

# Análisis de retención de clientes en el sector financiero mediante minería de datos

Brayan David Reyes Morales<sup>1</sup>, Nicolás Alejandro Fernández Espinosa<sup>2</sup>

Universidad Central  
Maestría en Analítica de Datos  
Curso de Bases de Datos  
Bogotá, Colombia  
breyesm@ucentral.edu.co, nfernandez@ucentral.edu.co

May 24, 2023

## Contents

<b>1</b>	<b>Introducción</b>	<b>3</b>
<b>2</b>	<b>Características del proyecto de investigación que hace uso de Bases de Datos</b>	<b>4</b>
2.1	Titulo del proyecto de investigación . . . . .	5
2.2	Objetivo general . . . . .	5
2.2.1	Objetivos específicos . . . . .	5
2.3	Alcance . . . . .	5
2.4	Pregunta de investigación . . . . .	6
2.5	Hipótesis . . . . .	6
<b>3</b>	<b>Reflexiones sobre el origen de datos e información</b>	<b>7</b>
3.1	¿Cual es el origen de los datos e información? . . . . .	7
3.2	¿Cuales son las consideraciones legales o éticas del uso de la información? . . . . .	7
3.3	¿Cuales son los retos de la información y los datos que utilizara en la base de datos en términos de la calidad y la consolidación? . . . . .	8
3.4	¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? (Max 100 Palabras) - ( <i>Primera entrega</i> ) . . . . .	10
<b>4</b>	<b>Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos)(<i>Primera entrega</i>)</b>	<b>12</b>
4.1	Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto ( <i>Primera entrega</i> ) . . . . .	12

4.2	Diagrama modelo de datos ( <i>Primera entrega</i> ) . . . . .	13
4.3	Imágenes de la Base de Datos ( <i>Primera entrega</i> ) . . . . .	14
4.4	Código SQL - lenguaje de definición de datos (DDL) ( <i>Primera entrega</i> ) . . . . .	15
4.5	Código SQL - Manipulación de datos (DML) ( <i>Primera entrega</i> ) . .	17
4.6	Código SQL + Resultados: Vistas ( <i>Primera entrega</i> ) . . . . .	19
4.7	Código SQL + Resultados: Triggers ( <i>Primera entrega</i> ) . . . . .	20
4.8	Código SQL + Resultados: Funciones ( <i>Primera entrega</i> ) . . . . .	21
4.9	Código SQL + Resultados: procedimientos almacenados ( <i>Primera entrega</i> ) . . . . .	21
<b>5</b>	<b>Bases de Datos No-SQL (<i>Segunda entrega</i>)</b>	<b>22</b>
5.1	Diagrama Bases de Datos No-SQL ( <i>Segunda entrega</i> ) . . . . .	22
5.2	SMBD utilizado para la Base de Datos No-SQL ( <i>Segunda entrega</i> )	22
<b>6</b>	<b>Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos</b>	<b>23</b>
6.1	Ejemplo de aplicación de ETL y Bodega de Datos . . . . .	24
<b>7</b>	<b>Proximos pasos</b>	<b>25</b>
<b>8</b>	<b>Lecciones aprendidas</b>	<b>25</b>
<b>9</b>	<b>Bibliografía</b>	<b>26</b>

# 1 Introducción

La motivación principal en la creación y mantenimiento de las empresas, sin importar el sector al que pertenezcan, ha sido el relacionamiento con los clientes, ya que ellos son los generadores de ingresos para las compañías. Aunque la ecuación de ingresos a partir de clientes puede ser simple y lógica, a menudo se olvida, ya que se prioriza la obtención de nuevos clientes en lugar de mantener a los existentes. No se le da suficiente importancia a los niveles de deserción en las compañías.

En la actualidad, debido a la globalización de servicios y la competencia, los productos en demanda son cada vez más similares en términos de calidad y precio [1], [2]. Ante esta situación, las compañías se vuelven dinámicas en función de los clientes, por lo cual es indispensable conocer a los clientes con el fin de explicar y anticiparse a sus acciones en el futuro.

A pesar de contar con una amplia oferta de productos y servicios, los clientes tienen la facilidad de cambiar de una compañía a otra. Por esta razón, las entidades priorizan la atracción de nuevos clientes, pero esta no es la mejor manera de solucionar la pérdida de clientes. En su lugar, se busca un relacionamiento orientado a fidelizar a los clientes de manera correcta y responsable, adaptado a los productos y servicios ofrecidos por la industria. Aquellos clientes que dejan de usar los productos de una empresa se denominan desertores, y al identificarlos se pueden aplicar estrategias de retención [3]. Por tanto, es indispensable conocer el entorno de las compañías y generar propuestas de retención efectivas que se traduzcan en una menor tasa de deserción y una mayor base de clientes.

Dado que la obtención de nuevos clientes ha cambiado dinámicamente, en la actualidad se busca retener de manera efectiva a los clientes. Se ha demostrado que las estrategias de retención son más rentables en comparación con los esfuerzos para atraer nuevos clientes. Algunas investigaciones muestran que el costo de obtener un nuevo cliente es de 5 a 7 veces mayor que el costo de retener a uno existente [4]. Además, el "efecto de lealtad" ha demostrado que un aumento del 5 por ciento en la tasa de retención de clientes puede generar incrementos del 35 por ciento y 95 por ciento en el valor actual neto de los clientes en una empresa desarrolladora de software y una agencia de publicidad, respectivamente [5].

Por los motivos mencionados anteriormente, se propone un modelo estadístico para identificar los comportamientos de los clientes y lograr un modelo de retención que pueda predecir el nivel de deserción. Este modelo estará enfocado en los usuarios del sistema financiero y se apoyará en modelos estudiados e implementados en diferentes trabajos e investigaciones similares.

Para abordar el problema, es necesario definir los criterios de evaluación, que

corresponden a las variables del dataset. Entre estos criterios, el más importante es la identificación de un cliente como desertor dentro de la compañía. El concepto de clientes desertores varía según los comportamientos de los clientes en diferentes industrias y tipos de mercado a los que están expuestos actualmente los consumidores. En el caso del sector bancario, se considera como cliente desertor a aquel que cierra todas sus cuentas bancarias y cesa los negocios con el banco en estudio [6].

En este estudio, se utilizarán los datos disponibles y se hará uso de la minería de datos, ya que esta técnica estadística es considerada una herramienta estratégica importante para las empresas. Permite analizar grandes volúmenes de información desde diferentes perspectivas y sintetizarla en información valiosa, lo cual facilita la toma de decisiones temprana y efectiva en las organizaciones empresariales. La minería de datos también es útil para crear un perfil preciso de los clientes basado en su comportamiento [7].

La minería de datos proporciona información comprimida que ayuda a comprender lo que ha ocurrido o está ocurriendo. El modelamiento de los datos es importante en las compañías, ya que genera información concisa y precisa que se puede utilizar para aplicar modelos estadísticos. En este caso, los modelos estadísticos utilizados serán algoritmos de programación como:

Árboles de Decisión, Redes Neuronales, Reglas de Asociación, Regresión Logística, Árboles Aleatorios y SVM.

