**ESPECIALIDAD EN CIENCIA DE DATOS**

**Proyecto Terminal**

**Clasificación de nuevos clientes para elaboración de campañas personalizadas**



**Profesor**: Dr. Carolina del Valle Soto

**Alumno**: Jesús Octavio Venegas Avalos

**ID**: [0230559@up.edu.mx](mailto:0230559@up.edu.mx)

**Problema de investigación**

En el área de inteligencia de datos de la empresa Bodesa S.A.P.I. de C.V., se desconoce a que segmento pertenecerá cada cliente nuevo una vez que es aprobado su crédito. El no conocer el perfil del cliente afecta la eficiencia operativa debido a que no es posible crear campañas de mercadotecnia dirigidas de acuerdo con el perfil de estos clientes y en su lugar se lanzan campañas de manera masiva y generales ocasionando un incremento en los costos de operación y una baja redención de compra.

**Justificación del problema**

Se cuenta con un sistema de crédito propio con aproximadamente 400,000 clientes activos, los cuales están clasificados en seis segmentos previamente establecidos por el área de mercadotecnia, y mes con mes se crean alrededor de 3,500 cuentas nuevas para las dos unidades de negocio que se tienen: La Marina y El bodegón.

Es importante tener campañas publicitarias bien dirigidas de acuerdo con el perfil de cada cliente, esto incluye a los clientes nuevos, por lo que es necesario saber a qué perfil o segmento pertenecerán éstos, y poder enviar publicidad dirigida de acuerdo con esta clasificación.

De la misma manera es necesario conocer a que segmento de clientes logramos incrementar la redención de compra y a cuáles no, con el fin de replantear nuevas estrategias en torno a las campañas enviadas para cada clasificación.

El tener más clientes activos, representa tener más transacciones y por ende aumento en la utilidad de la empresa.

Actualmente se destinan recursos tecnológicos y humanos para la realización de llamadas telefónicas, envíos de correos y mensajes por celular (SMS).

Este proyecto pretende por medio de un algoritmo de inteligencia artificial, obtener un modelo que clasifique los nuevos clientes y nos ayude a incrementar en un 40% la redención de compra y reducir en 30% los costos operativos en las campañas publicitarias.

**Hipótesis**

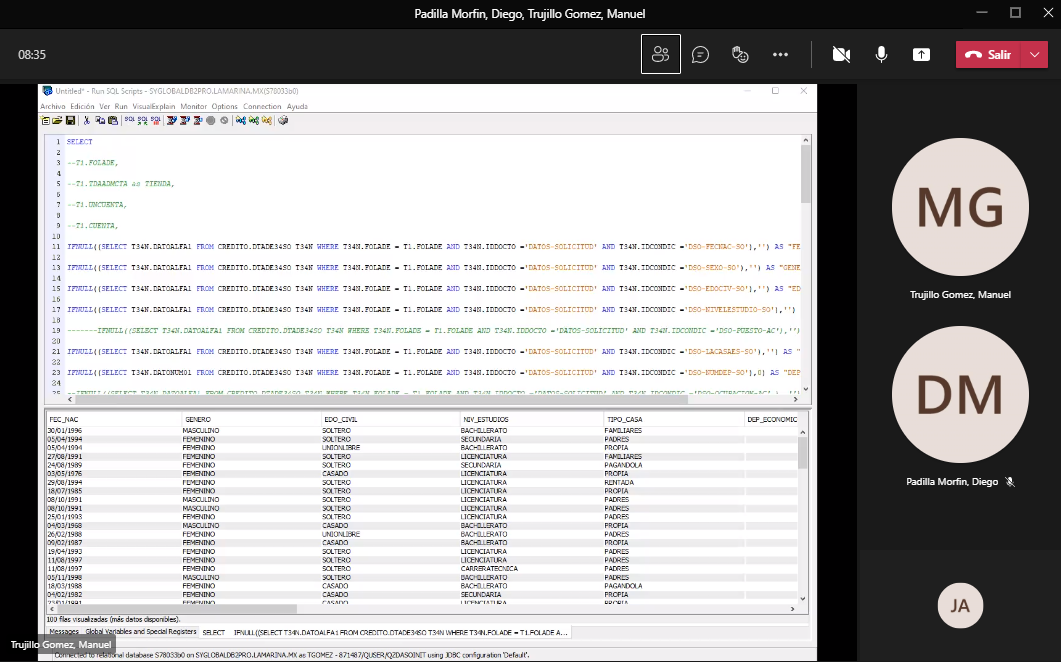
Si se implementa un algoritmo de aprendizaje automático, será posible identificar a qué segmento pertenecen los nuevos clientes, personalizando promociones de acuerdo con el perfil de cada cliente y aumentando la redención de compra.

