

**UNIVERSIDAD TECNOLÓGICA DEL VALLE DE TOLUCA**

**DIRECCIÓN DE CARRERA DE TECNOLOGÍAS DE LA INFORMACIÓN Y  
COMUNICACIÓN**

**INGENIERÍA EN DESARROLLO Y GESTIÓN DE SOFTWARE**

**NOMBRE LA ASIGNATURA:**

**EXTRACCIÓN DEL CONOCIMIENTO EN BASE DE DATOS**

**NOMBRE DE LA PROFESORA:**

**DOLORES NELLY GUTIERREZ MATA**

**“TERCER ENTREGABLE”**

**CYNTHIA PAOLA DURO SANCHEZ      221811737**

**CAROLINA DÍAZ ROMERO              221811725**

**FRANCISCA CAMPOS QUIÑONES      221811678**

**BRENDA XIMENA DURO SÁNCHEZ    221712243**

**GRUPO**

**IDGS – 91**

**CUATRIMESTRE**

**90**

**LUGAR**

**SANTA MARÍA ATARASQUILLO, LERMA, MÉXICO.**

**MAYO -AGOSTO 2022**

## Tabla de contenido

<u>ANÁLISIS COMPARATIVO</u>	4
<u>INTELIGENCIA ARTIFICIAL</u>	4
<u>MACHINE LEARNING</u>	5
<u>DATA MINING</u>	6
<u>BIG DATA</u>	7
<u>CASOS DE APLICACIÓN Y LENGUAJES Y HERRAMIENTAS UTILIZADAS.</u>	8
<u>INTELIGENCIA ARTIFICIAL</u>	8
<u>MACHINE LEARNING</u>	9
<u>DATA MINING</u>	10
<u>BIG DATA</u>	11
<u>CASO DE ESTUDIO</u>	13
<u>OBJETIVO Y ALCANCE DE CASO.</u>	13
<u>JUSTIFICACIÓN DE LA METODOLOGÍA A UTILIZAR PARA EL ANÁLISIS DE DATOS.</u>	14
<u>PLANEACIÓN DE LAS ETAPAS PARA EL ANÁLISIS DE DATOS.</u>	16
<u>ANTECEDENTES</u>	16
<u>OBJETIVOS</u>	17
<u>METODOLOGÍA</u>	18
<u>RESULTADOS</u>	20
<u>ESQUEMA DATA WERE HOUSE</u>	21
<u>TIPOS Y FUENTES DE DATOS</u>	22
<u>TÉCNICAS DE LIMPIEZA DE DATOS</u>	24
<u>DATA CLEANING</u>	24
<u>ETL</u>	24
<u>JUSTIFICACION DEL ALGORITMO (CLASIFICACION)</u>	28
<u>DEFINICIÓN</u>	28
<u>DESCRIPCION DEL MODELO DE DISEÑO DE CLASIFICACION</u>	30
<u>ALGORITMOS SUPERVISADOS</u>	31
<u>IMPLEMENTACIÓN DEL ALGORITMO.</u>	32
<u>RESULTADOS DEL ALGORITMO.</u>	32
<u>MÉTRICAS DE CLASIFICACIÓN</u>	33

<u>CURVA ROC</u>	33
<u>EXHAUSTIVIDAD</u>	35
<u>CONCLUSIONES</u>	37



## ANÁLISIS COMPARATIVO

### INTELIGENCIA ARTIFICIAL

	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.
INTELIGENCIA ARTIFICIAL	<ol style="list-style-type: none"> <li>1. Eliminación de tareas monótonas</li> <li>2. Manejo de una gran cantidad de datos</li> <li>3. Imitación de la cognición humana</li> <li>4. Son futuristas</li> </ol>	<p>De acuerdo con la Revista Forbes, este avance tecnológico se emplea para la resolución de problemas en todos los ámbitos. Sus beneficios incluyen el aumento de ventas, la detección de fraudes y la automatización de procesos, entre otros. Entre sus aplicaciones más usuales, se encuentran:</p> <ul style="list-style-type: none"> <li>• La robótica.</li> <li>• La domótica o casas inteligentes.</li> <li>• Las redes neuronales artificiales.</li> <li>• Los chatbots.</li> <li>• El reconocimiento de voz o facial.</li> </ul>	<ul style="list-style-type: none"> <li>• Drones.</li> <li>• Big Data.</li> <li>• Prevención contra la corrupción.</li> <li>• Blockchain.</li> </ul>	<ol style="list-style-type: none"> <li>1. Python</li> <li>2. C++</li> <li>3. R</li> <li>4. Java</li> <li>5. Prolog</li> </ol>

## MACHINE LEARNING

MACHINE LEARNING	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.
	<ul style="list-style-type: none"> <li>• El machine learning está íntimamente relacionado con el reconocimiento de patrones.</li> <li>• El aprendizaje automático es un campo de las ciencias de la información.</li> <li>• Los algoritmos de machine learning aprenden de manera autónoma.</li> <li>• Un agente inteligente es capaz de predecir eventos a partir de datos históricos.</li> <li>• Un sistema de machine learning mejora constantemente con el tiempo.</li> <li>• Existen una gran diversidad de algoritmos de aprendizaje automático, pero unos son más utilizados que otros.</li> </ul>	<ol style="list-style-type: none"> <li>1. Mayor conocimiento de los clientes. Al contar con el Machine Learning, es posible determinar gustos, hábitos y necesidades de compra de los clientes. Esto mejora la experiencia del cliente y facilita su fidelización.</li> <li>2. Desarrollo del e-commerce. Mediante el conocimiento de los clientes, el Machine Learning puede determinar cuáles son los productos con mayor o menor demanda, así como las temporadas ideales para promociones y descuentos.</li> <li>3. Predicción de tendencias y necesidades. Permite anticipar los movimientos en la demanda, así como las necesidades que generará un producto o servicio, con la finalidad de desarrollar un auxiliar.</li> </ol>	<ul style="list-style-type: none"> <li>• Sistemas de reconocimiento facial.</li> <li>• Desarrollo de bots en el área de los videojuegos.</li> <li>• Reconocimiento automático del habla.</li> <li>• Motores de búsquedas.</li> <li>• Diagnósticos médicos.</li> <li>• Predicción de tránsito vehicular a una hora dada.</li> <li>• Entendimiento de textos.</li> <li>• Anticipación de fallos en maquinarias.</li> <li>• Sistemas de visión artificial.</li> <li>• Robótica.</li> </ul>	<ol style="list-style-type: none"> <li>1. R</li> <li>2. C ++</li> <li>3. JavaScript</li> <li>4. Java</li> <li>5. C</li> <li>6. Julia</li> <li>7. Shell</li> <li>8. TypeScript</li> <li>9. Scala.</li> </ol>