## **2 Características del proyecto de investigación que hace uso de Bases de Datos**

El estudio se lleva a cabo en una entidad del sector financiero en Colombia. Aunque los modelos son aplicables a diferentes industrias y sectores, cada empresa tiene características únicas tanto en sí misma como en relación a sus clientes. En el caso de estudio, que se basa en datos de una institución financiera, es necesario definir de dónde se obtendrán los datos y el período correspondiente. Como se muestra en la tabla 1, existen dos opciones: se pueden seleccionar una cantidad específica de clientes retirados o tomar a todos los clientes de los últimos meses, no superando los 6 meses, y luego identificar los retirados. En este caso de estudio, se tomarán los clientes de tres meses específicos, considerando la variable categórica "retirados", la calidad de los datos y los tipos de datos que se generen a partir de la base proporcionada.

Utilizando las fuentes disponibles en la organización, se procederá con el modelo CRISP-DM con el objetivo de responder a la pregunta de negocio: "¿Qué tipo de clientes tienen una alta probabilidad de retirarse en el futuro?". Para ello, se ha creado una base de datos con diferentes variables que ayudarán a explicar esta cuestión.

## 2.1 Título del proyecto de investigación

Análisis de retención de clientes en el sector financiero mediante minería de datos.

## 2.2 Objetivo general

Identificar comportamientos de los clientes en el sistema financiero y lograr un modelo de retención exitoso.

### 2.2.1 Objetivos específicos

- Definir los criterios de evaluación y variables del dataset, en especial la definición de un cliente desertor en el sector bancario.
- Realizar un análisis exploratorio de los datos para identificar patrones y tendencias en el comportamiento de los clientes.
- Estimar un modelo estadístico que permita predecir el nivel de deserción de clientes en una entidad financiera.
- Evaluar la efectividad del modelo propuesto y comparar el costo de obtener un nuevo cliente versus el costo de retener a uno.

## 2.3 Alcance

Se busca realizar una extracción y manipulación precisa de los datos con el fin de crear una base de datos sólida que permita obtener información financiera y sociodemográfica de los clientes retirados en una entidad financiera. Para ello, se utilizarán sistemas informáticos como SQL, ya que se dispone de datos estructurados.

El objetivo es obtener un modelo que explique y comprenda el comportamiento de un cliente al retirarse de una compañía. Dado que no existen estudios anteriores específicos para el negocio en cuestión, se parte de un conjunto de datos relativamente robusto, aunque con variables que posiblemente contribuyan a entender al cliente. Es importante destacar que este modelo puede fortalecerse mediante la inclusión y eliminación de variables. Se toman en cuenta variables explicativas relacionadas con el negocio, como saldos de productos y las interacciones del cliente con el entorno empresarial. Sin embargo, se podrían agregar variables relacionadas con las experiencias del cliente con los productos (actualmente no mapeadas), lo que contribuiría a generar una cultura de medición de sentimientos y generación de datos a partir de cualquier comportamiento del cliente.

## **2.4 Pregunta de investigación**

¿Se puede predecir y retener a clientes desertores en una entidad financiera en Colombia mediante el relacionamiento de bases dedatos?

## **2.5 Hipótesis**

La minería de datos funciona para identificar patrones de comportamiento en los clientes de una entidad financiera, que pueden desarrollar estrategias de retención efectivas para reducir la tasa de deserción y aumentar el valor actual neto de los clientes.

### **3 Reflexiones sobre el origen de datos e información**

Se espera que una entidad tenga sus datos centralizados en un único lugar y con poca información faltante. Sin embargo, en muchas ocasiones esto no ocurre en las compañías, ya sea en tiempo real o con datos atrasados. En el caso particular de este ejercicio, la recopilación y acceso a la información puede estar limitada por otras áreas debido a precauciones en el manejo de transacciones y acceso a los archivos.

Una vez que se supera el problema de los permisos, surge otro obstáculo al identificar las variables relevantes para generar una base de datos, ya que muchas de ellas contienen información similar o repetida en diferentes tablas. Esto causa molestias y requiere un esfuerzo adicional en la minería de datos, especialmente si los datos se distribuyen en múltiples tablas. Una vez que la información está organizada, se puede crear una tabla con cinco variables, cada una proveniente de una tabla diferente. Sin embargo, esto puede representar un gran desafío en las entidades, ya que se descuida el modelamiento y se requiere un esfuerzo significativo para crear una base comprensible y realizar análisis descriptivos, con la esperanza de que las variables proporcionen información sólida.

#### **3.1 ¿Cual es el origen de los datos e información?**

La información proviene de transacciones y creaciones de clientes que se guardan en repositorios diferentes. En general los repositorios son archivos en Excel e información recopilada a través del programa CRM.

#### **3.2 ¿Cuales son las consideraciones legales o éticas del uso de la información?**

Para el tratamiento de datos que se aplica en el presente trabajo, se hace una solicitud formal al área de gobierno de datos, solicitando el uso de las variables expuestas. Para lograr la autorización correcta se realizó el anonimato de las identificaciones de los clientes.

En cuanto al riesgo que se puede presentar sin la previa autorización se detalla gracias a la ley estatutaria 1581 del 2012, en el cual se presentan las siguientes restricciones por parte de las personas y las empresas que permiten la utilización de los datos "sensibles" que suministra una persona a una entidad.

Artículo 9°. Autorización del Titular. Sin perjuicio de las excepciones previstas en la ley, en el Tratamiento se requiere la autorización previa e informada del Titular, la cual deberá ser obtenida por cualquier medio que pueda ser objeto de consulta posterior.

Artículo 10. Casos en que no es necesaria la autorización. La autorización del Titular no será necesaria cuando se trate de:

a) Información requerida por una entidad pública o administrativa en ejercicio de sus funciones legales o por orden judicial;

- b) Datos de naturaleza pública;
  - c) Casos de urgencia médica o sanitaria;
  - d) Tratamiento de información autorizado por la ley para fines históricos, estadísticos o científicos;
  - e) Datos relacionados con el Registro Civil de las Personas.
- Quien acceda a los datos personales sin que medie autorización previa deberá en todo caso cumplir con las disposiciones contenidas en la presente ley.[19]

### 3.3 ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en términos de la calidad y la consolidación?

Tomando las variables identificadas en el diccionario se procede a identificar la calidad de estos, buscando que las variables cumplan con los estándares mínimos para el montaje de un modelo estadístico, se obtiene lo siguiente:

	Nombre Columna	Tipo	Cantidad de Únicos	Cantidad de Nulos	Cantidad Valores No Nulos
	OFICINA	object	37	0	46144
	TIPO	object	2	0	46144
	SEGMENTO	object	4	0	46144
	RETIRARON	object	2	0	46144
	NICHO	object	12	0	46144
	CANAL	object	6	0	46144
	OCUPACION	object	6	0	46144
	SALDO_CAPTA	float64	25024	0	46144
	SALDO_COLOCA	float64	23294	0	46144
	NUM_PRODUCTOS	int64	28	0	46144
	PAGADURIA	object	291	20025	26119
	ESTADO_CIVIL	object	6	5166	40978
	ESTRATO	int64	6	0	46144
	EDAD	float64	83	1273	44871
	GENERO	object	2	811	45333
	VINCULO	object	3	0	46144
	INGRESOS	object	16005	665	45479
	PRODUCTO	object	16	0	46144

Con la ayuda de Google Colab se visualiza el tipo de dato, el número de valores únicos y la cantidad de valores nulos. Se observa que el dataset tiene un tamaño de 46.144 x 18. En donde se tienen 6 variables cuantitativas y 12 variables cualitativas, de las cuales las variables de “PAGADURIA, ESTADO CIVIL, EDAD, GENERO e INGRESOS” presentan valores faltantes, es por ello por lo que la variable de “INGRESOS” no se toma ni se puede convertir a tipo numérica ya que presenta valores faltantes y esto genera un error en la identificación del tipo de dato.