**Evidencias de visitas o interacción con el problema de investigación**



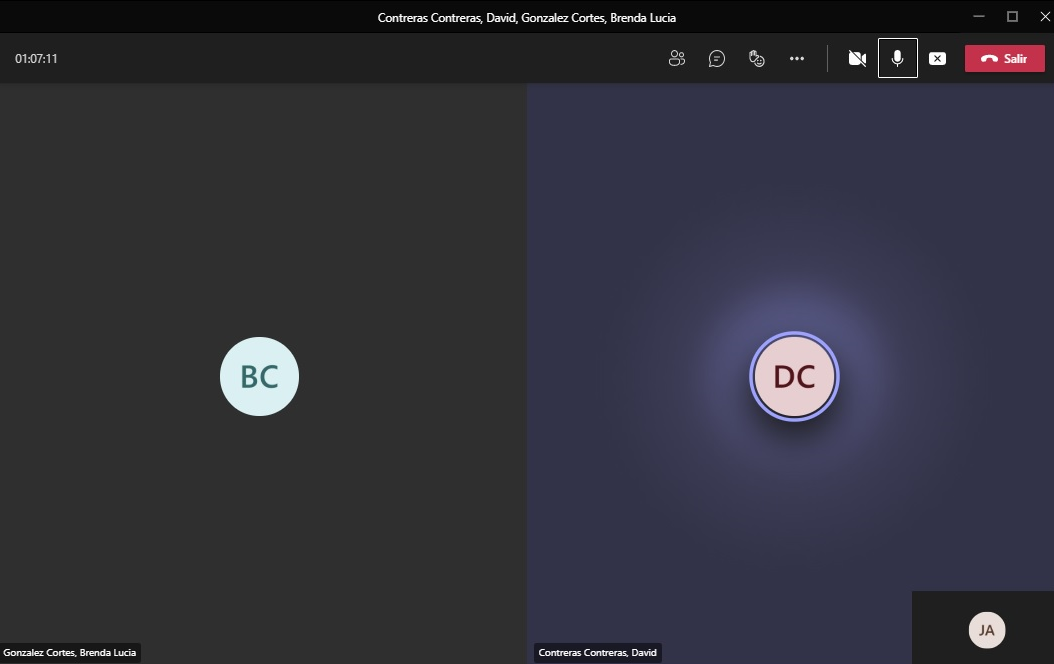
Reunión con la Lic. en mercadotecnia Brenda González Cortés, subdirectora de Inteligencia de clientes, para la revisión y explicación del problema, objetivos e hipótesis.

*Efectuada el 05 de abril del 2021 en la oficina de la subdirección*

****

Reunión con el equipo de desarrollo de T.I, Ing. en sistemas. Manuel Trujillo Gómez e Ing. en software Diego Padilla Morfin, para la extracción y análisis de variables demográficas a considerar para el proyecto.

*Efectuada el 14 de abril a través de la plataforma Microsoft Teams*



Junta con la Lic. en mercadotecnia Brenda Gonzales Cortés, subdirectora de inteligencia de clientes y con el Lic. David Contreras Analista de datos para revisión del modelo, ajustes y explicación de los resultados.

*Efectuada el 20 de mayo del 2021 a través de la plataforma Microsoft Teams*

**Marco teórico y estado del arte**

**Marco Teórico**

La inteligencia artificial tiene bastantes campos de aplicación científica y en los negocios no es la excepción. En las áreas de finanzas, Control y diseño, procesamiento de datos, desarrollo de software, etc.

Hablar de inteligencia artificial es remontarse a los años sesenta, sin embargo, en sus inicios no tuvo mucho éxito debido a los requerimientos tecnológicos necesarios para utilizarla, y solo las grandes corporaciones podían tener acceso a ésta. Hoy en día, gracias a los avances tecnológicos la inteligencia artificial está presente en muchas de las áreas comunes en nuestras vidas.

Es fundamental para entender el proyecto, conocer acerca de la inteligencia artificial. La inteligencia artificial abarca cualquier tecnología que permite a las computadoras imitar habilidades humanas, automatizando actividades para lograr un objetivo, para ello, es necesario dominar algunos campos de la educación como la estadística, análisis de datos, métodos científicos; en general, conocer de la ciencia de datos.

Ciencia de datos en la inteligencia artificial

La ciencia de datos es un subconjunto de la Inteligencia Artificial que utiliza la estadística, métodos científicos y principalmente el análisis de los datos, para extraer la información y tener una idea clara del comportamiento de los datos.

Las grandes cantidades de datos que existen actualmente representan un reto para el análisis y las dificultades inherentes a la manipulación de grandes conjuntos de datos. Día con día, los datos están adquiriendo más importancia debido a cantidad y complejidad para tratarlos.

La importancia de Big Data para el manejo de datos

La [Big Data](http://es.wikipedia.org/wiki/Big_data) consiste en encontrar los patrones de relaciones y comportamientos en el caos de los grandes volúmenes de datos que producimos. Los datos son tantos, que todos se vuelven estadísticamente significativos, lo que significa, que los métodos tradicionales que utilizamos para analizarlos se han vuelto obsoletos.

