Proyecto de Bases de datos para 2023

Brayan David Reyes Morales

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

April 18, 2023

Contents

1 Introducción

La motivación principal en la creación y mantenimiento de las empresas, (sin importar el sector a que pertenezca) 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 es olvidada dado a que se prioriza la obtención de clientes nuevos, pero no se prioriza los niveles de deserción en las compañías.

En la actualidad, con la globalización de los servicios y la amplia competencia, los productos brindados por las organizaciones son cada vez más similares en calidad y precio [1], [2], con la competencia que se genera naturalmente por las dinámicas empresariales los clientes suelen tener una alta rotación entre las compañías, ante esto las compañías se vuelven dinámicas en función de los clientes, es pro ello que se hace indispensable conocer a los clientes, con el fin de explicar sus acciones en el futuro.

Ante la problemática de una amplia oferta de productos y servicios, los clientes tienen la facilidad de cambiar de una compañía a otra, ante este problema de rotación las entidades han priorizado una alta competencia en cuanto la atracción de clientes nuevos, pero no es la manera correcta de solución del problema de retiros de una compañía. Ante esto se busca un relacionamiento orientado a fidelizar un cliente de manera correcta y responsable, dado los productos y servicios ofrecidos por cada compañía y cada industria. Aquellos clientes que dejen de usar los productos de una empresa son denominados desertores, identificarlos permite aplicar estrategias para retenerlos [3], es por ello por lo que se hace indispensable conocer el entorno de las compañías y así se puedan generar propuestas de retención efectivas, traduciendo en una deserción menor y una masa de clientes superior a sino se aplicaran las estrategias.

Dado a que la obtención de clientes ha cambiado dinámicamente ahora se pieza en la retención de una manera efectiva, ya que es demostrado que en la actualidad las industrian implementan mayores recursos a la atracción, mientras que las estrategias de retención son mejores en cuanto a costos, algunas investigaciones nos muestran que el costo de obtener de un nuevo cliente es de 5 a 7 veces mayor al costo de retención de uno antiguo [4], "The loyalty effect" donde se demostró que el aumento del 5% en la tasa de retención de clientes logró aumentos del 35% y el 95% en el valor actual neto de los clientes en una empresa desarrolladora de software y una agencia de publicidad, respectivamente. [5]

Por los motivos anteriormente descritos se propone un modelo estadístico para identificar los comportamientos de los clientes y así lograr un modelo de retención exitoso dirigido a usuarios del sistema financiero, apoyándose en modelos estudiados e implementados en diferentes trabajos o investigaciones pertinentes para predecir el nivel de deserción de clientes para una financiera.

Para abordar el problema se debe definir los criterios de evaluación (variables de dataset), ante estos criterios el mas importante a definir es cuando un cliente es marcado como desertor dentro de la compañía. La definición de clientes desertores es cambiante debido a los comportamientos de los clientes en diferentes industrias y tipos de mercado a los que hoy se exponen los consumidores.

Dentro del sector bancario, un cliente desertor es aquel que cierra todas sus cuentas bancarias y cesa los negocios con el banco en estudio [6].

Al tener una definición objetiva sobre lo que es un cliente desertor para el ejercicio, el siguiente paso es abordar los datos disponibles para la realización de los datos y se hace indispensable la minería de datos para el estudio, ya que a partir de este punto se puede presentar la mayoría de tiempo implementado para lograr un modelo exitoso. La importancia de la minería de datos en el abordaje de estudios es una técnica estadística considerada como una herramienta estratégica importante para las empresas, debido a que, este método permite analizar grandes volúmenes de información desde diferentes perspectivas y sintetizarlos en información valiosa que permite la temprana y efectiva toma de decisiones de las organizaciones empresariales, sirviendo de gran utilidad para crear un perfil de cliente preciso basado en el comportamiento del cliente [7].

La minería de datos se hace importante para la toma de decisiones en la compañía ya que aportan una información comprimida, para lograr un entendimiento de lo que está pasando o de algo que sucedió en el pasado. El modelamiento de los datos es de gran importancia en las compañías ya que permite generar una información concisa y precisa para aplicar modelos estadísticos. Los modelos estadísticos por aplicar para este caso, teniendo en cuenta que va dirigido hacia una compañía del sector financiero colombiano y estudios previamente realizados por diferentes autores, se obtiene que comúnmente se aplican algoritmos de programación de los siguientes modelos: Arboles de decisión, Redes Neuronales, Reglas de Asociación, Regresión Logística, Arboles Aleatorios y SVM. Los modelos mencionados permiten ajustarse a cualquier industria, dado como se haga la minería permitirá obtener un modelo exitoso para predecir el comportamiento de los clientes.