## DATA MINING

DATA MINING	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.
	<ul style="list-style-type: none"> <li>• Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.</li> <li>• En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet. El entorno de la minería de datos suele tener una arquitectura cliente/servidor.</li> <li>• Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados</li> <li>• El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias para efectuar preguntas adhoc y obtener rápidamente respuestas.</li> </ul>	<ol style="list-style-type: none"> <li>1. La minería de datos descubre información que no se esperaba obtener. Como muchos modelos diferentes son usados, algunos resultados inesperados tienden a aparecer. Las combinaciones de distintas técnicas otorgan efectos inesperados que se transforma en un valor añadido a la empresa.</li> <li>2. Enormes bases de datos pueden ser analizadas mediante la tecnología de data mining.</li> <li>3. Los resultados son fáciles de entender: personas sin un conocimiento previo en ingeniería informática pueden interpretar los resultados con sus propias ideas</li> <li>4. Contribuye a la toma de decisiones tácticas y estratégicas para detectar la información clave</li> </ol>	<ul style="list-style-type: none"> <li>• 'Marketing'. La minería de datos se utiliza para explorar bases de datos cada vez mayores y mejorar la segmentación del mercado</li> <li>• Comercio minorista. Los supermercados, por ejemplo, emplean los patrones de compra conjunta para identificar asociaciones de productos y decidir cómo situarlos en los diferentes pasillos y estanterías de los lineales</li> <li>• Banca. Los bancos recurren a la minería de datos para entender mejor los riesgos del mercado.</li> <li>• Medicina. La minería de datos favorece diagnósticos más precisos.</li> <li>• Televisión y radio. Hay cadenas que aplican la minería de datos en tiempo real a sus registros de audiencia en televisión online (IPTV) y radio.</li> </ul>	<ol style="list-style-type: none"> <li>1. Xplenty</li> <li>2. Weka</li> <li>3. Rapid Miner</li> <li>4. Teradata</li> <li>5. Orange</li> <li>6. Revolution</li> <li>7. Dundas</li> </ol>

## BIG DATA

BIG DATA	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes herramientas.	y
	<ul style="list-style-type: none"> <li>• Volumen. Cada vez son más los sistemas de producción de datos, desde las redes sociales hasta objetos como asistentes de voz para el hogar o pulseras de actividad.</li> <li>• Velocidad. Ritmo al que crece y se procesa la información.</li> <li>• Veracidad. Este factor es clave a la hora de llevar a cabo el análisis de datos</li> <li>• Variedad. La información recolectada puede venir de diferentes fuentes, así como ser de diversos tipos (estructurados, no estructurados).</li> <li>• Valor. Es una cualidad fundamental en el análisis de datos, ya que, a pesar de tener una gran cantidad de información, ésta pocas veces ofrece contenido útil.</li> <li>• Visualización. Todo este conocimiento no tiene sentido si no se establecen conclusiones sobre el mismo, creando gráficos que muestren de manera rápida y sencilla los resultados.</li> </ul>	<ul style="list-style-type: none"> <li>• Reducción de costes. Las nuevas tecnologías hacen que ya no se requiera de un servidor donde alojar los datos como se hacía tradicionalmente.</li> <li>• Rapidez. Hoy en día existe un mayor volumen de información, pero también herramientas de procesamiento de datos más ágiles.</li> <li>• Fidelización de clientes. Tener un alto volumen de información de los clientes permite conocer su comportamiento y necesidades</li> </ul>	<p>1. Personalización y transparencia hacia el consumidor</p> <p>Aunque todavía queda mucho por hacer, el ámbito del marketing y las relaciones con clientes y consumidores es uno de los que ha experimentado una mayor aplicación práctica del big data.</p> <p>2. Salud</p> <p>Sin duda, en una sociedad cada vez más envejecida y con una esperanza de vida creciente, este tipo de herramientas tienen mucho que aportar.</p> <p>3. Sostenibilidad e igualdad</p> <p>Según el informe <i>Smart Cities</i> de McKinsey, el big data puede tener importantes aplicaciones en el ámbito de la sostenibilidad, mejorando diversos indicadores de calidad de vida entre un 10% y un 30%. .</p>	<ul style="list-style-type: none"> <li>• Apache Hadoop.</li> <li>• Elasticsearch.</li> <li>• Apache Storm.</li> <li>• MongoDB.</li> <li>• Apache Spark.</li> <li>• Python.</li> <li>• Apache Cassandra.</li> <li>• Lenguaje R.</li> </ul>	



## CASOS DE APLICACIÓN Y LENGUAJES Y HERRAMIENTAS UTILIZADAS.

### INTELIGENCIA ARTIFICIAL

	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.	y
INTELIGENCIA ARTIFICIAL	<ul style="list-style-type: none"> <li>Manejo de una gran cantidad de datos: Almacenamiento de registros y eliminaciones de registros dentro del sistema, base de datos etc.</li> <li>Resiliencia: el sistema es apto para ser optimizado a futuro y tener una optimización constante.</li> <li>Buen rendimiento, esto es la posibilidad de manejar eficientemente gran cantidad de información.</li> </ul>	<ul style="list-style-type: none"> <li>Aumento de la eficacia, a la hora de manejar gran cantidad de información lo que garantiza.</li> <li>Gestión y control del sistema, dentro de la gestión de información.</li> <li>Desarrollo de nuevos y más innovadores productos.</li> <li>Generación de ventas y simplificación del ciclo de ventas.</li> </ul>	<ul style="list-style-type: none"> <li>Personalización de productos alojados dentro del sistema.</li> <li>Automatización de nuestros servicios.</li> <li>Optimización en la forma de ventas y personalización de cada producto y sus precios.</li> </ul>	<ol style="list-style-type: none"> <li>1. PHP</li> <li>2. JavaScript</li> </ol>	

## MACHINE LEARNING

MACHINE LEARNING	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.
	<ul style="list-style-type: none"> <li>• Agiliza los procesos de desarrollo de E-commerce.</li> <li>• Agilización de los procesos de venta de productos.</li> <li>• Ciberseguridad: Teniendo en cuenta que la mayoría de malwares utilizan código similar, el aprendizaje automático puede evitar fácilmente que los ataques se repitan.</li> </ul>	<ul style="list-style-type: none"> <li>• Machine Learning puede ser una herramienta para establecer nexos dinámicos entre negocios y clientes dentro del software de ventas.</li> <li>• Simplificación y rapidez en la obtención de datos e información de nuestros productos.</li> <li>• Mejoramiento en la relación con nuestros clientes.</li> <li>• Desarrollo y mejora de e-commerce.</li> </ul>	<ul style="list-style-type: none"> <li>• Nos arroja datos partiendo de información suministrada, encontrando patrones de comportamiento.</li> <li>• Uso de Gmail Para mantener al usuario protegido de virus y de recibir correos sospechosos o fraudulentos, la plataforma de email de Google integra el Machine Learning para evitar el correo no deseado (o spam) en la bandeja de entrada.</li> </ul>	10.PHP 11.JavaScript

## DATA MINING

DATA MINING	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.
	<ul style="list-style-type: none"> <li>• Explorar los datos se encuentran en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.</li> <li>• En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos; en otros, se mantienen en servidores de Internet e Intranet. El entorno de la minería de datos suele tener una arquitectura cliente/servidor.</li> <li>• Las herramientas de la minería de datos ayudan a extraer el mineral de la información enterrado en archivos corporativos o en registros públicos, archivados</li> <li>• El minero es, muchas veces un usuario final con poca o ninguna habilidad de programación, facultado por barrenadoras de datos y otras poderosas herramientas indagatorias para efectuar preguntas adhoc y obtener rápidamente respuestas.</li> </ul>	<ol style="list-style-type: none"> <li>1. La minería de datos descubre información que no se esperaba obtener. Como muchos modelos diferentes son usados, algunos resultados inesperados tienden a aparecer. Las combinaciones de distintas técnicas otorgan efectos inesperados que se transforma en un valor añadido a la empresa.</li> <li>2. Enormes bases de datos pueden ser analizadas mediante la tecnología de data mining.</li> <li>3. Los resultados son fáciles de entender: personas sin un conocimiento previo en ingeniería informática pueden interpretar los resultados con sus propias ideas</li> <li>4. Contribuye a la toma de decisiones tácticas y estratégicas para detectar la información clave</li> </ol>	<ul style="list-style-type: none"> <li>• 'Marketing'. La minería de datos se utiliza para explorar bases de datos cada vez mayores y mejorar la segmentación del mercado</li> <li>• Comercio minorista. Los supermercados, por ejemplo, emplean los patrones de compra conjunta para identificar asociaciones de productos y decidir cómo situarlos en los diferentes pasillos y estanterías de los lineales</li> <li>• Banca. Los bancos recurren a la minería de datos para entender mejor los riesgos del mercado.</li> <li>• Medicina. La minería de datos favorece diagnósticos más precisos.</li> <li>• Televisión y radio. Hay cadenas que aplican la minería de datos en tiempo real a sus registros de audiencia en televisión online (IPTV) y radio.</li> </ul>	<p><b>1.SQL</b> Los conjuntos de datos a gran escala pueden contener millones de filas, lo que dificulta encontrar con precisión los datos que se necesitan. SQL es un lenguaje de consulta que permite ajustar, localizar y comprobar conjuntos de datos masivos. Al ser un lenguaje de dominio específico, es conveniente para gestionar bases de datos relacionales.</p> <p><b>2.JAVASCRIPT</b> Está estrechamente relacionado con el desarrollo y las aplicaciones web, y aporta la capacidad de construir páginas web vibrantes al mundo de las visualizaciones de datos. Es otra opción de propósito general para los científicos de datos con una buena selección de paquetes y una gran integración web.</p>