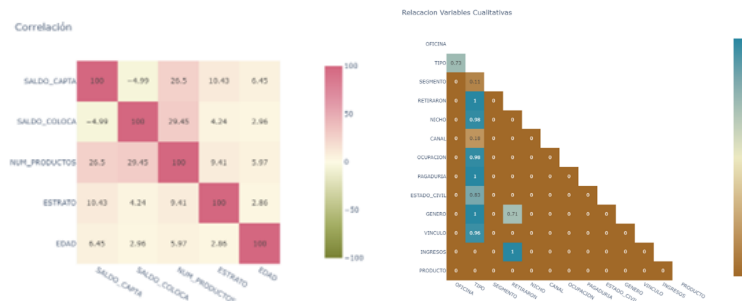


	SALDO_CAPTA	SALDO_COLOCA	NUM_PRODUCTOS	ESTRATO	EDAD
<b>count</b>	4.614400e+04	4.614400e+04	46144.00	46144.00	44871.00
<b>mean</b>	1.273259e+07	1.744456e+07	1.79	2.79	54.31
<b>std</b>	7.850941e+07	3.752638e+07	1.30	1.04	16.12
<b>min</b>	-1.214730e+05	0.000000e+00	1.00	1.00	13.00
<b>25%</b>	0.000000e+00	0.000000e+00	1.00	2.00	41.00
<b>50%</b>	1.005045e+05	2.085620e+06	1.00	3.00	56.00
<b>75%</b>	2.317506e+06	1.996970e+07	2.00	3.00	66.00
<b>max</b>	4.712371e+09	1.500000e+09	36.00	6.00	95.00

Generando un análisis de las variables numéricas se puede deducir que las variables de saldos necesitan una normalización, modificándolas con logaritmos o exponenciales, ya que sus valores al ser altos generan un poco compresión de los mismos, por otro lado los clientes de la entidad estudiada presentan una edad media de 54 años, un estrato 3 y productos adquiridos de 2. Hay que tener en cuenta que cuando se habla de saldos se debe tener presente que algunos pueden ser muy altos y otros bajos, dado el siguiente se puede estudiar la posibilidad de recortar los datos con medias recortadas, solo para dichas variables.



Con la generación de los gráficos de frecuencia se puede deducir que la base necesita ser balanceada y buscar que variables se les debe realizar una imputación de datos y otras solo perder la información, se debe tener cuidado ya que si se asume la pérdida de datos de una variable se puede estar eliminando información de los clientes retirados y así quitando la oportunidad de que los modelos no puedan aprender sobre los clientes retirados.



Se generan gráficos de correlación para las variables en donde las variables cualitativas en su mayoría se explican las unas a las otras, la variable TIPO, la cual contiene el tipo de cliente si es activo en sus saldos es la que no tiene una explicación frente a las demás, esta variable puede ser retirada ya que se tenía únicamente para generar la base con la que se genera el modelamiento, esto nos da una buena señal para plantear un modelo estadístico, en cuanto a las variables categóricas.

Por otro lado, las variables numéricas no presentan correlación es fuertes entre ellas, pero no se puede deducir que se tengan que dejar a un lado, ya que no se han realizado estudios anteriores en la compañía, al tener estos valores nos permite tener un punto inicial de como se comportan las variables dentro de la compañía y así generar o mejorar los estudios posteriores.

3.4 ¿Que espera de la utilización de un sistema de Bases de Datos para su proyecto? (Max 100 Palabras) - (*Primera entrega*)

Con el presente ejercicio se espera obtener un modelo que permita explicar y entender un cliente que se retira de una compañía. Como no se han hecho estudios anteriores en el negocio específico se inicia con un dataset medianamente robusto, pero con variables que posiblemente ayudaran a entender al cliente. Claramente este modelo se puede robustecer con la integración y salida de variables, es importante recalcar que se están tomando variables explicativas de negocio como saldos productos y conexiones del cliente con el entorno del negocio, pero se pueden agregar las variables de experiencias del cliente con los productos (actualmente no mapeadas), ayudaría a generar la cultura de medición de los sentimientos y generación de datos a partir de cualquier comportamiento del cliente.

Con la serie de modelos mencionados en el estudio se busca dar un valor agregado a una gestión y no tan solo un resultado. Es importante el resultado por que a partir de ello se tomaran decisiones, pero el resultado puede ser cambiante con cada modelo aplicado, el valor adicional se agrega cuando en la practica sin importar el resultado se pueden aplicar varias fórmulas “comerciales”

apoyadas de la estadística que permitan aumentar la fidelidad de un cliente generando una confianza de marca, teniendo como base principal el análisis de datos.

## 4 Diseño del Modelo de Datos del SMBD (Sistema Manejador de Bases de Datos) *(Primera entrega)*

### 4.1 Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto *(Primera entrega)*

Para el trabajo actual se utilizará MySQL el cual permite modificar y generar tablas que permiten realizar la conexión a través de llaves, para el caso se tiene una llave única que será el número de identificación del usuario.

MySQL Workbench es un sistema libre que permite trabajar con una cantidad limitada de datos, pero para el proyecto actual es suficiente ya que posee las siguientes características:

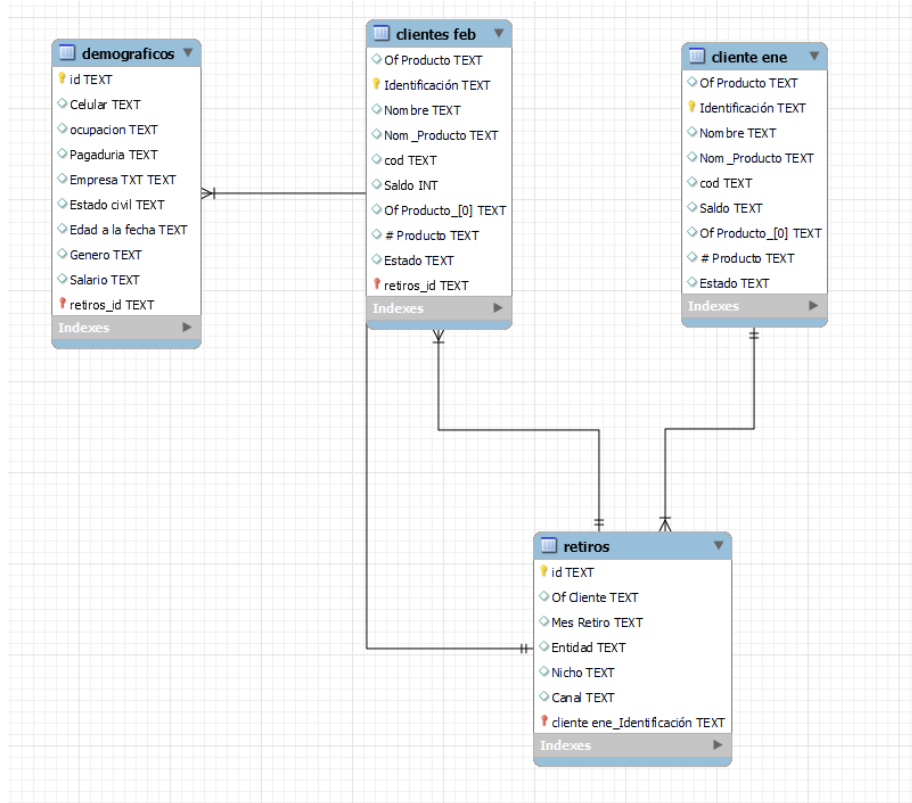
- Disponibilidad en gran cantidad de plataformas y sistemas operativos.
- Soporta gran cantidad de datos y posee diferentes opciones para su almacenamiento.
- Transacciones y claves primarias y foráneas.
- Conectividad segura.