Una vez que se conoce la información con la que se va a trabajar, se presenta otro reto adicional que consiste en estudiar los datos, su comportamiento y aprender de ellos mediante herramientas informáticas, a esta rama de la inteligencia artificial se le conoce como aprendizaje automático

Métodos y técnicas de aprendizaje automático

El aprendizaje automático es el diseño y estudio de las herramientas informáticas que utilizan la experiencia pasada como entrenamiento para la toma decisiones futuras; es el estudio de programas que pueden aprender de los datos. El objetivo fundamental es generalizar, o inducir una regla desconocida a partir de ejemplos donde esa regla es aplicada.

Los métodos para el aprendizaje automático se clasifican en tres grandes grupos: aprendizaje supervisado, no supervisado y aprendizaje por refuerzo. Dependiendo del tipo de problema que se quiere resolver se determina cuál de éstos aplicar en conjunto con algunas de las técnicas para implementarlo, por ejemplo:

1. Redes neuronales

Las redes neuronales son sistemas computacionales, inspirados en las neuronas naturales dotando a los ordenadores de inteligencia artificial. Están formadas por unidades básicas llamadas neuronas que se conectan entre sí formando la red neuronal. El objetivo de éstas es entender los datos del mundo real (imágenes, texto, voz, etc.), procesarlos y clasificarlos o etiquetarlos.

1. Máquinas de Vectores de soporte (SVM)

Son un conjunto de algoritmos de aprendizaje supervisado en el que mediante un conjunto de datos de entrenamiento (muestras) se etiquetan clases y se entrenan para construir un modelo que prediga la clase de una nueva muestra.

Las SVM construyen un [hiperplano](https://es.wikipedia.org/wiki/Hiperplano) o conjunto de hiperplanos en un espacio de [dimensionalidad](https://es.wikipedia.org/wiki/Dimensi%C3%B3n) muy alta que se utilizan para realizar una separación correcta de las clases.

Clasificación de clientes

Para analizar el comportamiento futuro de los clientes primero tenemos que entender la experiencia pasada, conocer cómo transaccionan, con qué frecuencia interactúan y cuanto invierten en sus compras.

Beneficios del análisis RFM

El análisis RFM consiste en segmentar a los clientes basado en tres medidas que nos ayudan a predecir la disposición de un cliente a participar en mensajes y ofertas de marketing. Para ello es necesario saber cuándo fue su última compra (Recency), con qué frecuencia (Frecuency) han comprado en el pasado y cuanto gastan en total (Money).

Realizar un análisis de este tipo beneficia a los clientes porque se crean segmentos que distinguen campañas publicitarias personalizadas de acuerdo con cada cliente, además de que mejora la tasa de redención que nos mide la respuesta de acción de un cliente ante una acción de marketing.

Como medir los resultados de un aprendizaje automático.

La Matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo  [de aprendizaje supervisado](https://es.wikipedia.org/wiki/Aprendizaje_supervisado). Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real., o sea en términos prácticos nos permite ver qué tipos de aciertos y errores está teniendo nuestro modelo a la hora de pasar por el proceso de aprendizaje con los datos.

Otro de los conceptos utilizados son los KPI’s, que miden el desempeño de alguna acción para validar si existen o no progresos. En el aprendizaje maquina existen varias métricas para conocer los resultados de los algoritmos, en este proyecto además de la matriz de confusión utilizaremos el Accuracy que determina la exactitud del modelo.

**Estado del arte**

De acuerdo con el problema de investigación de este proyecto, el análisis del estado del arte contempla investigaciones que abarcan desde la segmentación de los clientes (como parte inicial para entender el proyecto) hasta las técnicas y herramientas utilizadas para resolver problemas similares al tema en cuestión.

En un estudio de agrupación y creación de perfiles de clientes (Maryani, Ina. 2017) explica un análisis de clasificación utilizando un modelo o técnica llamado RFM (Recencia, Frecuencia y Valor monetario) y un algoritmo de aprendizaje maquina (machine learning); que clasifica a los clientes de acuerdo con el comportamiento de compra.

Para nuestro proyecto, la segmentación hecha por el área de mercadotecnia se basa en este mismo tipo de análisis para segmentar y clasificar a los clientes actuales en los seis segmentos que se establecieron para ellos: Potenciales, Campeones, Lealtad, Atención Alta, Atención media y Perdidos.

En una primera etapa el proceso de la investigación (Maryani, Ina 2017), otorga un valor a cada atributo (RFM) como se muestra en la tabla 1, tomando como base rangos de valores para cada uno de éstos y otorgando un valor resultado como se muestra en Tabla 2.