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

El estudio se realizará para una entidad financiera Colombia, se debe tener en cuenta que aunque los modelos pueden ser aplicables a diferentes industrias y empresas, no todas la empresas son iguales y tienen ciertos detalles que la hacen única y así mismo a sus clientes. Dado que para el siguiente estudio se hace basado en los datos de una financiera, se debe definir de donde se obtendrán los datos y el tiempo de estos, como se evidencia en la tabla 1, se pueden tomar datos a partir de una cantidad especifica definiendo a los clientes retirados, hay una segunda opción de tomar a los clientes totales de los últimos meses, no mayor a 6 meses, e identificar los retirados. En el caso de estudio se tomarán los clientes de tres meses específicos, con su variable categórica (retirados), calidad, y los tipos de datos que se generen a partir de la base entregada.

A partir de la identificación de las fuentes de datos disponibles en la organización, se procede con el modelo CRISP-DM, tratando de responder la pregunta de negocios “¿Qué tipo de clientes tienen una alta probabilidad de retirarse en

el futuro?", para ello se ha generado una base con diferentes variables, que ayudaran a explicar la pregunta de negocio.

2.1 Titulo del proyecto de investigación

Construccion correcta de bases de datos, para la aplicacion de modelos supervisados, que permitan explicar los retiros de clientes.

2.2 Objetivo general

Construccion correcta de bases de datos para la aplicacion de modelos estaditcos que aporten informacion de clientes para compañias de financiemmiento.

2.2.1 Objetivos especificos

- Estructura y creacion de tablas de datos que permitan extraer informacion de datos.
- Aprendizaje de lenguaje SQL para consulta y extracion de datos.
- Comprension de datos estructurados para consulta de informancion de distintas bases de datos.

2.3 Alcance

Con la correcta extraccion y manipulacion de los datos se busca crear una base de datos solida que permita extraer informacion finaniciera y sociodemografica de clientes retirados por parte de una entidad financiera, manejnado sistemas informaticos como SQL ya que el manjeo de los datos deben ser estructurados.

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 especifico 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.

2.4 Pregunta de investigación

¿Como generar una base de datos correctas con diferentes fuentes de informacion utilizando el concepto de bases estructuradas o SQL?

2.5 Hipotesis

A traves de lenguajes de progrmacion como SQL se puede obtener informacion de funtes de diferentes areas de una organizacion para la generacion de modelos estaditicos que permitan brindar visiones sobre el comportamiento de clientes.

3 Reflexiones sobre el origen de datos e información

(Max 400 Palabras) - (*Primera entrega*)

Se espera que una entidad tenga sus datos ubicados en un solo lugar y con poca información faltante, sin embargo, muchas veces en las compañías esto no ocurre, ya sea en tiempo real o tiempo vencido. Para este ejercicio en particular, la recopilación y acceso a la información puede estar limitada por otras áreas, ya que existe cierta precaución en el manejo de transacciones y acceso a los archivos.

Una vez superado el problema de los permisos, surge otro obstáculo al identificar las variables relevantes para generar una base de datos, ya que muchas de ellas manejan información parecida o repetida en tablas diferentes. Esto provoca molestias y un desgaste de tiempo en la minería de datos, especialmente si los datos se distribuyen en múltiples tablas. Una vez organizada la información, se puede crear una tabla de cinco variables conformadas por una variable de cada tabla, lo cual puede ser un gran obstáculo en las entidades, ya que se deja de lado el modelamiento y se requiere de un esfuerzo significativo para la creación de una base comprensible y análisis descriptivo, con la esperanza de que las variables aporten 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 eticas del uso de la información?

No se encuentra ninguna consideracion etica o legal que impidan realizar el estudio de los datos.

3.3 ¿Cuales son los retos de la información y los datos que utilizara en la base de datos en terminos 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:

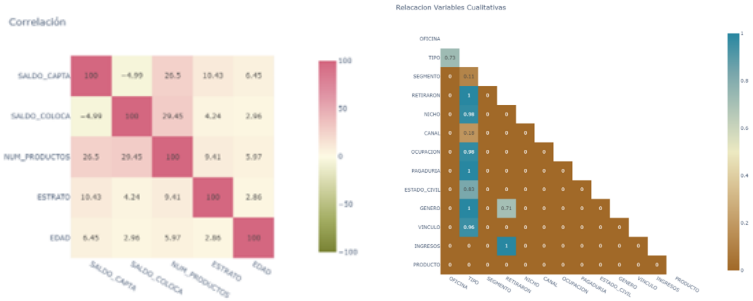
	Nombre Columna	Tipo	Cantidad de Únicos	Cantidad de Nulos	Cantidad Valores No Nulos
OFICINA	OFICINA	object	37	0	46144
TIPO	TIPO	object	2	0	46144
SEGMENTO	SEGMENTO	object	4	0	46144
RETIRARON	RETIRARON	object	2	0	46144
NICHO	NICHO	object	12	0	46144
CANAL	CANAL	object	6	0	46144
OCUPACION	OCUPACION	object	6	0	46144
SALDO_CAPTA	SALDO_CAPTA	float64	25024	0	46144
SALDO_COLOCA	SALDO_COLOCA	float64	23294	0	46144
NUM_PRODUCTOS	NUM_PRODUCTOS	int64	28	0	46144
PAGADURIA	PAGADURIA	object	291	20025	26119
ESTADO_CIVIL	ESTADO_CIVIL	object	6	5166	40978
ESTRATO	ESTRATO	int64	6	0	46144
EDAD	EDAD	float64	83	1273	44871
GENERO	GENERO	object	2	811	45333
VINCULO	VINCULO	object	3	0	46144
INGRESOS	INGRESOS	object	16005	665	45479
PRODUCTO	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
count	4.614400e+04	4.614400e+04	46144.00	46144.00	44871.00
mean	1.273259e+07	1.744456e+07	1.79	2.79	54.31
std	7.850941e+07	3.752638e+07	1.30	1.04	16.12
min	-1.214730e+05	0.000000e+00	1.00	1.00	13.00
25%	0.000000e+00	0.000000e+00	1.00	2.00	41.00
50%	1.005045e+05	2.085620e+06	1.00	3.00	56.00
75%	2.317506e+06	1.996970e+07	2.00	3.00	66.00
max	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 presenta 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 perdida 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 especifico 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 a 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*)

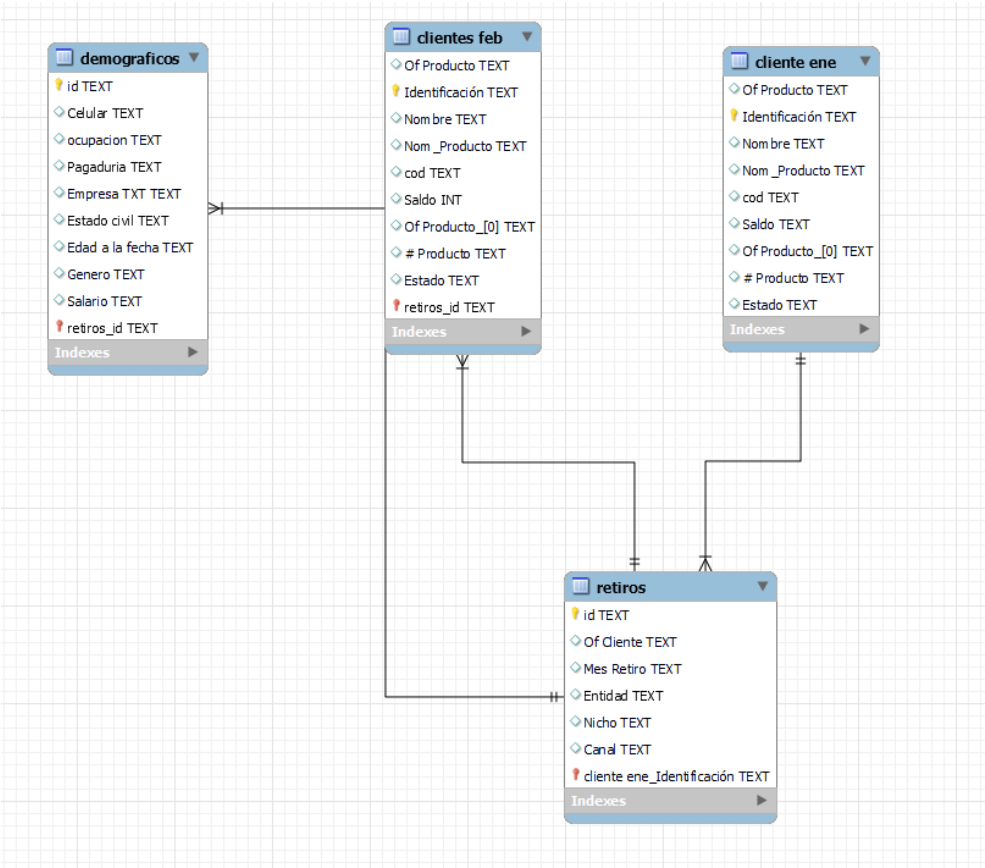
4.1 Características del SMBD (Sistema Manejador de Bases de Datos) para el proyecto (*Primera entrega*)

PAra el trabajo actual se utilizara MySQL el cual permite modificar y generar tablas que permiten realizar la coneccion a traves de llaves, para el caso se tiene una llave unica que sera el numero de identificacion del usuario.