Como segunda opción se cuenta con el programa de Oracle SQL para suplir la contingencia de datos pesados (registros mayores de 200.000) para la generación de las tablas de la base de datos.

- Oracle SQL es un sistema de gestión de bases de datos relacionales (RDBMS) que proporciona una plataforma sólida y escalable para el almacenamiento, gestión y procesamiento de grandes cantidades de datos. Algunas de las características y ventajas de Oracle SQL incluyen:

- Escalabilidad: Oracle SQL puede manejar grandes cantidades de datos y es altamente escalable, lo que significa que puede manejar el aumento de datos y usuarios sin comprometer el rendimiento.
- Fiabilidad: Oracle SQL es conocido por su alta disponibilidad y confiabilidad, lo que lo hace ideal para aplicaciones críticas de misión crítica.
- Seguridad: Oracle SQL ofrece un alto nivel de seguridad para proteger los datos almacenados, incluyendo funciones de autenticación, autorización y cifrado de datos.
- Rendimiento: Oracle SQL está optimizado para ofrecer un rendimiento rápido y eficiente, lo que lo hace ideal para aplicaciones que requieren un acceso rápido a grandes cantidades de datos.
- Funcionalidad avanzada: Oracle SQL ofrece una amplia gama de funcionalidades avanzadas, como particionamiento de tablas, indexación avanzada, gestión de transacciones y más, lo que lo hace ideal para aplicaciones empresariales complejas.
- Soporte de la comunidad: Oracle SQL tiene una gran comunidad de usuarios y desarrolladores que ofrecen soporte y recursos para ayudar a los usuarios a resolver problemas y aprender más sobre el sistema.

## 4.2 Diagrama modelo de datos (*Primera entrega*)



### 4.3 Imágenes de la Base de Datos (*Primera entrega*)

Query 1 x DML DDL

Limit to 1000 rows

```

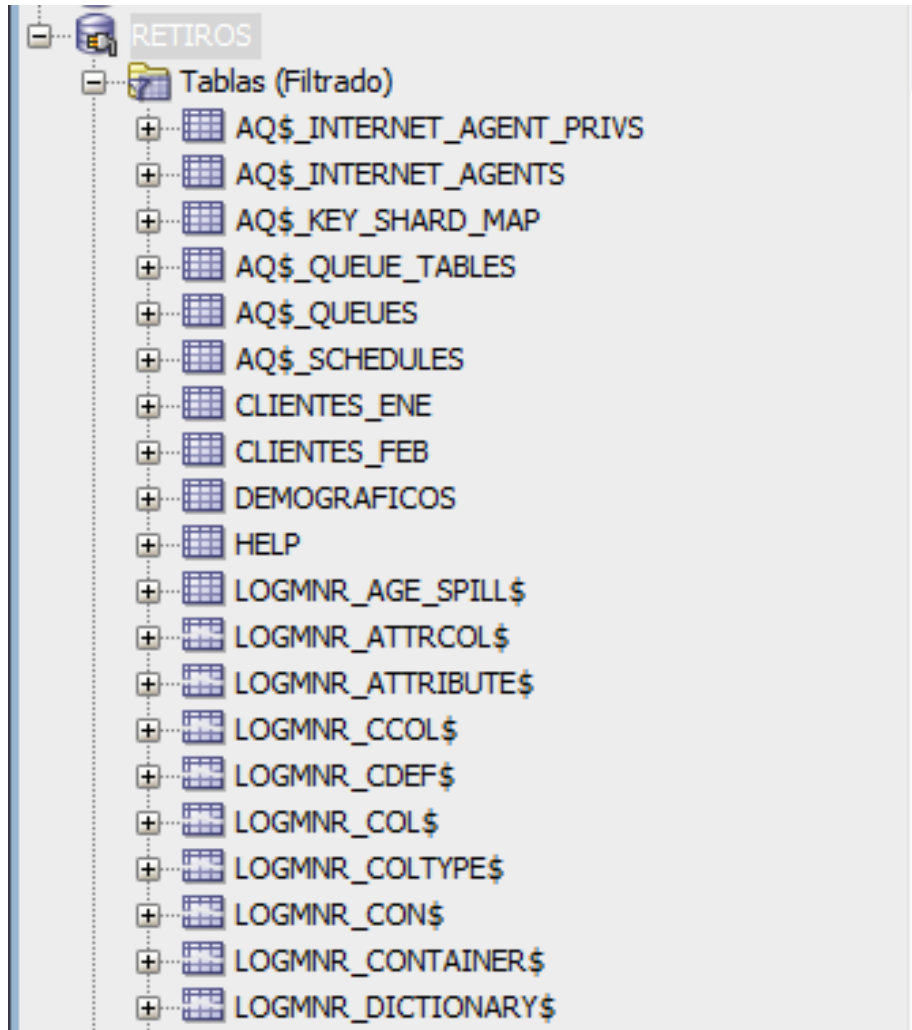
1 • select * from demograficos;
2 • select * from retiros;

```

Result Grid Filter Rows: Export: Wrap Cell Content: Fetch rows:

	id	Of Cliente	Mes Retiro	Entidad	Nicho	Canal
▶	196600	Bogota - Galerías	nov-22	Colpensiones	Pensionados	E-Credit
	588574	Duitama	nov-22	Empleado Empresa Privada	Empleado Empresa Privada	Oficinas
	1098375	Duitama	nov-22	Colpensiones	Pensionados	Oficinas
	1279325	Pereira	nov-22	Colpensiones	Pensionados	E-Credit
	2295624	Ibague	nov-22	Colpensiones	Pensionados	E-Credit
	2909718	San Gil	nov-22	Fopep	Pensionados	Oficinas
	3069502	Bogota - Galerías	nov-22	Rama Judicial	Sistema Nacional de Justicia	Oficinas
	3171460	Bogota - Centro	nov-22	Militares y Policías	Pensionados	Oficinas
	4313498	Armenia	nov-22	Otros Segmentos **	Otros Segmentos **	Oficinas
	4322090	Manizales	nov-22	Colpensiones	Pensionados	E-Credit
	4327460	Manizales	nov-22	Colpensiones	Pensionados	E-Credit
	4378988	Pereira	nov-22	Otros Pensionados	Pensionados	Oficinas
	4508852	B'bermeja	nov-22	Colpensiones	Pensionados	E-Credit
	4627172	Popayan	nov-22	Independientes	Independientes	Oficinas
	5149854	Riohacha	nov-22	Empleado Empresa Privada	Empleado Empresa Privada	Oficinas
	5202301	Bogota - Centro	nov-22	Militares y Policías	Pensionados	E-Credit
	5588376	B'lmanga	nov-22	Fiduprevisora (Magisterio)	Pensionados	E-Credit
	5765303	Socorro	nov-22	Independientes	Independientes	Oficinas
	6092561	Calli	nov-22	Colpensiones	Pensionados	E-Credit
	6243747	Pitalito	nov-22	Fiduprevisora (Magisterio)	Pensionados	Oficinas
	6759758	Tunja	nov-22	Independientes	Independientes	Oficinas
	6771669	Tunja	nov-22	Rama Judicial	Sistema Nacional de Justicia	Oficinas
	7162534	Tunja	nov-22	Empleado Empresa Privada	Empleado Empresa Privada	Oficinas
	7438550	Tunja	nov-22	Empleado Empresa Privada	Empleado Empresa Privada	Oficinas