## BIG DATA

BIG DATA	Características.	Beneficios, restricciones y retos	Casos de aplicación.	Lenguajes y herramientas.
	<ul style="list-style-type: none"> <li>• Volumen. Cada vez son más los sistemas de producción de datos, desde las redes sociales hasta objetos como asistentes de voz para el hogar o pulseras de actividad.</li> <li>• Velocidad. Ritmo al que crece y se procesa la información.</li> <li>• Veracidad. Este factor es clave a la hora de llevar a cabo el análisis de datos</li> <li>• Variedad. La información recolectada puede venir de diferentes fuentes, así como ser de diversos tipos (estructurados, no estructurados).</li> <li>• Valor. Es una cualidad fundamental en el análisis de datos, ya que, a pesar de tener una gran cantidad de información, ésta pocas veces ofrece contenido útil.</li> <li>• Visualización. Todo este conocimiento no tiene</li> </ul>	<ul style="list-style-type: none"> <li>• Reducción de costes. Las nuevas tecnologías hacen que ya no se requiera de un servidor donde alojar los datos como se hacía tradicionalmente.</li> <li>• Rapidez. Hoy en día existe un mayor volumen de información, pero también herramientas de procesamiento de datos más ágiles.</li> <li>• Fidelización de clientes. Tener un alto volumen de información de los clientes permite conocer su comportamiento y necesidades</li> </ul>	<ul style="list-style-type: none"> <li>• La ciberseguridad también puede sacar partido de las aplicaciones del Big Data, puesto que gracias al análisis inteligente de datos es posible establecer posibles relaciones ocultas, detectar patrones de conducta y prevenir amenazas a la seguridad. No solo hablamos de predecir y prevenir ataques en base a la información extraída de los datos, sino también prevenir el fraude gracias a la comprobación en tiempo real del historial de una cuenta, lo que puede ayudar a detectar el comportamiento anómalo de un usuario o una transacción sospechosa. La información que aporta el Big Data en el</li> </ul>	<p><b>1.PHP</b></p> <p>es un lenguaje de programación muy utilizado para el desarrollo web gracias a su sintaxis sencilla y ser multiplataforma. Esto hace que sea ideal para proyectos de Inteligencia Artificial que tengan que ejecutarse en un navegador.</p> <p><b>2.SQL</b></p> <p>utilizado en una amplia gama de aplicaciones, por lo que es un lenguaje muy útil para estar familiarizado.</p>

sentido si no se establecen conclusiones sobre el mismo, creando gráficos que muestren de manera rápida y sencilla los resultados.

sector de la seguridad informática gracias al análisis del tráfico de la red puede ayudar a descubrir amenazas y prevenir ataques de hackers, el espionaje industrial, el fraude cibernético e incluso el ciberterrorismo. En este caso, podemos decir que el análisis de datos masivos puede ayudar a salvar vidas cuando se evita el ciberataque a sistemas e infraestructuras críticas.

SQL es más útil como lenguaje de procesamiento de datos que como herramienta analítica avanzada.

### 3. JAVA

Hay mucho que decir para aprender Java como un lenguaje de ciencia de datos de primera elección. Muchas compañías apreciarán la capacidad de poder integrar el código de producción de ciencia de datos directamente en la base de un código ya existente, y además encontramos que el rendimiento de Java

# CASO DE ESTUDIO

## OBJETIVO Y ALCANCE DE CASO.

### OBJETIVO Y ALCANCE

#### OBJETIVO

##### PRINCIPAL:

Brindarle al cliente la venta de zapatos por medio del comercio electrónico, pretende ayudar a sus clientes a que puedan realizar compra de calzado en línea de forma rápida, sencilla y práctica, como permitir al usuario realizar la selección de los productos que más se adapten a sus necesidades y de esa manera, tener una experiencia agradable a la hora de interactuar con nuestro sistema.

##### ESPECIFICOS:

- Controla el inventario de las zapaterías, sucursales y almacenes en tiempo real y sincronizado.
- Gestiona el stock en tiempo real, que puedes conectar a tu tienda online, si la posees.
- Genera informes de requerimientos que te indican cuales productos debes adquirir para reponer tus inventarios.
- Realiza proyecciones, usando la información histórica de tus ventas, y las tendencias recientes. También te ayuda a mantener tus inventarios con un índice de rotación ideal.
- Ofrece trazabilidad a tus productos en el almacén y sucursales. Si es requerido, puede generar un proceso de etiquetado sencillo y eficiente.
- Brinda seguridad a tu inventario, facilitando la detección de faltantes en tu mercancía. Puedes comparar de forma sencilla el inventario físico y registrado en el sistema, para realizar oportunamente los ajustes necesarios.

#### ALCANCE

El planteamiento de este proyecto se realizará para generar y desarrollar un sistema que permita a los usuarios brindarles la posibilidad de realizar la compra de zapatos en línea, de forma rápida, sencilla y efectiva.

## JUSTIFICACIÓN DE LA METODOLOGÍA A UTILIZAR PARA EL ANÁLISIS DE DATOS.

La metodología para utilizar es data mining ya que el sistema está dentro de la categoría e-commerce por lo tanto esta metodología nos ayuda a cumplir con las siguientes tareas:

- Segmentación de mercado
- Analizar las demandas
- Crear perfiles de compradores
- Analizar carrito de compra
- Calcular los precios de los productos
- Identificar fallos en los procesos de venta
- Elaborar un pronóstico sobre el vencimiento de los contratos

Sus métodos nos ayudan en:

- Clasificación: clasifica los datos individuales en categorías específicas definidas previamente.
- Análisis de valores atípicos o de desviaciones: identifica a los objetos que no cumplen las reglas de dependencia en objetos emparentados.
- Análisis de clústeres: identifica la concentración de similitudes y construye grupos de objetos que comparten una serie de características comunes en comparación con otros grupos
- Análisis de correlación: descubre correlaciones entre dos o más objetos independientes que, aunque no muestran ningún tipo de relación directa.
- Análisis de la regresión: destapa las relaciones entre una variable dependiente (por ejemplo, los análisis de cifras de ventas de productos) y una o varias variables independientes (el precio del producto o los ingresos del comprador) con el objetivo de realizar una serie de pronósticos sobre la variable dependiente (pronóstico de ventas).