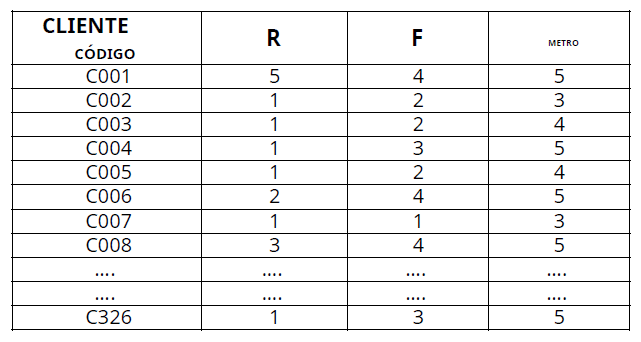


Tabla 1

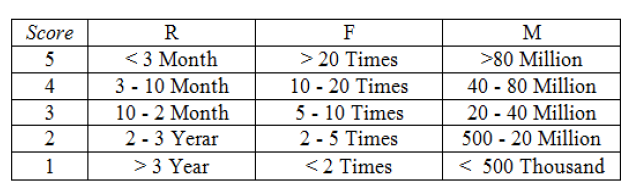


Tabla 2

Como segunda etapa, una vez obtenidos estos valores se crean clúster o segmento mediante el algoritmo K-Means para obtener las agrupaciones correspondientes Tabla 3

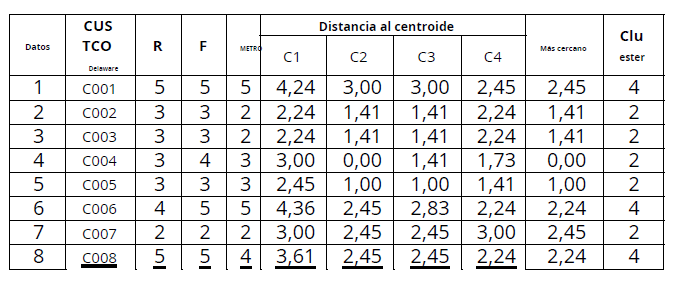


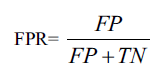
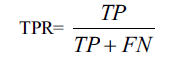
Tabla 3

Así mismo, otro de los estudios analizados compara la efectividad de tres algoritmos de aprendizaje máquina para predecir una clasificación de clientes (T.K. Das, 2015), e implementa los algoritmos K-Means, Clasificación Bayesiana y Maquinas de vectores de soporte (SVM) para una muestra de datos compuesta por 17 características de tipo numéricas, categóricas y binarias.

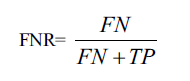
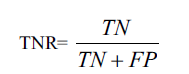
Este análisis se desarrolló para encontrar el mejor clasificador de clientes que ayude con los datos históricos que se tienen, a dirigir una oferta específica de marketing, en lugar de hacerla apuntando a todos los tipos de clientes que se tienen y para todos los tipos de promociones; es un estudio muy similar a lo que se desea en este proyecto terminal.

Para medir el rendimiento de aprendizaje del sistema se utilizó una matriz de confusión (tabla 4), como base para determinar los valores de Accuracy, Tasa de error y especificidad principalmente; sin embargo, implementa además estadísticas Kappa para validar las siguientes tasas:

Verdaderos positivos Falsos positivos



Verdaderos negativos Falsos negativos



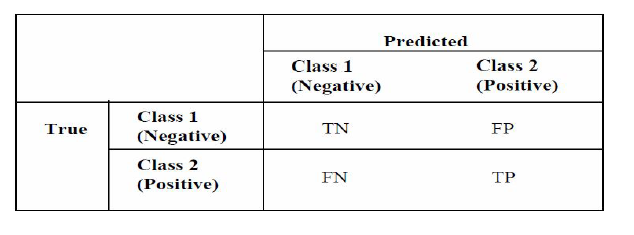
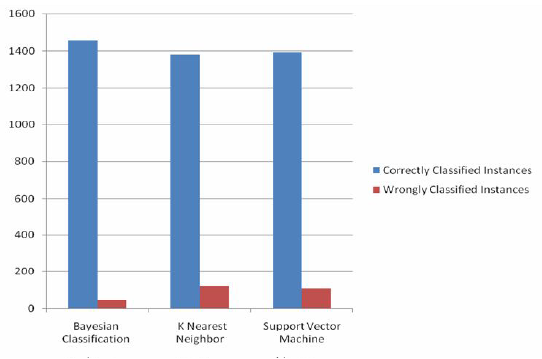


Tabla 4

Este estudio hace hincapié en la importancia del preprocesamiento de datos antes de ejecutar cualquier algoritmo: limpieza de datos, transformaciones, selecciones de características, etc.

Los resultados obtenidos para esta investigación (T.K. Das, 2015), son los siguientes:



Entre los tres clasificadores en el experimento, la clasificación de Bayes fue el más preciso y la peor fue el de VSM con un 88%. Así mismo, la clasificación SVM arroja la tasa de error más alta con un gran número de falsos positivos y falsos negativos en comparación con los demás.

“Esta técnica ayudaría al departamento de marketing a identificar a los encuestados para que fueran básicamente el objetivo de una actividad de campaña específica. También evita el gasto innecesario de enviar ofertas de promoción a los no encuestados”

Con base a estos resultados, podremos formular una hipótesis en la que desarrollando un algoritmo híbrido logremos una más alta precisión y una tasa de error mínima.

