

# Evaluación del abandono de clientes de una compañía de telecomunicaciones por medio de cuatro modelos de aprendizaje maquina

Castro Rodríguez Jorge Octavio, Pérez Vázquez Ernesto

Universidad Panamericana, Facultad de Ingeniería,  
Augusto Rodin 498, Ciudad de México, 03920, México  
{0234588,0234686}@up.edu.mx

## Resumen

Debido a que los clientes son el activo maspreciado de las compañías, la retención de clientes es un objetivo esencial dentro de la estrategia de desarrollo de cualquier organización. En un entorno donde los servicios, productos y valor añadido están en constante evolución se vuelve cuestión de supervivencia contar con una estrategia de retención de clientes. Basados en información histórica existente en combinación con herramientas tecnológicas de procesamiento de información el aprendizaje de maquina (ML) nos permite realizar la extracción de conocimiento de una manera poderosa. Este proyecto implementó las siguientes técnicas de ML: Regresión logística, Random forest, SVM y XGBoost. Los resultados representan las características de los clientes que abandonaron. El negocio pueden identificar los abandonos que sucederán usando los resultados de las técnicas empleadas. Con base a dicha información la organización debe implementar estrategias de retención para aquellos clientes mas propensos a abandonar al negocio y cuyas características estén apegadas a los objetivos de desarrollo organizacional.

**Keywords:** abandono de clientes · regresión logística · SVM · Random forest · XGBoost.

## 1. Introducción

Haciendo énfasis en la base de que es mucho mas costoso captar un nuevo cliente que retener a los clientes actuales y el hecho que a periodos mas largos de permanencia los clientes generan mayores ganancias, Verbeke et al. (2011) afirma que la retención de clientes incrementa la rentabilidad.

Hoy día las compañías invierten grandes cantidades de recursos económicos, tecnológicos, humanos, estratégicos para aumentar el numero de clientes que se tienen registrados en la base de datos histórica del negocio, sin embargo no se tienen una estrategia clara y enfocada para retener a los clientes actuales pero

sobre todo retener a los clientes ideales (clientes que encajen en el tipo y estrategia de nuestro negocio).

con base a lo anterior es sumamente importante tener conocimiento de los siguientes puntos ya que en cualquier negocio donde los costos de captación son elevados y los márgenes de ganancia del producto o servicio son pequeños, la manera mas eficiente para generar rentabilidad es con base al consumo recurrente de los productos y/o servicios de la compañía, por tanto minimizar la tasa de abandono de los clientes adecuados se vuelve un pilar de supervivencia de las compañías.

- Conocer como se segmentan nuestros clientes para tener claridad de sus gustos, comportamiento e intereses y de esta manera enfocar la creación de nuevos productos y/o servicios de manera dirigida a captar dichas necesidades.
- Análisis de las acciones históricas que se han realizado para fidelizar a los clientes y a partir de ello identificar la efectividad de dichas estrategias y de esta manera enriquecer la base de información que alimente tanto a el conocimiento de nuestros clientes (segmentación) y de esta manera maximizar las opciones y niveles de retención de nuestros clientes.
- Mejora continua de nuestros servicios, productos, promociones, ofertas comunicados, etc., que genere un continuo y genuino interés en nuestra compañía y en lo que ofrece (Optimizar valor agregado de la compañía).

Con base a lo anterior, el objetivo establecido para este artículo es ejecutar una metodología de ciencia de datos que nos permita llevar a cabo la implementación de cuatro modelos distintos de aprendizaje automático, a partir de un set de datos históricos en común, y con base a los resultados de cada uno de los modelos realizar un análisis comparativo que nos permita finalmente determinar el modelo mas optimo.

Para llevar a cabo la implementación de los modelos elaboraremos para cada uno de ellos los siguientes pasos generales (en la sección numero cuatro de este documento se describe a mayor detalle las etapas de la metodología implementada):

- Recolección de datos (set de datos compartidos)
- Pre procesamiento
- Extracción de características
- Selección de características
- Creación de modelos
- Evaluación de modelos

Este documento se encuentra estructurado de la siguiente manera:

- **Trabajos relacionados.** En esta sección se observan los trabajos históricos de autores distintos que se tomaron como referencia base para el desarrollo de este artículo, dichos trabajos y autores fueron seleccionados debido a las similitudes con respecto a nuestro objetivo inicial (financiero-banca) como con nuestro objetivo final (Telecomunicaciones - considerando las limitaciones)
- **Limitaciones.** En este apartado se describen los elementos restrictivos para llevar el estudio presentado en este documento
- **Material y método.** En este bloque narra la propuesta del trabajo, los materiales utilizados y la metodología llevada a cabo
- **Experimentación.** Se describen los experimentos realizados y las métricas de evaluación
- **Resultados y discusión.** En este modulo se plasman los resultados obtenidos por cada uno de los modelos con la finalidad de abrir el debate dirigido a determinar el modelo mas adecuado para el set de datos utilizados
- **Conclusiones.** Sección que se utiliza para detallar las conclusiones del proyecto y consideraciones futuras del estudio.
- **Referencias.** En este apartado se encontraran las referencias bibliográficas que se utilizaron para el desarrollo de este artículo.