Las herramientas que se utilizan son:

El Framework por utilizar es Laravel: Nos ayuda en muchas cosas al desarrollar una aplicación, por medio de su sistema de paquetes y de ser un framework del tipo MVC (Modelo-Vista-Controlador) da como resultado que podamos

“despreocuparnos” por instanciar clases y métodos para usarlos en muchas partes de nuestra aplicación sin la necesidad de escribirlo y repetirlo muchas veces con lo que eso conlleva a la hora de modificar algo en el código.

Lenguaje para el desarrollo es JavaScript: Es un lenguaje de programación ligero, interpretado, o compilado justo-a-tiempo (just-in-time) con funciones de primera clase. Lenguaje de programación basada en prototipos, multiparadigma, de un solo hilo, dinámico, con soporte para programación orientada a objetos, imperativa y declarativa.

Servidor de base de datos SQL: SQL es un lenguaje de computación para trabajar con conjuntos de datos y las relaciones entre ellos. Los programas de bases de datos relacionales. Se usa para describir conjuntos de datos que pueden ayudarle a responder preguntas. Al usar SQL, debe usar la sintaxis correcta. La sintaxis es el conjunto de reglas mediante las que se combinan correctamente los elementos de un idioma.



## **PLANEACIÓN DE LAS ETAPAS PARA EL ANÁLISIS DE DATOS.**

### **ANTECEDENTES**

Calzado Ma Jo es un negocio de calzado de dama desarrollado por la familia Fernández, con el objetivo de ofrecer a los clientes calzado de buena calidad, se dieron a conocer por medio de ventas dentro de la plaza azul en el municipio de San Mateo Atento.

En el inicio de la pandemia por protocolos de seguridad cerraron dicha plaza, los encargados de calzado Ma Jo decidieron ofrecer sus servicios mediante la aplicación Facebook, dentro de esta aplicación ellos dan a conocer sus productos, mediante el chat atienden a sus clientes otorgando información acerca de los productos de su interés y haciendo entrega en puntos medios del municipio de San Mateo Atenco.

Nombre: Zapatería “Calzado Ma Jo”

Ubicación: 5 de Mayo 613, Barrio de la Concepción, 52105 San Mateo Atenco, Méx.

Teléfono: 722 512 4825

Facebook: <https://www.facebook.com/Calzado-Ma-Jo-100486658304339/>

## OBJETIVOS

### PRINCIPAL:

Brindarle al cliente la venta de zapatos por medio del comercio electrónico, pretende ayudar a sus clientes a que puedan realizar compra de calzado en línea de forma rápida, sencilla y práctica, como permitir al usuario realizar la selección de los productos que más se adapten a sus necesidades y de esa manera, tener una experiencia agradable a la hora de interactuar con nuestro sistema.

### ESPECIFICOS:

- Controla el inventario de las zapaterías, sucursales y almacenes en tiempo real y sincronizado.
- Gestiona el stock en tiempo real, que puedes conectar a tu tienda online, si la posees.
- Genera informes de requerimientos que te indican cuales productos debes adquirir para reponer tus inventarios.
- Realiza proyecciones, usando la información histórica de tus ventas, y las tendencias recientes. También te ayuda a mantener tus inventarios con un índice de rotación ideal.
- Ofrece trazabilidad a tus productos en el almacén y sucursales. Si es requerido, puede generar un proceso de etiquetado sencillo y eficiente.
- Brinda seguridad a tu inventario, facilitando la detección de faltantes en tu mercancía. Puedes comparar de forma sencilla el inventario físico y registrado en el sistema, para realizar oportunamente los ajustes necesarios.

## METODOLOGÍA

La metodología para utilizar es data mining ya que el sistema esta dentro de la categoría e-commerce por lo tanto esta metodología nos ayuda a cumplir con las siguientes tareas:

- Segmentación de mercado
- Analizar las demandas
- Crear perfiles de compradores
- Analizar carrito de compra
- Calcular los precios de los productos
- Identificar fallos en los procesos de venta
- Elaborar un pronóstico sobre el vencimiento de los contratos

Sus métodos nos ayudan en:

- Clasificación: clasifica los datos individuales en categorías específicas definidas previamente.
- Análisis de valores atípicos o de desviaciones: identifica a los objetos que no cumplen las reglas de dependencia en objetos emparentados.
- Análisis de clústeres: identifica la concentración de similitudes y construye grupos de objetos que comparten una serie de características comunes en comparación con otros grupos
- Análisis de correlación: descubre correlaciones entre dos o más objetos independientes que, aunque no muestran ningún tipo de relación directa.

- Análisis de la regresión: destapa las relaciones entre una variable dependiente (por ejemplo, los análisis de cifras de ventas de productos) y una o varias variables independientes (el precio del producto o los ingresos del comprador) con el objetivo de realizar una serie de pronósticos sobre la variable dependiente (pronóstico de ventas).

Las herramientas que se utilizan son:

El Framework por utilizar es Laravel: Nos ayuda en muchas cosas al desarrollar una aplicación, por medio de su sistema de paquetes y de ser un framework del tipo MVC (Modelo-Vista-Controlador) da como resultado que podamos “despreocuparnos” por instanciar clases y métodos para usarlos en muchas partes de nuestra aplicación sin la necesidad de escribirlo y repetirlo muchas veces con lo que eso conlleva a la hora de modificar algo en el código.

Lenguaje para el desarrollo es JavaScript: Es un lenguaje de programación ligero, interpretado, o compilado justo-a-tiempo (just-in-time) con funciones de primera clase. Lenguaje de programación basada en prototipos, multiparadigma, de un solo hilo, dinámico, con soporte para programación orientada a objetos, imperativa y declarativa.

Servidor de base de datos SQL: SQL es un lenguaje de computación para trabajar con conjuntos de datos y las relaciones entre ellos. Los programas de bases de datos relacionales. Se usa para describir conjuntos de datos que pueden ayudarle a responder preguntas. Al usar SQL, debe usar la sintaxis correcta. La sintaxis es el conjunto de reglas mediante las que se combinan correctamente los elementos de un idioma.