La siguiente investigación (Miranda, Jaime. 2005) nos dice como con un modelo predictivo de SVM, podemos identificar clientes con tendencias a la fuga de un banco utilizando variables de comportamiento bancario, sociodemográficas, de percepción de calidad del servicio y finalmente variables de entorno.

Esta investigación se enfocó en las fugas voluntarias, que se refieren cuando el cliente por propia cuenta desea abandonar su afiliación con el banco. Se observó que entre más transacciones realiza un cliente es menor la probabilidad de fuga de este.

Lo que se pretende es comparar el resultado del modelo actual con el resultado del modelo anterior que utiliza la empresa bancaria.

Dentro de las variables sociodemográficas implementa características como edad, renta, estado civil, nivel educacional, etc. que son atributos esenciales para identificar patrones de comportamiento de los clientes.

En su investigación (Miranda, Jaime. 2005) aplica transformaciones de datos categóricos a numéricos, como por ejemplo el nivel educacional, como se muestra en la tabla 5.

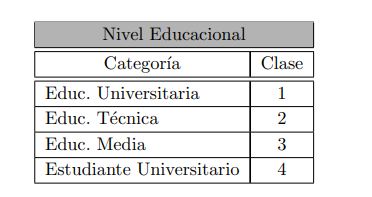


Tabla 5

Además, sugiere la creación de nuevas variables a partir de otras, eliminación de datos atípicos como edades mayores de 200 o saldos negativos y llenado de promedios y modas en datos faltantes, etc., y concluye el preprocesamiento estandarizando los datos entre 0 y 1debido a las variaciones de valores entre todos sus atributos.

Para el entrenamiento y obtención del modelo, utiliza como clasificador de Support Vector Machines, para encontrar un hiperplano que divida el espacio en dos regiones: clientes con fuga y clientes sin fuga, para este caso.

El resultado del modelo SVM se comparó con el que se tenía del modelo de Red Neuronal Multicapa (MLP), en donde se utilizaron las mismas variables y los mismos conjuntos de entrenamiento y prueba, arrojando lo mostrado en la tabla 6:

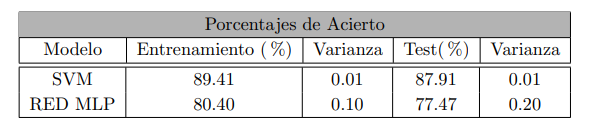


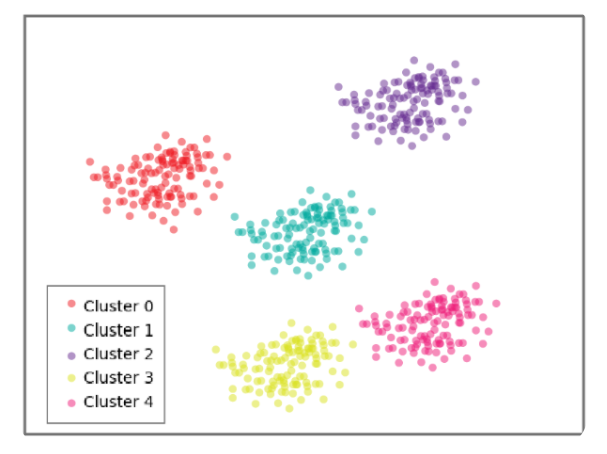
Tabla 6

“Cuanto más conozcan las empresas a sus clientes, mejor” (Romana, Chuprina. 2020).

En el siguiente análisis de la periodista, explica que el obtener la mayor de cantidad de información para conocer al cliente cada vez es más difícil y más aún si se utilizan métodos tradicionales para obtenerla y analizarla, sin embargo, con IA es mucho más efectivo.

El 75% de las empresas que utilizan IA y aprendizaje automático han aumentado hasta en 10% la satisfacción del cliente. (Forbes). Desafortunadamente solo el 37% de organizaciones minoristas están trabajando con IA y aprendizaje automático, debido a que estas organizaciones están acostumbradas a ser reactivas y necesitan un cambio cultural, comenzando por no considerar a la IA como algo complejo.

Predecir el comportamiento de un cliente basado en inteligencia artificial inicia con la segmentación (llamada clusterización en aprendizaje automático) que identifica datos unidos por algunos patrones similares o características que los agrupan.



Los datos demográficos de los clientes, como el sexo, la edad, los ingresos anuales y el puntaje de gastos, generalmente se tienen en cuenta para evaluar su similitud, pero un buen científico de datos puede detectar rasgos más ocultos para contribuir significativamente a la agrupación de clústeres. (Romana, Chuprina. 2020)

En su análisis, explica el proceso que se necesita para trabajar con el modelo predictivo:

* Limpieza de datos faltantes
* Imputación con valores medios o frecuentes
* Reducción de la dimensionalidad
* Extracción de nuevas características

En conclusión, de este análisis, se puede decir que es posible detectar grupos o clústeres de clientes en base a características similares por medio de un modelo que ayude a las organizaciones principalmente a regular las estrategias de marketing para diferentes clientes y ofrecerles descuentos y productos relevantes.