#### 4.4 Código SQL - lenguaje de definición de datos (DDL) (Primera entrega)



```

CREATE TABLE Retiros (
    ID_Cliente NUMBER(10) PRIMARY KEY,
    Of_Cliente VARCHAR2(50),
    Mes_Retiro VARCHAR2(10),
    Entidad_Nicho VARCHAR2(50),
    Canal VARCHAR2(50)
);

alter table Retiros
add Entidad VARCHAR2(50);

SELECT * FROM Retiros

CREATE TABLE Demograficos (
    ID_Cliente NUMBER(10) PRIMARY KEY,
    Celular VARCHAR2(50),
    Ocupacion VARCHAR2(50),
    Pagaduria VARCHAR2(70),
    Empresa VARCHAR2(50),
    Estado_civil varchar2(50),
    Edad number(10),
    Genero varchar2(50),
    Salario number(20)
);

ALTER TABLE Demograficos MODIFY Pagaduria varchar2(100);
ALTER TABLE Demograficos MODIFY Empresa varchar2(100);
ALTER TABLE Demograficos MODIFY Ocupacion varchar2(50);

```



```

CREATE TABLE Clientes_ENE (
    ID_Cliente NUMBER(10) PRIMARY KEY,
    Of_producto VARCHAR2(50),
    Producto VARCHAR2(50),
    COD VARCHAR2(70),
    Saldo number(20),
    Num_producto number(20),
    Estado varchar2(10)
);

CREATE TABLE Clientes_FEB (
    ID_Cliente NUMBER(10) PRIMARY KEY,
    Of_producto VARCHAR2(50),
    Producto VARCHAR2(50),
    COD VARCHAR2(70),
    Saldo number(20),
    Num_producto number(20),
    Estado varchar2(20)
);

```

#### 4.5 Código SQL - Manipulación de datos (DML) *(Primera entrega)*

```

INSERT INTO CLIENTES_ENE (OF_PRODUCTO, ID_CLIENTE, PRODUCTO, COD, SALDO, NUM_PRODUCTO, ESTADO) VALUES ('30053',123334,'A Término','ED',20.000.000 ,9813,'Activo');
--Fila 2
INSERT INTO CLIENTES_ENE (OF_PRODUCTO, ID_CLIENTE, PRODUCTO, COD, SALDO, NUM_PRODUCTO, ESTADO) VALUES ('30053',4456454,'A Término','ED',44.388.502 ,9813,'Activo');
--Fila 3
INSERT INTO CLIENTES_ENE (OF_PRODUCTO, ID_CLIENTE, PRODUCTO, COD, SALDO, NUM_PRODUCTO, ESTADO) VALUES ('30053',77879899,'A Término','ED',18.800.000 ,10243,'Activo');
--Fila 4
INSERT INTO CLIENTES_ENE (OF_PRODUCTO, ID_CLIENTE, PRODUCTO, COD, SALDO, NUM_PRODUCTO, ESTADO) VALUES ('30053',1909090768541,'A Término','ED',5.000.000 ,10243,'Activo');
--Fila 5
INSERT INTO CLIENTES_ENE (OF_PRODUCTO, ID_CLIENTE, PRODUCTO, COD, SALDO, NUM_PRODUCTO, ESTADO) VALUES ('30053',52343450301,'A Término','ED',6.000.000 ,10243,'Activo');
--Fila 6
INSERT INTO CLIENTES_ENE (OF_PRODUCTO, ID_CLIENTE, PRODUCTO, COD, SALDO, NUM_PRODUCTO, ESTADO) VALUES ('30053',9045354317045,'A Término','ED',5.000.000 ,10243,'Activo');

```

Tabla Normalizada:

```

CREATE TABLE Norma (
    ID_Cliente NUMBER(10) PRIMARY KEY,
    Of_producto VARCHAR2(50),
    Producto VARCHAR2(50),
    COD VARCHAR2(70),
    Saldo number(20),
    Num_producto number(20),
    Estado varchar2(20),
    Mes varchar2(10)
);

select * from Norma;

INSERT INTO Norma (ID_Cliente, Of_producto, Producto, COD, Saldo, Num_producto, Estado, Mes)
SELECT ID_Cliente, Of_producto, Producto, COD, Saldo, Num_producto, Estado, 'Enero'
FROM Clientes_ENE;

INSERT INTO Norma (ID_Cliente, Of_producto, Producto, COD, Saldo, Num_producto, Estado, Mes)
SELECT ID_Cliente, Of_producto, Producto, COD, Saldo, Num_producto, Estado, 'Febrero'
FROM Clientes_FEB;

DROP TABLE Clientes_ENE;
DROP TABLE Clientes_FEB;

select * from Norma;

```

#### 4.6 Código SQL + Resultados: Vistas (*Primera entrega*)

OF_PRODUTO	PRODUCTO	COD	SALDO	NUM_PRODUCTO	ESTADO	MES
30030	A Término	CD	242	206158	Activo	Enero
30005	Programado	Programado	35	58500054652	Activo	Enero
30005	Programado	Programado	68	58500054653	Activo	Enero
30005	Programado	Programado	109	58500054654	Activo	Enero
30005	Programado	Programado	162	58500054655	Activo	Enero
30005	Programado	Programado	510	58500054656	Activo	Enero
30005	Programado	Programado	53	58500054657	Activo	Enero
30005	Programado	Programado	393	58500054658	Activo	Enero
30006	Permanente	Permanente	968	58500055676	Activo	Enero
30017	Permanente	Permanente	604	58500060019	Activo	Enero
30017	Permanente	Permanente	524	58500060020	Activo	Enero
30017	Permanente	Permanente	950	58500060021	Activo	Enero
30019	Permanente	Permanente	803	58500061100	Activo	Enero
30028	Permanente	Permanente	23	58500064362	Activo	Enero
30028	Permanente	Permanente	751	58500064363	Activo	Enero
30028	Permanente	Permanente	49	58500064364	Activo	Enero
30028	Permanente	Permanente	712	58500064365	Activo	Enero
30028	Permanente	Permanente	58	58500064366	Activo	Enero
30028	Permanente	Permanente	780	58500064367	Activo	Enero
30032	Permanente	Permanente	729	58500066575	Activo	Enero
30032	Permanente	Permanente	953	58500066576	Activo	Enero
30032	Permanente	Permanente	791	58500066577	Activo	Enero
30032	Permanente	Permanente	216	58500066578	Activo	Enero
30032	Permanente	Permanente	164	58500066579	Activo	Enero
30034	Permanente	Permanente	111	58500067655	Activo	Enero
30034	Permanente	Permanente	319	58500067656	Activo	Enero
.....	-	-	---	-----	-	-

