:** Contiene 80181 valores únicos.
- **metodo:** Contiene 81254 valores únicos.
- **canal:** Contiene 81254 valores únicos.

Dataset 2:

- **rut cliente:** Contiene 40224 valores únicos.
- **fecha evento:** Contiene 80470 valores únicos.
- **metodo:** Contiene 81546 valores únicos.
- **canal:** Contiene 81546 valores únicos.

La librería pandas ofrece una buena cantidad de estadísticas descriptivas gracias a su función `.describe`, las cuales se muestran a continuación:

- **Recuento (count):** Calcula el número de valores no nulos en cada columna.
count: 58009.000000 valores no nulos.
- **Media (mean):** Calcula la media de las columnas numéricas. **mean: 504.854902**

- **Desviación estándar (std):** Calcula la desviación estándar de las columnas numéricas. **std: 553.964179**
- **Mínimo (min):** Calcula el valor mínimo de las columnas numéricas. **min: 1.000000**
- **Cuartiles:** Calcula los cuartiles de las columnas numéricas
 - **25 %:** 8.000000
 - **50 %:** 303.000000
 - **75 %:** 884.000000
- **Máximo (max):** Calcula el valor máximo de las columnas numéricas. **max: 1815.000000**

REFERENCIAS

- Afp capital*. (2023). Sitio web. Descargado de <https://www.afpcapital.cl/Paginas/default.aspx>
- Ajitesh, K. (2023). *Different types of time-series forecasting models*. Sitio web. Descargado de <https://vitalflux.com/different-types-of-time-series-forecasting-models/>
- Amazon. (2023). *¿qué es una red neuronal?* Sitio Web. Descargado de <https://aws.amazon.com/es/what-is/neural-network/#:~:text=Una%20red%20neuronal%20es%20un,lo%20hace%20el%20cerebro%20humano.>
- Arana, C. (2021). *Modelos de aprendizaje automático mediante árboles de decisión*.
- Baker, K. (2023). *What is attribution modeling and why its so important*. Sitio web. Descargado de <https://blog.hubspot.com/marketing/attribution-modeling>
- Contreras Arteaga, A., y Sánchez Cotrina, F. (2019). *Analítica predictiva para conocer el patrón de consumo de los clientes en la empresa cienpharma s.a.c utilizando ibm spss modeler y la metodología crisp-dm*. (p.15)
- dask: Scale the python tools you love*. (2023). Sitio Web. Descargado de <https://www.dask.org/>
- Elizabeth, F. (2023). *Cómo crear un modelo de recomendación basado en machine learning*. Sitio web. Descargado de <https://aws.amazon.com/es/blogs/aws-spanish/como-crear-un-modelo-de-recomendacion-basado-en-machine-learning/>

- Eric. (2023). *Data science life cycle: Crisp-d, and osemn frameworks*. Sitio web. Descargado de <https://datarundown.com/data-science-life-cycle/>
- Filzinger, T. (2023). *Lstm long short-term memory - memoria a corto plazo de larga duración*. Sitio Web. Descargado de <https://konfuzio.com/es/lstm/#:~:text=La%20LSTM%20es%20un%20tipo,periodos%20de%20tiempo%20m%C3%A1s%20largos>.
- Garcia-Estrella, C., y Barón Ramírez, E. (2021). *La inteligencia de negocios y la analítica de datos en los procesos empresariales*.
- Hyndman, R. J., y Athanasopoulos, G. (2023). *Forecasting: principles and practice, 2nd edition*. Sitio Web. Descargado de [OTexts.com/fpp2](https://otexts.com/fpp2)
- IBM. (2023). *¿qué es el random forest?* Sitio web. Descargado de <https://www.ibm.com/mx-es/topics/random-forest>
- Kimball, R., y Caserta, J. (2004). *The data warehouse etl toolkit*. Wiley Publishing, Inc.
- Lemon, K. N., y Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*.
- Mckinney, W. (2011). pandas: a foundational python library for data analysis and statistics. *Python for high performand and scientific computing*, 14(9), 1-9.
- Morales Oñate, V. (2022). *Series de tiempo*. Sitio Web. Descargado de https://bookdown.org/victor_morales/SeriesdeTiempo/%7D%7D
- Parra, J. (2002). Análisis exploratorio y análisis confirmatorio de datos. *Espacio Abierto*, 11(1), 115-124.
- Rasekhi, S., Fard, H., y Kim, D. (2016). *Understanding the role of trust in pension system: A literature review*. The Third Wave. William Morrow and company Inc.

- Rodriguez, J. (2023). *¿qué es un cliente?* Sitio web. Descargado de <https://blog.hubspot.es/sales/que-es-un-cliente>
- Ruiz, N. (2022). *Una aproximación al análisis exploratorio de datos* (Tesis Doctoral no publicada). Universidad de Valladolid.
- Sánchez, W. (2011). La usabilidad en ingeniería de software: definición y características. *Revista de Ingeniería e Innovación de la Facultad de Ingeniería, Universidad Don Bosco*, 1(2), 7-21.
- Toffler, A. (1980). *Understanding the role of trust in pension system: A literature review*.
- van der Walt, S., Colbert, S. C., y Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science and Engineering, Institute of Electrical and Electronics Engineers*, 13(2), 22-30. doi: 10.1109/MCSE.2011.37
- van Rossum, G., y Drake, F. L. (2003). *An introduction to python*. Network Theory Limited.
- Vatsal. (2021). *Recommendation systems explained*. Sitio web. Descargado de <https://towardsdatascience.com/recommendation-systems-explained-a42fc60591ed>
- Zheng, B., Thompson, K., Lam, S. S., Yoon, S. W., y Gnanasambandam, N. (2013). *Customers' behavior prediction using artificial neural network*.