**Infografía con el contenido principal del proyecto**

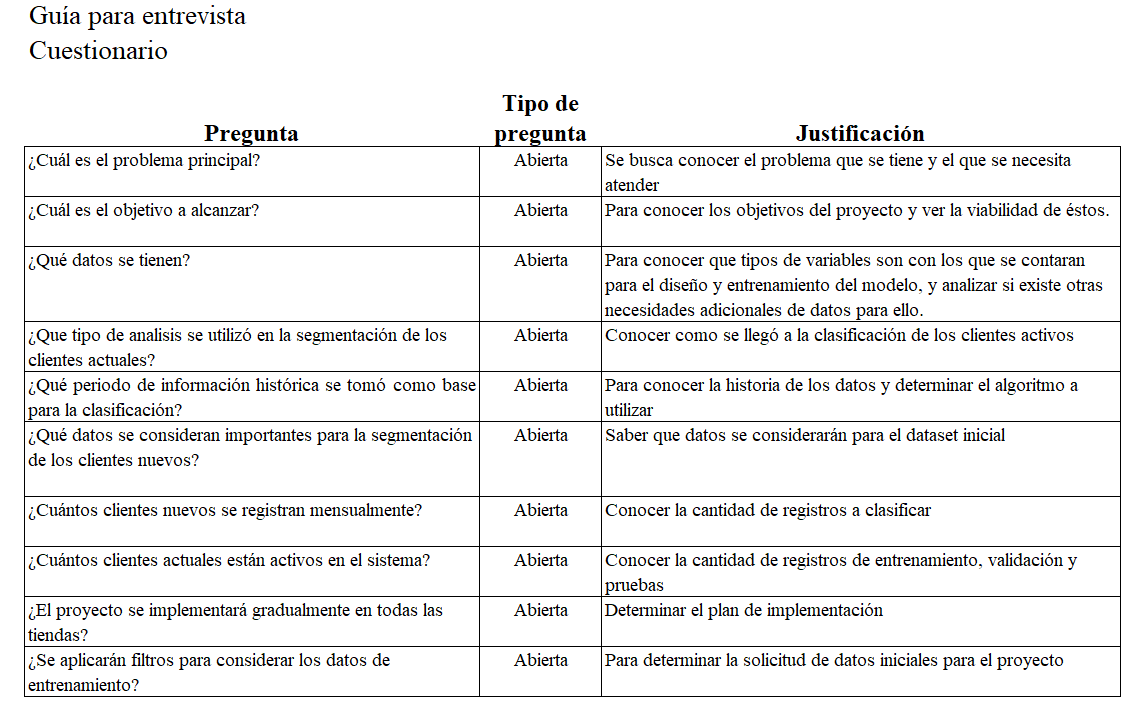
(Archivo anexo)

**Explicación, diseño y ejecución de técnicas de recolección de datos**

La recolección de datos para el proyecto fue desarrolla utilizando las siguientes técnicas:

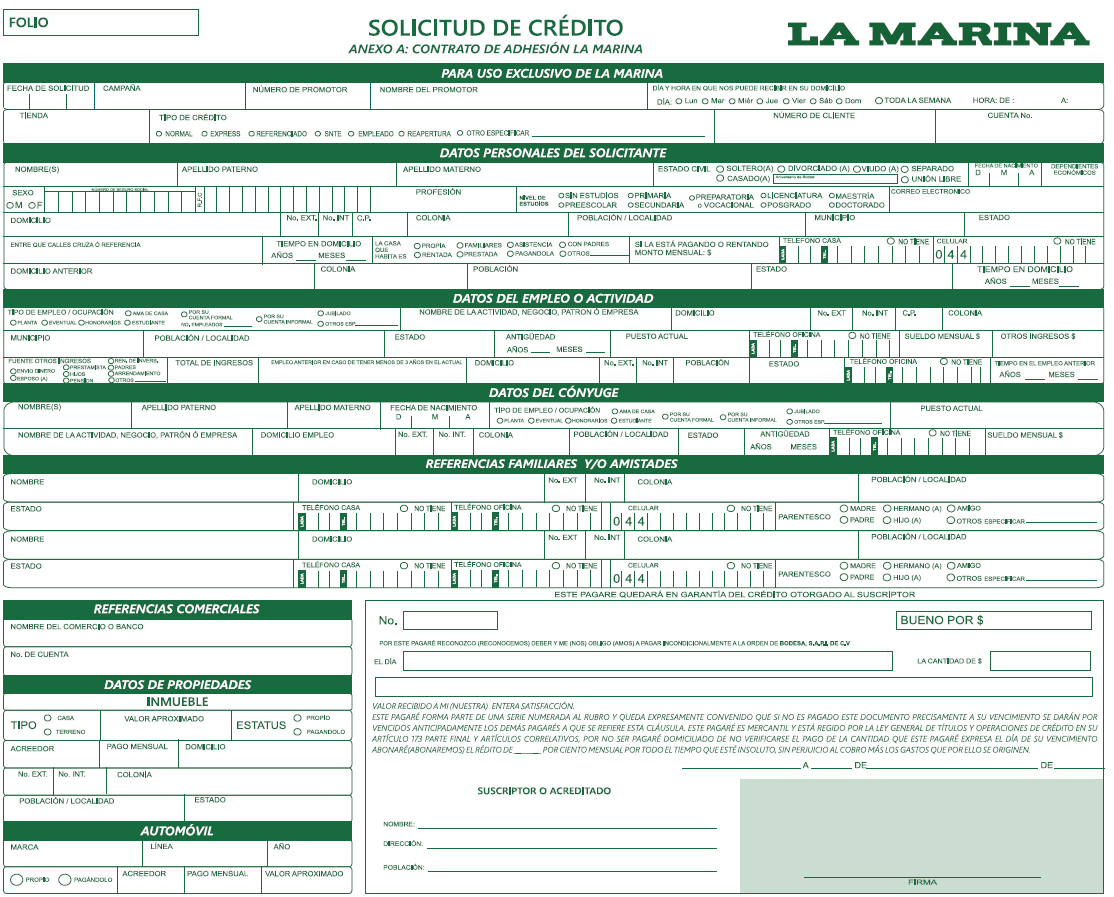
* Entrevistas presenciales
* Entrevistas virtuales
* Cuestionarios