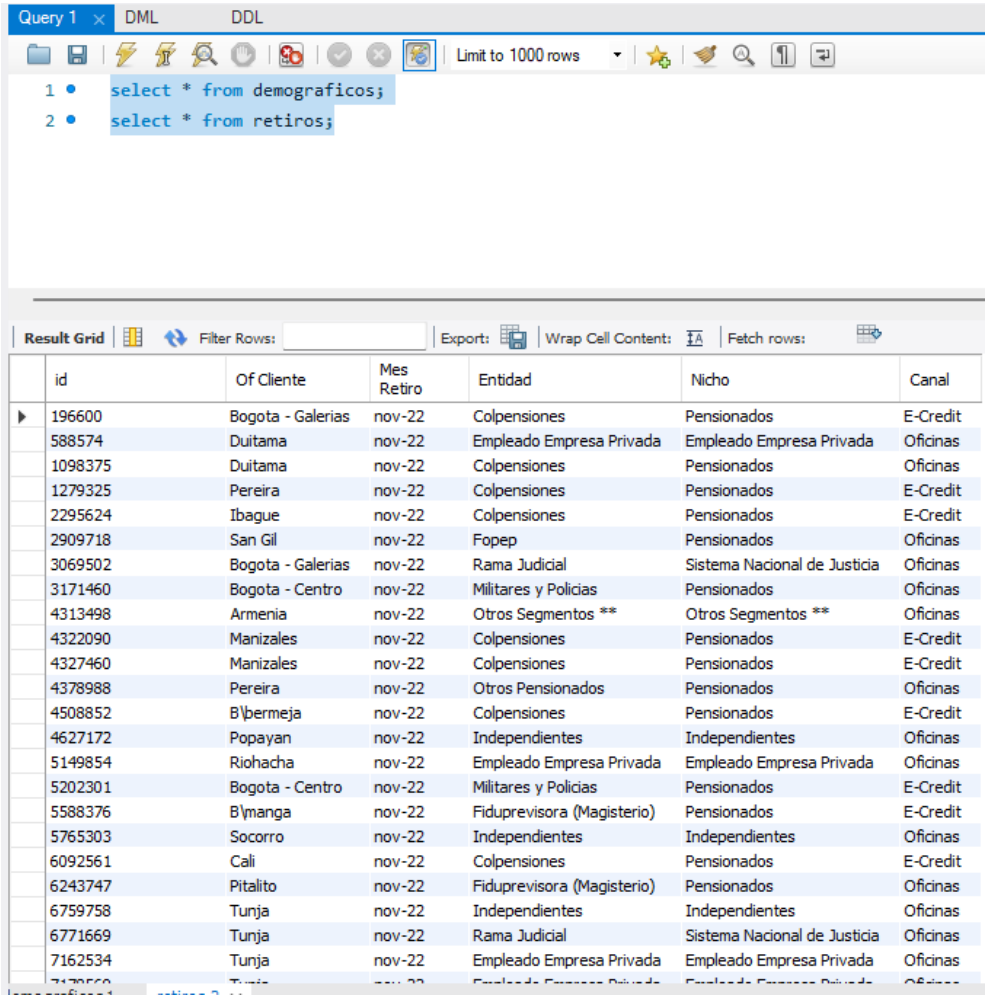
MySQL Workbench es un sistema libre que permite trabajar con una cantidad limitada de datos, pero para el proyecto actual es suficiente ya que posee las siguientes caracteristicas:

- 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.

4.2 Diagrama modelo de datos (*Primera entrega*)



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



id	Of Cliente	Mes Retiro	Entidad	Nicho	Canal
196600	Bogota - Galerias	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 - Galerias	nov-22	Rama Judicial	Sistema Nacional de Justicia	Oficinas
3171460	Bogota - Centro	nov-22	Militares y Policias	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 Policias	Pensionados	E-Credit
5588376	B\manga	nov-22	Fiduprevisora (Magisterio)	Pensionados	E-Credit
5765303	Socorro	nov-22	Independientes	Independientes	Oficinas
6092561	Cali	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

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

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

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

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

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

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

5 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.