CELULAR	OCCUPACION	PAGADURIA	EMPRESA	ESTADO_CIVIL	EDAD	GENERO	SALARIO
6910880	(null)	SEC DE EDUCACION SANTAND	(null)	(null)	(null)	(null)	(null)
3163202474	(null)	(null)	(null)	(null)	(null)	(null)	(null)
6578988	(null)	(null)	(null)	(null)	(null)	(null)	(null)
5,73144E+11	Independiente	(null)	(null)	Casado/a	26	Femenino	0
(null)	(null)	(null)	(null)	(null)	(null)	(null)	0
5,73204E+11	Empleado Público	INSTITUTO DISTRITAL PARA LA PROTECCION DE LA NIÑEZ Y LA JUVENTUD	IDIFRON	Soltero/a	27	Femenino	1700000
(null)	Empleado Privado	(null)	(null)	(null)	22	Femenino	0
5,73016E+11	Independiente	DATO DEFURABLE	(null)	Sin información	32	Femenino	0
5,73135E+11	Empleado Privado	(null)	ROYAL SEGURIDAD	Soltero/a	30	Masculino	1100000
5,73014E+11	(null)	(null)	(null)	Soltero/a	(null)	Femenino	846000
5,73215E+11	(null)	(null)	(null)	Unión libre	28	Femenino	0
5,73183E+11	Estudiante	(null)	SIN INFORMACION	Soltero/a	20	Femenino	900000
3138341716	Independiente	(null)	INDEPENDIENTE	Soltero/a	22	Masculino	1000000
(null)	(null)	(null)	(null)	(null)	37	(null)	0
5,73142E+11	(null)	(null)	(null)	(null)	(null)	(null)	(null)
5,73126E+11	(null)	(null)	(null)	(null)	(null)	(null)	(null)
5,73123E+11	Empleado Público	(null)	FISCALIA GENERAL	Soltero/a	47	Masculino	2363065
5,73174E+11	Independiente	(null)	DEPARTAMENTO ADM NACION	Soltero/a	34	Femenino	0
5,73144E+11	Empleado Privado	(null)	STAFF	Soltero/a	23	Masculino	908526
5,73133E+11	Empleado Privado	(null)	Indra Colombia	Soltero/a	22	Masculino	877000
5,73505E+11	Estudiante	(null)	NO APLICA	Soltero/a	21	Masculino	877503
5,73173E+11	Estudiante	(null)	ESTUDIANTE	Soltero/a	22	Masculino	3900000
3015447087	Empleado Privado	(null)	SERVICES Y CONSULTING	Soltero/a	19	Masculino	1100000
5,73164E+11	Estudiante	SIN PAGADURIA	NO APLICA	Soltero/a	21	Femenino	100000
5,73137E+11	Empleado Público	(null)	Sin Información	Soltero/a	47	Masculino	6810316
(null)	Independiente	(null)	SIN INFORMACIÓN	(null)	47	Masculino	0
5,73165E+11	Estudiante	(null)	NO APLICA	Soltero/a	23	Femenino	1000000
5,73144E+11	Estudiante	DATO DEFURABLE	(null)	Soltero/a	22	Masculino	670000
5,73116E+11	Estudiante	(null)	NO APLICA	Soltero/a	21	Masculino	500000
5,73156E+11	Empleado Privado	DATO DEFURABLE	(null)	Sin información	22	Masculino	828116

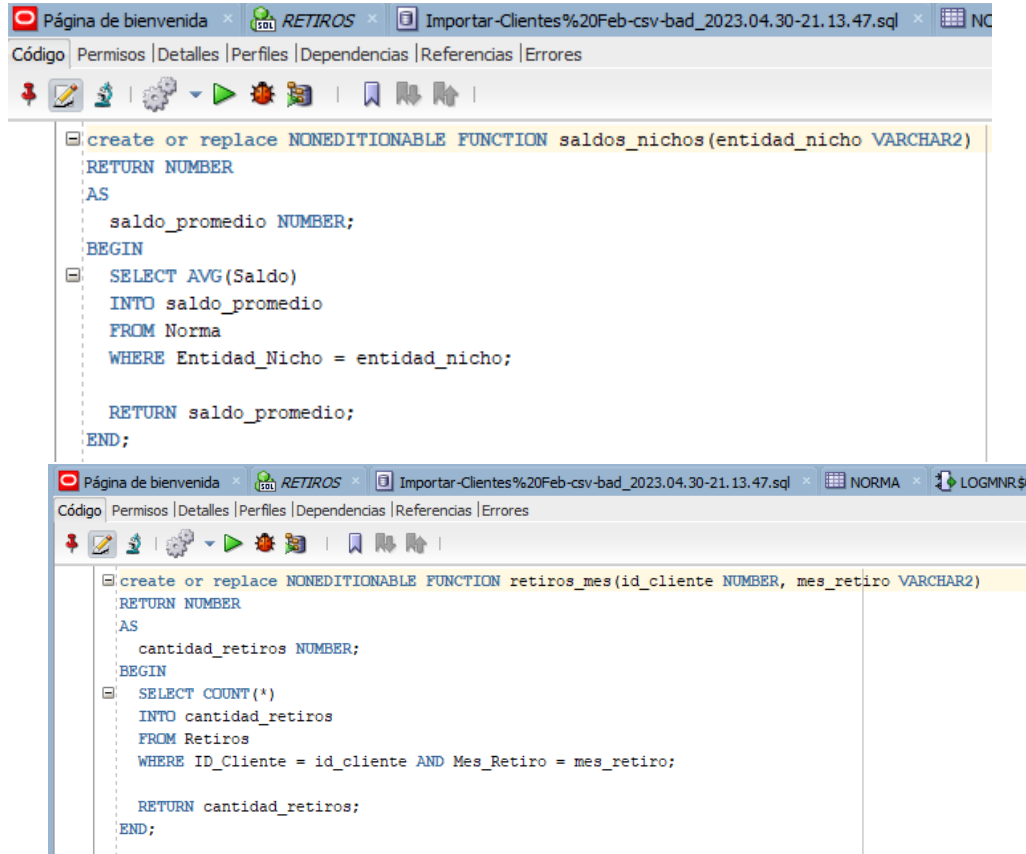
#### 4.7 Código SQL + Resultados: Triggers (*Primera entrega*)

```

## tigger
CREATE OR REPLACE TRIGGER actualiza_entidad
AFTER INSERT OR UPDATE ON Demograficos
FOR EACH ROW
BEGIN
UPDATE Retiros
SET Entidad = :new.Pagaduria
WHERE ID_Cliente = :new.ID_Cliente;
END;

```

#### 4.8 Código SQL + Resultados: Funciones (*Primera entrega*)



The image displays two screenshots of an SQL IDE interface. The top screenshot shows the creation of a function named `saldos_nichos`. The bottom screenshot shows the creation of a function named `retiros_mes`.

```
create or replace NONEDITIONABLE FUNCTION saldos_nichos(entidad_nicho VARCHAR2)
RETURN NUMBER
AS
    saldo_promedio NUMBER;
BEGIN
    SELECT AVG(Saldo)
    INTO saldo_promedio
    FROM Norma
    WHERE Entidad_Nicho = entidad_nicho;

    RETURN saldo_promedio;
END;
```