Mediante varias entrevistas presenciales y virtuales con la subdirectora de inteligencia de clientes, la Lic. Brenda González Cortés y su equipo de trabajo, se logró conocer en una primera etapa el problema, los requisitos, los datos que se tienen y los objetivos a alcanzar, mediante la implementación de un pequeño cuestionario guía:



Así mismo una vez iniciado el proyecto se continuó con entrevistas abiertas presenciales y virtuales dentro de la etapa de ajustes y pruebas, finalizando con la revisión de resultados y precisión del modelo obtenido.

Otra técnica de recolección de datos para este proyecto son los cuestionarios, representados para el proyecto por solicitudes de crédito donde cada prospecto de cliente otorga sus datos personales, demográficos, laborales, crediticios y referencias.



En la primera entrevista se definió el problema que se tiene respecto a las campañas mal dirigidas por parte de la empresa; campañas que se realizan de manera general sin importar a que cliente se le enviará una promoción.

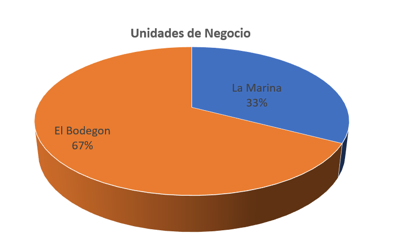
Así mismo, se recopiló información sobre el proceso para llegar a los segmentos de clientes que se establecieron por la misma área, este proceso se realizó con los datos de comportamiento de compra y pago de cada uno de los clientes actuales utilizando un análisis de segmentación RFM (Recency, Frecuency, Money).

Como segunda parte, se identificaron que datos de la solicitud de crédito nos proporciona información que contribuyan a la identificación de cada cliente en el sistema, y que datos se consideran importantes para el entrenamiento, validación y pruebas de los modelos, para posteriormente obtener estos mismos datos de los clientes nuevos y poder predecir las clases o segmentos.

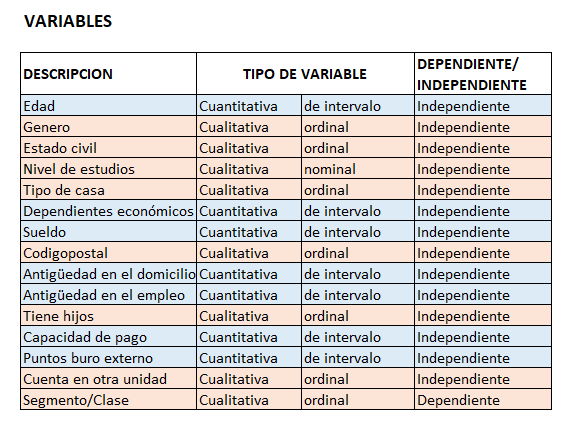
**Al menos una propuesta de experimento en desarrollo**

Con base a la recolección de datos y acercamiento al problema visto anteriormente, propongo hacer un primer experimento partiendo de una muestra de la información.

Como se mencionó en la justificación, la empresa se compone de dos unidades de negocio, y para esta primera experimentación se propone tomar los datos de las tiendas de La Marina que representan el 33% de la base de clientes; y de esta muestra solo tomar los datos de los clientes que pertenecen a las tiendas locales ubicadas en la ciudad de Colima, que representan un total de 36,116 registros.