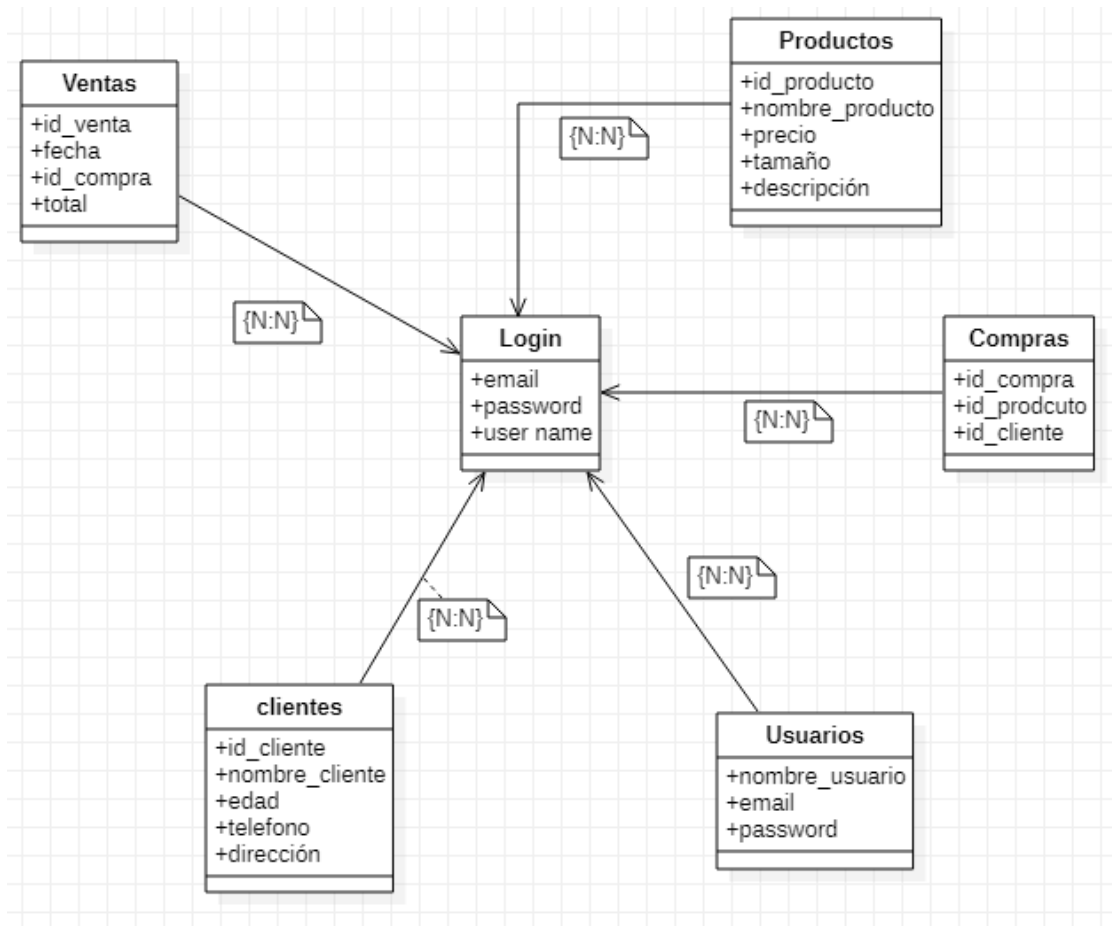
## RESULTADOS

RESULTADOS	
ESPERADOS	<ol style="list-style-type: none"><li>1. Agilizar el manejo del inventario y ventas de una zapatería local.</li><li>2. Implementar un sistema desde el cual se pueda hacer la creación, modificación y eliminación de clientes y usuarios.</li><li>3. Implementar un sistema que permita el alta, pausa y venta de productos.</li><li>4. Implementar un sistema que cuente con un carrito de compras que agilice el proceso de la venta.</li><li>5. Implementar un sistema que cuente con un módulo de logueo para la verificación de datos del cliente.</li></ol>
OBTENIDOS	<ol style="list-style-type: none"><li>1. Prototipo de sistema para el manejo del inventario y ventas de una zapatería local.</li><li>2. Sistema que permite la creación y manipulación de información de clientes y usuarios.</li><li>3. Sistema que permite la creación y eliminación de productos.</li><li>4. Sistema que cuenta con un carrito de compras para agilizar procesos.</li><li>5. Sistema con un módulo de logueo para la verificación de datos del cliente.</li></ol>

## ESQUEMA DATA WERE HOUSE

Dentro del Esquema DWH de estrella se encuentra una tabla de hechos que contiene los datos para el análisis, rodeada de las tablas de dimensiones.

Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales.

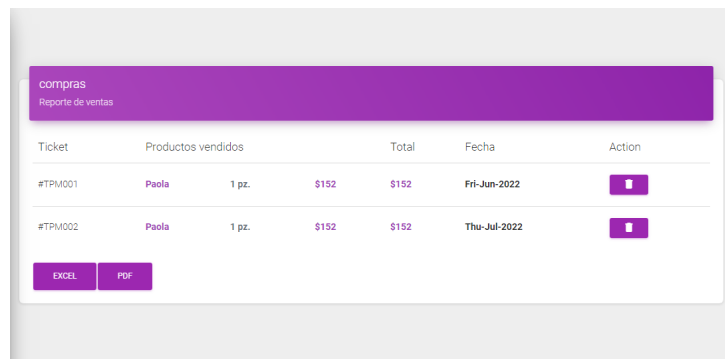


Este esquema nos ayuda a separar los datos del proceso de negocios en: hechos y dimensiones. Los hechos contienen datos medibles, cuantitativos, relacionados con la transacción del negocio, y las dimensiones son atributos que describen los datos indicados en los hechos

## TIPOS Y FUENTES DE DATOS

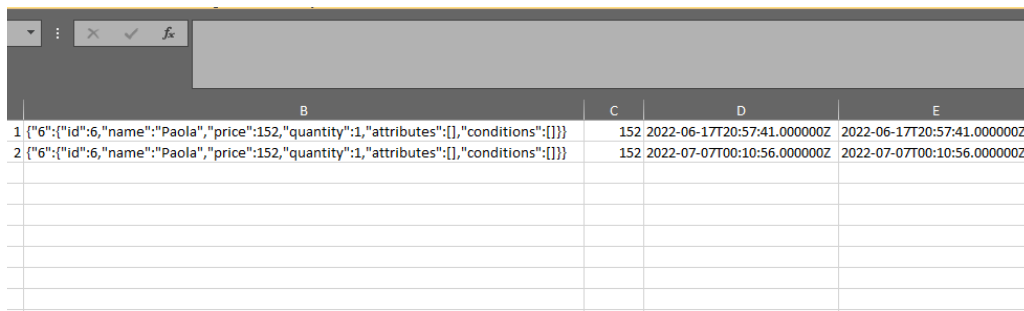
Tipos:

- Datos de transacciones



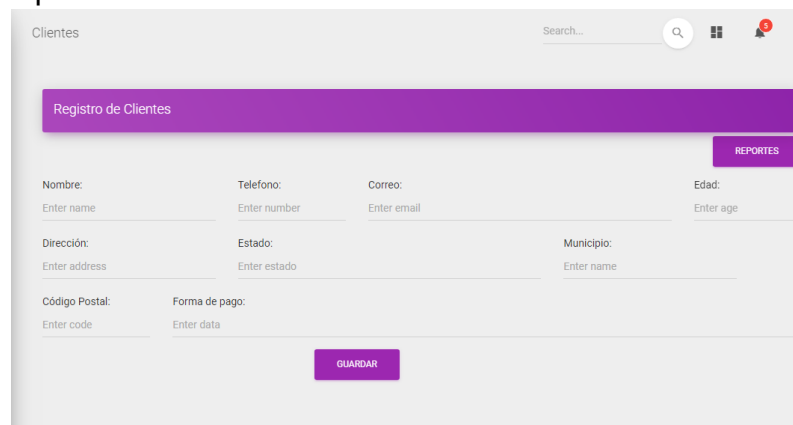
Ticket	Productos vendidos	Total	Fecha	Action
#TPM001	Paola 1 pz.	\$152	Fri-Jun-2022	
#TPM002	Paola 1 pz.	\$152	Thu-Jul-2022	

EXCEL PDF



	B	C	D	E
1	[{"id":6,"name":"Paola","price":152,"quantity":1,"attributes":[],"conditions":[]}]	152	2022-06-17T20:57:41.000000Z	2022-06-17T20:57:41.000000Z
2	[{"id":6,"name":"Paola","price":152,"quantity":1,"attributes":[],"conditions":[]}]	152	2022-07-07T00:10:56.000000Z	2022-07-07T00:10:56.000000Z

- Generados por los humanos



Clientes Search...

Registro de Clientes

REPORTES

Nombre: Enter name Telefono: Enter number Correo: Enter email Edad: Enter age

Dirección: Enter address Estado: Enter estado Municipio: Enter name

Código Postal: Enter code Forma de pago: Enter data

GUARDAR

- Web y medios sociales



Web y medios sociales para el desarrollo de sistema de ventas:

Este tipo de fuente de datos nos beneficia ya que Son los que se originan en la red y configuran, según los expertos, el trozo más grande del pastel llamado Big Data y es una de las fuentes de datos más utilizadas en la actualidad. Hablamos de la información que se genera sobre clics en vínculos y elementos.