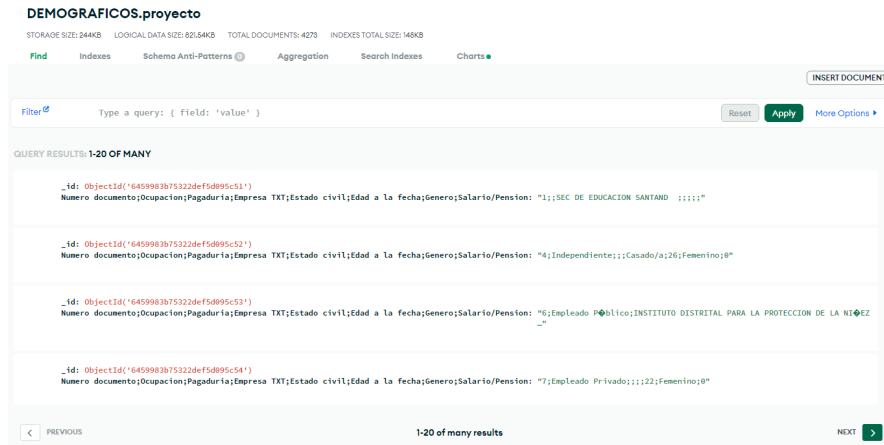
```
create or replace NONEDITIONABLE FUNCTION retiros_mes(id_cliente NUMBER, mes_retiro VARCHAR2)
RETURN NUMBER
AS
    cantidad_retiros NUMBER;
BEGIN
    SELECT COUNT(*)
    INTO cantidad_retiros
    FROM Retiros
    WHERE ID_Cliente = id_cliente AND Mes_Retiro = mes_retiro;

    RETURN cantidad_retiros;
END;
```

#### 4.9 Código SQL + Resultados: procedimientos almacenados (*Primera entrega*)

## 5 Bases de Datos No-SQL (*Segunda entrega*)

### 5.1 Diagrama Bases de Datos No-SQL (*Segunda entrega*)



```
BD proyecto NoSQL.ipynb
Archivo Editor Ver Insertar Entorno de ejecución Herramientas Ayuda Se han guardado todos los cambios

+ Código + Texto

[1] 1 !pip install pymongo
    2 !pip install dnspython

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pymongo
  Downloading pymongo-4.3.3-cp310-cp310-manylinux_2_17_x86_64_manylinux2014_x86_64.whl (492 kB)
    492.0 kB/s eta 0:00:00
Collecting dnspython<3.0.0,=>1.16.0 (from pymongo)
  Downloading dnspython-2.3.0-py3-none-any.whl (283 kB)
    283.7 kB/s eta 0:00:00
Installing collected packages: dnspython, pymongo
Successfully installed dnspython-2.3.0 pymongo-4.3.3
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: dnspython in /usr/local/lib/python3.10/dist-packages (2.3.0)

1 import pymongo
2 from pymongo import MongoClient
3 from pymongo.server_api import ServerApi
4
5 # Conectarse a la base de datos de MongoDB Atlas
6 client = MongoClient("mongodb+srv://nfernandez@cluster0.arzkuw.mongodb.net/?retryWrites=true&majority", server_api=ServerApi('1'))

[ ] 1 !mongoimport --host <ac-m8sgqci-shard-00-01.arzkuw.mongodb.net:27017> --port <27017> --username <nfernandez> --password <Nikis1984> --db <DEMOGRAFICOS> --collection

[3] 1 db = client.DEMOGRAFICOS
    2 collection = db.proyecto

[4] 1 # Send a ping to confirm a successful connection
    2 try:
    3     client.admin.command('ping')
    4     print("Pinged your deployment. You successfully connected to MongoDB!")
    5 except Exception as e:
    6     print(e)

Pinged your deployment. You successfully connected to MongoDB!
```

### 5.2 SMDB utilizado para la Base de Datos No-SQL (*Segunda entrega*)

Se usa MongoDB Atlas como SMDB para la Base de Datos No-SQL al ser una solución completa para la gestión de bases de datos NoSQL, al adaptarse a las necesidades de una amplia variedad de aplicaciones y permitir escalar, asegurar y administrar los datos de manera efectiva y eficiente.

La tabla que se usa en MongoDB es la de demográficos cargando 4273 registros por importación con las herramientas de MongoDB, se conecta a python y se

trabaja con códigos desde esta aplicación.

## 6 Aplicación de ETL (Extract, Transform, Load) y Bodega de Datos

Con la implementación de nuestro modelo de datos y considerando que recibimos información de forma mensual, utilizamos la aplicación Power BI para realizar la lectura y transformación de datos, facilitando su visualización y accesibilidad para personas fuera del equipo de análisis.

Las ventajas de utilizar Power BI como herramienta de almacenamiento de datos y para la realización de ETL en nuestra base de datos, creada con Oracle SQL, son las siguientes:

- Conectividad amplia: Power BI ofrece una amplia gama de conectores de datos, lo que permite extraer datos de diversas fuentes, como bases de datos, archivos, servicios en la nube, aplicaciones web, entre otros. Esto facilita la integración de múltiples fuentes de datos en el proceso de ETL.

- Transformación de datos: Power BI proporciona una funcionalidad robusta para transformar datos. Mediante su editor de consultas, los usuarios pueden realizar operaciones de limpieza, filtrado, combinación, agregación y enriquecimiento de datos. También es posible aplicar transformaciones más avanzadas utilizando el lenguaje M o lenguaje DAX.

- Automatización y programación: Power BI permite automatizar el proceso de ETL mediante la creación de consultas y scripts. Los usuarios pueden programar la ejecución de consultas en horarios específicos o configurar actualizaciones automáticas de los datos.

- Modelado de datos: Power BI incluye una funcionalidad de modelado de datos intuitiva y potente. Permite crear relaciones entre tablas, definir medidas y cálculos personalizados, y diseñar modelos de datos eficientes para el análisis y visualización.

- Visualización de datos: Una vez que los datos han sido transformados y cargados en el modelo de datos de Power BI, se pueden crear visualizaciones interactivas y atractivas. Power BI ofrece una amplia variedad de gráficos, tablas y visualizaciones personalizadas que permiten explorar y presentar los datos de manera efectiva.

- Publicación y distribución: Power BI permite publicar y compartir los informes y paneles creados con otros usuarios. Además, ofrece opciones de colaboración y acceso a través de la nube, lo que facilita la distribución de los

resultados del ETL a las partes interesadas.