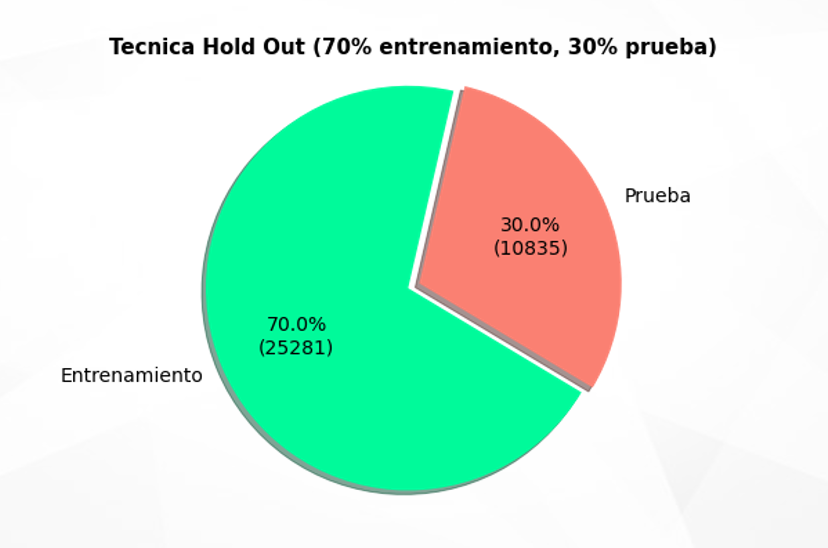
Después de un análisis de las variables que tenemos, se tomarán en cuenta las variables demográficas y crediticias siguientes,



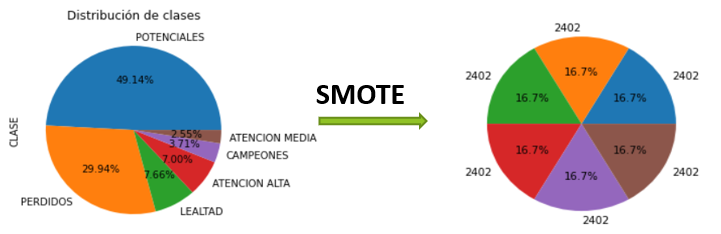
Parte de la propuesta en desarrollo consiste en analizar tres algoritmos de aprendizaje automático para comparar la eficiencia de cada uno y seleccionar el que mejor resultado nos arroje. Los algoritmos que se proponen son:

* Redes neuronales
* Regresión Logística
* Máquinas de vectores de soporte (SVM)

Para cada uno de los algoritmos se sugiere implementar datos de entrenamiento y prueba de acuerdo con la siguiente distribución:



Además de aplicar la técnica de sobre muestreo SMOTE para balancear los datos que se tienen en cada uno de los segmentos de los clientes.



Por último, se propone utilizar el lenguaje de programación de Python para el desarrollo y evaluación de los modelos.

**Referencias**

*Aplicaciones de Inteligencia Artificial para los negocios*. (2019, 30 julio). Brainvestigations. https://www.brainvestigations.com/inteligencia-artificial/aplicaciones-para-los-negocios/

*Aprendizaje automático: Cuándo usar cada método y técnica*. (2018, 23 octubre). CambioDigital. https://cambiodigital-ol.com/2018/10/aprendizaje-automatico-cuando-usar-cada-metodo-y-tecnica/

Chuprina, R. (2021, 28 abril). *Artificial Intelligence for Customer Behavior Analysis: A Practical Use Case*. Technology Partner for Innovative Companies. https://spd.group/artificial-intelligence/ai-for-customer-behavior-analysis/#Predicting\_customer\_behavior

Das, T. K. (2015, 31 octubre). *A customer classification prediction model based on machine learning techniques*. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7456903

García, L. A. (2014, 1 diciembre). *La ciencia y el hombre*. www.uv.mx. https://www.uv.mx/cienciahombre/revistae/vol17num3/articulos/inteligencia/index.htm

Gonzalo, A. (2020, 19 enero). *Tratamiento de clases desbalanceadas*. MachineLearningParaTodos.com. https://machinelearningparatodos.com/tratamiento-de-clases-desbalanceadas/

Maryani, I. (2018, 17 octubre). *Customer segmentation based on RFM model and clustering techniques with K-means algorithm*. IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/8780570

Miranda, J. (2005, 31 octubre). *Predicción de Fugas de Clientes para una Institución Financiera mediante Support Vector Machines*. Revista Ingenierıa de Sistemas. http://dii.uchile.cl/~ris/RISXIX/RISXIXpaper4.pdf

Serrano, S. (2020, 6 diciembre). *Guía de análisis por RFM: 6 segmentos clave para el RFM basado en Marketing*. Barilliance. https://www.barilliance.com/es/guia-de-analisis-por-rfm-6-segmentos-clave-para-el-rfm-basado-en-marketing/

Sethi, A. (2020, 25 junio). *One-Hot Encoding vs. Label Encoding using Scikit-Learn*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/

*Técnicas de inteligencia artificial: ¿cuáles son y para qué se usan?* (2021, 8 enero). APD España. https://www.apd.es/tecnicas-de-la-inteligencia-artificial-cuales-son-y-para-que-se-utilizan/