Pero también de toda aquella contenida en las búsquedas que realizamos por ejemplo en Google, las publicaciones en las Redes sociales (Twitter, Facebook, etc.) y el contenido web como páginas, enlaces o imágenes.

En este caso ya que Actualmente a través de Internet puede accederse a consultar fuentes de información imprescindibles en cualquier investigación.

Entre las primeras fuentes de información disponibles en Internet se encuentran los catálogos de las grandes bibliotecas, a estas se han unido importantes bibliografías, directorios e instituciones.

Además de los registros bibliográficos, en las bibliotecas virtuales se pueden consultar y leer textos completos. Para facilitar a nuestros usuarios sus consultas en Internet se han seleccionado diversos recursos que se han agrupado por tipología y materias.

Lo que nos ha ayudado como fuentes de información principal para el desarrollo de nuestro sistema de ventas.



## **TÉCNICAS DE LIMPIEZA DE DATOS**

La limpieza de datos ha hecho que la dependencia de la información de datos sea manejable al mantener la calidad de los datos y mantener la integridad como una prioridad principal para las empresas. El proceso de limpieza de datos puede ser complejo si tiene diferentes conjuntos de datos provenientes de fuentes dispares. Tener una estrategia de limpieza de datos eficiente mantiene la integridad de los datos durante un proyecto de limpieza de datos.

Limpieza de datos, también conocida como depuración de datos o limpieza, es el primer paso en la preparación de datos. Implica identificar errores en un conjunto de datos y corregirlos para garantizar que solo se transfieran datos limpios y de alta calidad a los sistemas de destino lo que nos ayuda con el manejo más fácil y rápido de información.

### **DATA CLEANING**






#### **ETL**

Los procesos ETL son un término estándar que se utiliza para referirse al movimiento y transformación de datos. Se trata del proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos (denominada data mart o data warehouse) con el objeto de analizarlos. También pueden ser enviados a otro sistema operacional para apoyar un proceso de negocio.






Fases de un proceso ETL

Las distintas fases o secuencias de un proceso ETL son las siguientes:

- Extracción de los datos desde uno o varios sistemas fuente.

Clientes											
Reporte de clientes											
No	Nombre	Telefono	Correo	Edad	Dirección	Estado	Municipio	Código Postal	Forma de pago	Action	
6	Oscar	66666	sofi@gmail.com	66	av. principal	México	lerma	255555	efectivo	EDIT	
1	sandra	7224327417	sandra123@gmail.com	35	16 de sep #701	mexico	toluca	52106	efectivo	EDIT	
2	lilian	7224327455	54k3@gmail.com	22	16 de sep #701	mexico	toluca	52106	efectivo	EDIT	
3	sofia	7225648917	555gttthgh3@gmail.com	45	16 de sep #701	mexico	toluca	52106	efectivo	EDIT	
4	cassandra	7212365417	vcfvq2555@gmail.com	25	16 de sep #701	mexico	toluca	52106	efectivo	EDIT	
<div>EXCEL</div> <div>PDF</div>											

- Transformación de dichos datos, es decir, posibilidad de reformatear y limpiar estos datos cuando sea necesario.

Reportes									
employed deleted successfully									
Empleados									
Reporte de empleados									
No	Nombre	Edad	Telefono	Correo	Dirección	puesto	Action		
1	sandra	35	7224327417	sandra123@gmail.com	16 de sep #701	vendedor	EDIT		
2	lilian	22	7224327455	54k3@gmail.com	16 de sep #701	gerente	EDIT		
3	sofia	45	7225648917	555gttthgh3@gmail.com	16 de sep #701	administrador	EDIT		
4	cassandra	25	7212365417	vcfvq2555@gmail.com	16 de sep #701	administrador	EDIT		
5	estefania	50	7224378965	sajj3@gmail.com	16 de sep #701	vendedor	EDIT		

- Carga de dichos datos en otro lugar o base de datos, un data mart o un data warehouse, con el objeto de analizarlos o apoyar un proceso de negocio.

The screenshot shows a data management interface. On the left, a sidebar lists various database components like 'Tables', 'Columns', 'Indexes', etc. The main area displays a table with the following data:

id	nombre_empleado	edad_empleado	telefono_empleado	correo_empleado	direccion_empleado	puesto_empleado
1	sandra	35	7224327417	sandra123@gmail.com	16 de sep #701	vendedor
2	lillian	22	7224327455	541k3@gmail.com	16 de sep #701	gerente
3	sofia	45	7225648917	555gttFgh3@gmail.com	16 de sep #701	administrador
4	cassandra	25	7212365417	vofvg2555@gmail.com	16 de sep #701	administrador
5	estefania	50	7224378965	sa333@gmail.com	16 de sep #701	vendedor
(Auto)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

## La limpieza de datos como etapa separada de los procesos ETL

Aunque podría entenderse como una acción integrada en la fase de transformación de datos, en la actualidad la tendencia es considerar la limpieza de datos como una fase separada del proceso ETL.

Esta visión corresponde a una concepción más moderna y práctica del proceso. Para ahorrar tiempo y ganar en efectividad es conveniente unificar criterios, por ejemplo, introduciendo “av” en vez de “avenida” en todos los registros de una base de datos de direcciones postales, ANTES de empezar el proceso ETL propiamente dicho.

Tan importante es tener la información consolidada como que todos los datos sean correctos y con una visión única para todos los usuarios. Solo así se pueden lograr unos circuitos de trabajo y análisis de dichos datos realmente óptimos y efectivos.

### Pasos que se realizan para la limpieza de datos:

Paso 1: elimine las observaciones duplicadas o irrelevantes

Elimine las observaciones no deseadas de su conjunto de datos, incluidas las observaciones duplicadas o las observaciones irrelevantes. Las observaciones duplicadas ocurrirán con mayor frecuencia durante la recopilación de datos. Cuando combina conjuntos de datos de varios lugares, extrae datos o recibe datos de clientes o varios departamentos, existen oportunidades para crear datos

duplicados. La reduplicación es una de las áreas más importantes a considerar en este proceso. Las observaciones irrelevantes son cuando notas observaciones que no encajan en el problema específico que estás tratando de analizar. Por ejemplo, si desea analizar datos sobre clientes de la generación del milenio, pero su conjunto de datos incluye generaciones anteriores, puede eliminar esas observaciones irrelevantes.

## **Paso 2: corregir errores estructurales**

Los errores estructurales ocurren cuando mide o transfiere datos y observa convenciones de nomenclatura extrañas, errores tipográficos o mayúsculas incorrectas. Estas inconsistencias pueden causar categorías o clases mal etiquetadas. Por ejemplo, puede encontrar que aparecen "N/A" y "No aplicable", pero deben analizarse como la misma categoría.

## **Paso 3: filtre los valores atípicos no deseados**

A menudo, habrá observaciones únicas que, de un vistazo, no parecen encajar en los datos que está analizando. Si tiene una razón legítima para eliminar un valor atípico, como una entrada de datos incorrecta, hacerlo ayudará al rendimiento de los datos con los que está trabajando. Sin embargo, a veces es la aparición de un valor atípico lo que probará una teoría en la que está trabajando. Recuerde: el hecho de que exista un valor atípico no significa que sea incorrecto. Este paso es necesario para determinar la validez de ese número. Si un valor atípico resulta ser irrelevante para el análisis o es un error, considere eliminarlo.

## **Paso 4: Manejar los datos que faltan**

- No puede ignorar los datos faltantes porque muchos algoritmos no aceptarán valores faltantes. Hay un par de maneras de lidiar con los datos que faltan. Ninguno es óptimo, pero ambos pueden ser considerados.