## 6.1 Ejemplo de aplicación de ETL y Bodega de Datos

Una vez establecida la conexión con nuestra base de datos, procedemos a realizar transformaciones en los datos. Esto se hace con el objetivo de asegurar que los datos nuevos se ajusten adecuadamente a los cargados inicialmente. Configuramos la base de datos utilizando códigos del lenguaje DAX.

```
let
    Origen = Excel.Workbook(File.Contents("C:\Users\ldreyes\Documents\COOPERATIVA\MODELO\MODELO.xlsx"), null, true),
    MODELO_Sheet = Origen[Item="MODELO (2)",Kind="Sheet"][Data],
    #"Encabezados promovidos" = Table.PromoteHeaders(MODELO_Sheet, [PromoteAllScalars=true]),
    #"Tipo cambiado" = Table.TransformColumnTypes(#"Encabezados promovidos",{{"Nit", type text}, {"PROBABILIDAD_BOSQUE", Percentage.Type}, {"PROBABILIDAD_LOGIT", Percentage.Type}})
in
    #"Tipo cambiado"
```

Una vez establecida la conexión, realizamos la transformación de los datos para adaptarlos a un modelo de datos panel o para su visualización en Power BI. Esto se debe a que tanto Oracle como Power BI funcionan con la estructura de bases de datos SQL.

```
1 RANGO_EDAD = IF (MODELO[EDAD]>=70,"70-90",IF (AND(MODELO[EDAD]<70,MODELO[PROBABILIDAD_BOSQUE]>=50),"70-50",IF (AND(MODELO[EDAD]<50,MODELO[EDAD]>=40),"50-40",IF (AND(MODELO[EDAD]<40,MODELO[EDAD]>=25),"40-25","menor a 25"))))
```

Una vez creados los códigos de transformación y carga de información, se presenta un panel de control o una visualización gráfica de los datos estructurados. Esto permite una interpretación fluida de los datos y facilita el acceso a personas que no están directamente involucradas en el área de análisis.





## 7 Proximos pasos

Como parte del proceso de carga y transformación de datos, es fundamental implementar modelos de inteligencia artificial utilizando los datos almacenados en la base de datos. Esto se debe a que los datos por sí solos tienen un gran valor, pero sin configuraciones estadísticas no se puede obtener una comprensión más profunda de la información.

El objetivo de este estudio es aplicar modelos estadísticos para crear estrategias de información u ofertas dirigidas a los clientes de una compañía financiera. De esta manera, se busca lograr la fidelización de clientes existentes y atraer a nuevos clientes. Para lograrlo, es esencial leer y cargar los datos en los modelos de inteligencia artificial disponibles en la actualidad.

## 8 Lecciones aprendidas

Las lecciones aprendidas en este trabajo son las siguientes:

- Manipulación de datos: Es fundamental comprender cómo estructurar una base de datos correctamente para realizar análisis descriptivos o predictivos de los datos proporcionados.
- Conexión de bases de datos: Es necesario entender las variables adecuadas y reconocer las claves primarias y foráneas en la estructura de las bases de datos.
- Conexión de archivos planos (CSV o Excel): Se debe ser capaz de leer y transformar archivos planos utilizando programas de bases de datos como MySQL, Oracle y SQL Server mediante la escritura de consultas relacionales.
- Creación de usuarios y máquinas remotas: Se requiere la capacidad de configurar usuarios y máquinas remotas para realizar transformaciones de datos en diferentes lenguajes de programación. Esto permite aplicar modelos estadísticos a la estructura modificada de la base de datos.

Estas lecciones aprendidas son fundamentales para desarrollar un análisis de datos efectivo y utilizar herramientas y técnicas avanzadas en el proceso de transformación y análisis de datos.

## 9 Bibliografía

- [1] H. Arellano Díaz, La calidad en el servicio como ventaja competitiva, Dominio de las Ciencias, vol. III, pp. 72-83, 2017.
- [2] D. P. Puerto Becerra, La globalización y el crecimiento empresarial a través de estrategias de internacionalización, Pensamiento Gestión, nº 28, pp. 171-195, 2010.
- [3] Guangli, N., Lingling, Z., Xingsen, L., Yong, S. (2011). The Analysis on the Customers Churn of Charge Email Based on Data Mining Take One Internet Company for Example. Institute of Electrical and Electronics Engineers, 843-847.
- [4] J. Rozum, Defining and Understanding Software Measurement Data, Software Engineering Institute Carnegie Mellon University, 2002.
- [5] R. Feinberg y M. Trotter, Immaculate deception: the unintended negative effects of the CRM revolution: maybe we would be better off without customer relations management., Defying the Limits, pp. 26-31, 2001.
- [6] Chitra, K., Subashini, B. (2011). Customer Retention in Banking Sector using Predictive Data Mining Technique. The 5th International Conference on Information Technology.
- [7] Polo Redondo, Y., Sesé Olivan, F. J. (2009). La retención de los clientes un estudio empírico de sus determinantes. Revista Española de Investigación de Marketing, 117-137
- [8] Sarkar, D., Bali, R., Sharma, T. (2018). Practical machine learning with Python. A Problem-Solvers Guide To Building Real-World Intelligent Systems. Berkely: Apress.
- [9] Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." Journal of Educational Psychology 66 (5): 688-701.
- [10] Little, R. J. A., D. B. Rubin, and S. Z. Zangeneh. 2017. "Conditions for Ignoring the Missing-Data Mechanism in Likelihood Inferences for Parameter Subsets." Journal of the American Statistical Association 112 (517): 314-20.
- [11] Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and Regression Trees. New York: Wadsworth Publishing.
- [12] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, O'Reilly Media, Inc., 2019.
- [13] Bruno. M., (2022). Implementación de un modelo de minería de datos para predecir la deserción de los clientes en una empresa de telecomunicaciones. Universidad católica santo toribio de Mogrovejo. 20-28.
- [14] Bohorquez. M., Torys. J., Paredes. M., (2020). MODELOS DE PREDICCIÓN DE DESERCIÓN DE CLIENTES PARA. Revista Compendium: Cuadernos de Economía y Administración., Vol 7, No 1, 1-11.
- [15] R. A. Barrueta Meza and E. J. P. Castillo Villarreal, "Modelo de análisis predictivo para determinar clientes con tendencia a la deserción en bancos peruanos," Universidad Peruana de Ciencias Aplicadas(UPC)., Lima, Perú, 2018. Doi: <http://doi.org/10.19083/tesis/626023>.
- [16] Torrado, M. y Berlanga, V. (2013). Análisis Discriminante mediante SPSS. [En línea] REIRE, Revista d'Innovació i Recerca en Educació, 6 (2),

150-166.

[17] Hastie, T., Friedman, J., y Tibshirani, R. (2001). The Elements of Statistical Learning. Nueva York, Estados Unidos: Springer New York. DOI: 10.1007/978-0-387-21606-5.

[18] Rossiter, D. 1994. Basic Concepts and Procedures on Land Evaluation. Cornell University course Soil, Crop Atmospheric Sciences. 'Special Topics in Soil, Crop Atmospheric Sciences: Land evaluation, with emphasis on computer applications', Spring Semester 1994.

[19] Ley 1584 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. Publicada en el Diario Oficial 48587 de octubre 18 de 2012.

[20] Russo, M. y Ferrari, F. 2016. Analyzing Data with Power BI and Power Pivot for Excel (ISBN 9781509302765). Microsoft. 2016.

[21] Grus, J. (2019). Data science from scratch: first principles with python. O'Reilly Media.

[22] Nina-Alcocer, V., Blasco-Gil, Y., Peset, F. (2013). Datasharing: guía práctica para compartir datos de investigación. El profesional de la información, 22(6), 562-568.

[23] Downey, A. (2014). Think stats: exploratory data analysis. " O'Reilly Media, Inc."