- Como primera opción, puede eliminar las observaciones a las que les faltan valores, pero al hacerlo eliminará o perderá información, así que tenga esto en cuenta antes de eliminarlo.
- Como segunda opción, puede ingresar valores faltantes basados en otras observaciones; nuevamente, existe la posibilidad de perder la integridad de los datos porque puede estar operando a partir de suposiciones y no de observaciones reales.
- Como tercera opción, puede modificar la forma en que se usan los datos para navegar de manera efectiva por los valores nulos.

## **JUSTIFICACION DEL ALGORITMO (CLASIFICACION)**

### **DEFINICIÓN**

Los algoritmos de clasificación se usan cuando el resultado deseado es una etiqueta discreta, en otras palabras, son útiles cuando la respuesta al problema cae dentro de un conjunto finito de resultados posibles.

En el caso de que el modelo entrenado es para predecir cualquiera de las dos clases objetivos, verdadero o falso, por ejemplo, se le conoce como clasificación binaria. Algunos ejemplos de esto son: predecir si un alumno aprobará o no, predecir si un cliente comprará un producto nuevo o no.

El algoritmo elegido fue el de clasificación, ya que fue el que más se adapto a las necesidades del proyecto, el algoritmo de clasificación muestra patrones en los datos que le damos y los clasifica en grupo, porque dentro del sistema se encuentra diferentes tipos de usuario, como lo son (administrador, cliente, empleado) en donde se clasifican de acuerdo con el tipo que administrador y cada uno tiene acceso a diferentes partes del sistema como, por ejemplo:

- Vista usuario en donde solo el puede ver los productos y realizar compra.

MAJO

Inicio

Productos

Tenis Negros se ha agregado con éxito al carrito!

Tabla de productos

Listado de productos registrados

VER CARRITO 1

ID	Nombre	Tamaño	Descripción	Precio	Opciones	ADMIN Opciones
9	Tenis Negros	6	Tenis negros	250.00		
10	Mocasines cafe	24	Mocasines color café de piel	450.00		
11	Valerinas	23	valerinas beige	150.00		
12	ximena	24	Tenis rojos	-5.00		

EXCEL PDF

- Vista administrador en donde puede acceder a clientes, empleados, productos, compras y catálogo.

MAJO

Inicio

Cientes

Empleados

Productos

Compras

Catálogo

Paola

Paola

Inicio

Search...

Llamanos: Tel: 722-432-7417

Entregas a domicilio: Las 24 Horas









Followers +750

Unete a nosotros.

Followers +24

Actualizaciones diarias.

El mejor servicio de entrega

Tabla de productos						
Listado de productos registrados						
ID	Nombre	Tamaño	Descripción	Precio	Opciones	ADMIN Opciones
9	Tenis Negros	6	Tenis negros	250.00		<a href="#">EDIT</a> 
10	Mocasines cafe	24	Mocasines color café de piel	450.00		<a href="#">EDIT</a> 
11	Valerinas	23	valerinas beige	150.00		<a href="#">EDIT</a> 
12	ximena	24	Tenis rojos	-5.00		<a href="#">EDIT</a> 
<a href="#">EXCEL</a> <a href="#">PDF</a>						

## DESCRIPCION DEL MODELO DE DISEÑO DE CLASIFICACION

Para aplicar un método de clasificación, es la partición del conjunto de datos en dos conjuntos de datos más pequeños que serán utilizadas con los siguientes fines: entrenamiento y test. El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de prueba se emplea para comprobar el comportamiento del modelo estimado.

Cada registro de la base de datos debe de aparecer en uno de los dos subconjuntos, y para dividir el conjunto de datos en ambos subconjuntos, se utiliza un procedimiento de muestreo: muestreo aleatorio simple o muestreo estratificado.

Como resultado de aplicar un método de clasificación, se cometerán dos errores, en el caso de una variable binaria que toma valores 0 y 1, habrá ceros

que se clasifiquen incorrectamente como unos y unos que se clasifiquen incorrectamente como ceros. A partir de este recuento se puede construir el siguiente cuadro de clasificación:

<b>Valor real</b> $Y_i$ <b>Valor estimado</b> $\hat{Y}_i$	$Y_i=0$	$Y_i=1$
$Y_i=0$	P <sub>11</sub>	P <sub>12</sub>
$Y_i=1$	P <sub>21</sub>	P <sub>22</sub>

Donde P<sub>11</sub> y P<sub>22</sub> corresponderán a predicciones correctas (valores 0 bien predichos en el primer caso y valores 1 bien predichos en el segundo caso), mientras que P<sub>12</sub> y P<sub>21</sub> corresponderán a predicciones erróneas (valores 1 mal predichos en el primer caso y valores 0 mal predichos en el segundo caso)

## ALGORITMOS SUPERVISADOS

Los algoritmos que parten de un conjunto de datos etiquetados se denominan supervisados pues se supone que un "instructor" o supervisor está mostrando al aprendiz los datos de entrenamiento al mismo tiempo que le indica cuál es la respuesta correcta en cada caso. Si el nombre de "supervisado" está o no bien escogido es algo que invita a discusión...

En cualquier caso este tipo de algoritmos son los que más éxito tienen pues son bien conocidos y es muy sencillo evaluar su rendimiento (partimos de datos etiquetados, podemos dividirlos en los bloques de entrenamiento y prueba, entrenar nuestro modelo y ver hasta qué punto es capaz de predecir las etiquetas del conjunto de pruebas). Tal y como se ha explicado, estos algoritmos tienen



como objetivo encontrar la forma de relacionar las características que forman el conjunto de entrenamiento con sus etiquetas, y esta relación, una vez identificada, es la que se utilizará con nuevos datos para predecir sus etiquetas.

- La predicción del precio de venta de nuestros productos a partir de los datos relacionados. Para realizar el análisis deberemos partir de un conjunto de entrenamiento que incluya dichos datos y el precio real por el que se vendió
- En este caso el conjunto de entrenamiento está formado por las imágenes
- Los programas de correo electrónico incluyen clasificadores de spam que, basándose en datos conocidos de emails que han sido manualmente clasificados como spam, crean un modelo de Machine Learning para la identificación de nuevos mensajes de spam

El agrupamiento de K-means es el algoritmo de agrupamiento más utilizado. Es un algoritmo basado en centroides y es el algoritmo de aprendizaje no supervisado más simple. Este algoritmo intenta minimizar la varianza de los puntos de datos dentro de un grupo. También es la forma en que la mayoría de las personas se familiarizan con el aprendizaje automático sin supervisión. K-means se usa mejor en conjuntos de datos más pequeños porque itera sobre *todos* los puntos de datos. Eso significa que tomará más tiempo clasificar los puntos de datos si hay una gran cantidad de ellos en el conjunto de datos. Dado que así es como k-means agrupa los puntos de datos, no se escala bien.

## **IMPLEMENTACIÓN DEL ALGORITMO.**

Requiere que se indique de antemano el número de clusters que se van a crear. Las agrupaciones resultantes pueden variar dependiendo de la asignación aleatoria inicial de los centroides. Presenta problemas de robustez frente a outliers.

El clustering K Means es un buen lugar para empezar a explorar un conjunto de datos sin etiquetas. La K en K Means denota el número de

clúster. Este algoritmo está destinado a converger hacia una solución después de algunas iteraciones.

Tiene 4 pasos básicos:

- Inicializar los clústeres centroides, escoger los 3 libros para empezar.
- Asignar los puntos de datos a los clústeres, colocar los libros restantes uno por uno.
- Actualizar los centros de clúster, empezar de nuevo con 3 libros diferentes.
- Repetir el paso 2 y 3 hasta que se cumpla la condición de parada.

## **RESULTADOS DEL ALGORITMO.**

El algoritmo k-means mediante imágenes o procesamiento de datos determina cierta categoría obtenida dentro de las regiones de entrega. Implementando el algoritmo de agrupación k-means dentro se centra en el estándar de las regiones principales de entrega que abarca el proyecto observando esto con un módulo de entregas a domicilio. Eligiendo el algoritmo k-means si es necesario determinar una predicción de valores futuros entre los followers que se tendrán o se tienen en cierto tiempo o bien con imágenes hacer predicción de la marca de algún producto. Se utiliza el algoritmo k-means para hacer la predicción de los followers que tiene la tienda en sus diferentes redes sociales con ayuda de la implementación de la api de cada red social.

## **MÉTRICAS DE CLASIFICACIÓN**

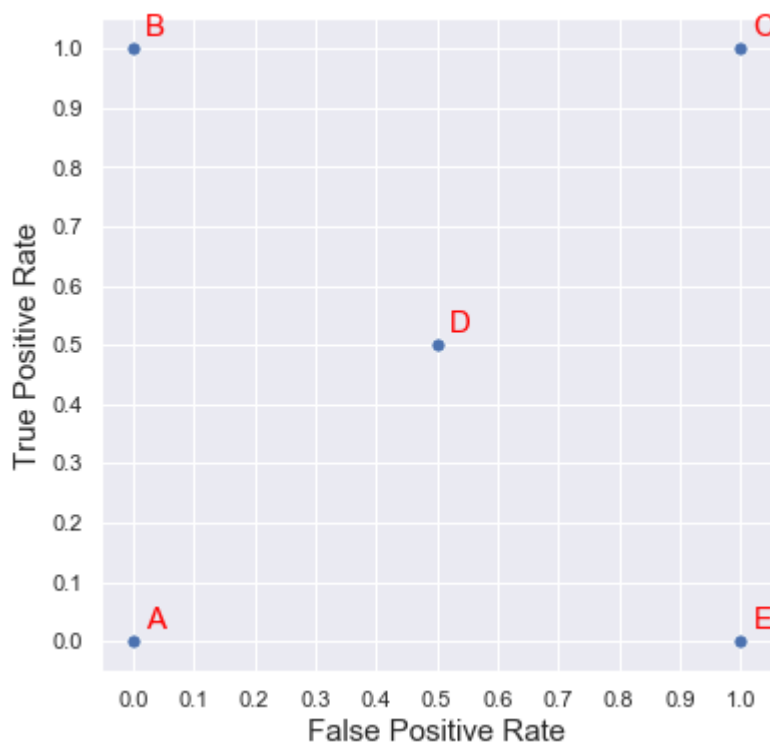
### **CURVA ROC**

Estas métricas son independientes: podemos tener un clasificador que detecte como positivos todos los que existan, pero que no sea muy exacto pues esté detectando como positivos muchas muestras que no lo son. O al revés: el

clasificador puede estar detectando como positivas pocas muestras, pero lo puede estar haciendo de forma muy exacta, marcando como positivas solo las que realmente lo son.

Un clasificador ideal detectaría todas las muestras positivas y solo esas. Es decir, si hablamos de las métricas de precisión y exhaustividad, el clasificador ideal tendría una precisión de 1 y una exhaustividad (recall o TPR) de 1. Si, por el contrario, estamos refiriéndonos a la métrica de FPR, el clasificador ideal tendría un FPR de 0 (no habría falsos negativos).

El clasificador con el siguiente esquema, en el que hemos marcado los valores de TPR y FPR para cinco clasificadores distintos (A, B, C, D y E):



Ya hemos visto que el FPR toma valores en el rango 0 (para el clasificador ideal) y 1 (para el peor clasificador posible), y que el TPR toma valores en el rango 0 (para el peor clasificador posible) y 1 (para el clasificador ideal). Esto significa que:

- El clasificador **A** tiene el mejor FPR posible (no detecta como positivos ninguna muestra negativa) pero tiene el peor TPR (ninguna muestra positiva la identifica como tal).
- El clasificador **B** tiene el mejor FPR y el mejor TPR. Se trata del clasificador ideal.
- El clasificador **C** tiene el peor FPR (todas las muestras negativas son identificadas como positivas) pero su TPR es el mejor posible (todas las muestras positivas son identificadas como tal).
- El clasificador **D** tiene un FPR de 0.5 (la mitad de las muestras negativas las identifica como tal. La otra mitad las identifica erróneamente como positivas), y un TPR también de 0.5 (la mitad de las muestras positivas las identifica como tal, e identifica la otra mitad erróneamente como negativas).
- Por último, el clasificador **E** tiene el peor FPR posible (todas las muestras negativas son identificadas como positivas) y el peor TPR posible (ninguna muestra positiva es identificada como tal). Es el peor clasificador posible (todas las predicciones son erróneas).

## EXHAUSTIVIDAD

La exhaustividad de una clasificación es la ratio entre los verdaderos positivos y la suma de los verdaderos positivos y los falsos negativos, o dicho con otras palabras: el ratio entre los verdaderos positivos (los que se han detectado) y los positivos reales (los hayamos detectado o no).

Es decir, si nuevamente estamos detectando transacciones fraudulentas, la exhaustividad sería el número de transacciones fraudulentas detectadas dividido por el número de transacciones fraudulentas reales totales: si hemos detectado 3 y había 5, la exhaustividad ha sido de  $3 / 5 = 0.6$ . El clasificador ideal tendría una exhaustividad de 1, pues todos los positivos reales serían detectados como positivos (verdaderos positivos), y el peor clasificador posible tendría una exhaustividad de 0, pues ninguno de los positivos reales sería identificado como positivo.

Lógicamente, el clasificador ideal tendría un FPR de cero (pues no tendría falsos positivos), y el peor clasificador posible tendría un FPR de uno (todos los negativos reales serían identificados erróneamente como positivos):

```
y_real = [1, 1, 1, 1, 1, 1, 1, 0, 0, 0]
```

```
y_pred = [1, 1, 1, 1, 0, 0, 0, 0, 1, 1]
```

Tenemos 10 elementos de los que 7 son positivos y 3 son negativos. Hemos marcado en nuestra predicción como positivos 6 elementos, de los que 4 son verdaderos positivos y 2 son falsos positivos. De forma semejante, hemos marcado 4 negativos, de los que 1 es un verdadero negativo y los otros 3 son falsos negativos.

La matriz de confusión devuelta por sklearn es:

```
confusion_matrix(y_real, y_pred)
```

```
array([[1, 2],
```

```
       [3, 4]], dtype=int64)
```

El FPR se calcula, , como el ratio entre verdaderos negativos y el número de negativos:  $2 / 3$ , y representa la proporción de negativos que no estamos identificando. Mostrando este cálculo en la matriz de confusión.

		Positive	Negative
Prediction	Positive	True positives 4	False positives 2
	Negative	False negatives 3	True negatives 1

Es decir, 2/3.

## CONCLUSIONES

- La implementación de un sistema ha ayudado mucho al negocio ya que se les mostró una vista previa de lo que sería manejar su negocio de manera digital.
- Algunos de los pequeños y medianos negocios que se vieron afectados por la pandemia recurrieron a la implementación de comercio electrónico, el cual fue una gran herramienta, sin embargo, ahora que debemos regresar a

la normalidad y adaptarnos a la vida postpandemia muchos de ellos recurrieron a mantener el manejo de sus negocios de esta forma ya que agiliza procesos y les brinda una oportunidad de crecimiento el mercado.

- El e-commerce brinda a los negocios una gran oportunidad de crecimiento y posicionamiento dentro del mercado ya que permite la difusión de información en línea.