"La infidelidad siempre tiene como constantes la insatisfacción, desapego y desconfianza".

## 2. Trabajos relacionados

AbbasKeramati, HajarGhaneei and SeyedMohammadMirmohammadi (2016) se enfocan en la predicción de abandono de clientes en los servicios de banca electrónica utilizando el método de árbol de decisiones de aprendizaje maquina. Pavan Raj (2018) utiliza aprendizaje maquina para predecir el abandono de clientes de telecomunicaciones haciendo uso de KNN, random forest, Gaussian Naive Bayes, Support Vector Machine, LightGBMClassifier, XGBoost. Leon Palafox muestra el proceso para el entendimiento de negocio perfilado al abandono de cliente, entendimiento de datos y su exploración, Guo-en and Wei-dong (2008) se enfocan en construir un modelo de predicción de abandono de clientes usando SVM en la industria de telecomunicaciones. De manera tabular los podemos resumir en la figura 1:

Investigador y/o año	Industria	Metodo(s)
AbbasKeramati, HajarGhaneei and SeyedMohammadMirmohammadi (2016)	Banca	Árbol de decisión
Pavan Raj (2018)	Telecomunicaciones	KNN, random forest, Gaussian Naive Bayes, Support Vector Machine, LightGBMClassifier, XGBoost
Leon Palafox	Telecomunicaciones	Análisis exploratorio

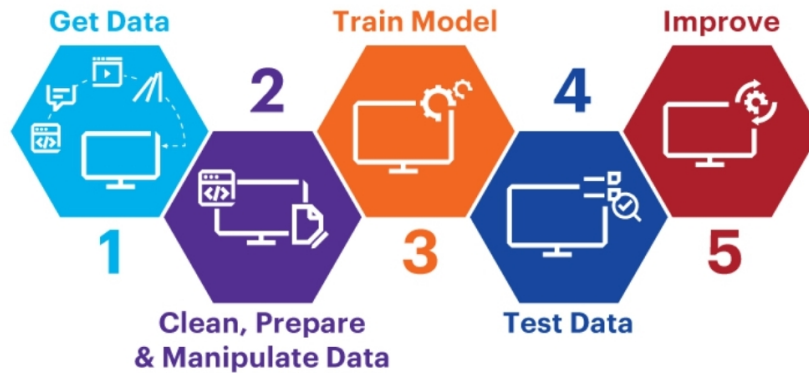
**Figura 1.** Trabajos relacionados.

### 3. Limitaciones

Inicialmente se tenía planeado realizar el ejercicio con una base de datos de Nacional Monte de Piedad sin embargo no fue posible debido a las restricciones de la institución para compartir información productiva lo que represento un limitante respecto una limitante para este estudio, por lo que se opto por trabajar con información pública del sitio Kaggle, La investigación futura tendrá como objetivo realzar este ejercicio con información productiva esperando lograr superar los problemas de privacidad asociados, de igual manera se planea programar un despliegue el cual no podremos ejecutar debido al corto tiempo con el cual se conto para obtener las autorizaciones correspondientes.

### 4. Materiales y método

El estudio se llevo a cabo conforme al esquema denominado CRISPDM, (en la figura 2 podemos observar el flujo general utilizado):



**Figura 2.** Flujo de trabajo.

#### ■ Entendimiento del Negocio

El entendimiento del negocio debe partir del conocimiento de los objetivos estratégicos de las organizaciones ya que a partir de dicho conocimiento se pueden establecer líneas base de observación y/o acción. Debido a las limitantes de el set de datos se entiende que este apartado queda pendiente para futuros ejercicios, sin embargo cabe recalcar que con base al entendimiento de la metodología empleada este estudio podrá ser ejecutado con mayor exactitud y experiencia en cuanto se logre obtener información productiva.

#### ■ Entendimiento de los Datos

Como parte de la exploración de datos descubrimos comportamientos valiosos para la toma de decisiones, por ejemplo figura 3:

- El abandono temprano de clientes jóvenes, figura 3
- Las personas que pagan menos permanecen por mayor tiempo
- Las personas que cuentan con el servicio de telefonía tienden a tener una taza alta de abandono
- Las personas que utilizan fibra óptica tienden a presentar una taza alta de abandono
- El abandono no se relaciona con el genero del cliente, figura 3
- Al finalizar el primer mes de servicio es donde se nota la mayor tasa de abandono
- La correlación de las variables pertenecientes al set de datos, figura 3
- Con forme los clientes permanecen mayor tiempo con el servicio la tasa de abandono disminuye

#### ■ Preparación de los Datos

La información se preparo mediante a python ya que encontramos casos en los cuales tuvimos que eliminar espacios (TotalCharges) o eliminar valores vacíos (TotalCharges), se opto por excluir 11 observaciones por contener en una columna valores vacíos, normalizamos los valores, excluimos las variables gender y customerID ya que con base al análisis exploratorio observamos que no tienen valor representativo, como se muestra en la figura 3.

Para efectos de este estudio el set de datos fue dividido en dos subsets de datos, el primero de ellos contiene el 70 porciento de las observaciones y fue reservado para la etapa de entrenamiento, el 30 porciento de la información fue utilizada para la etapa de evaluación.

#### ■ Modelado

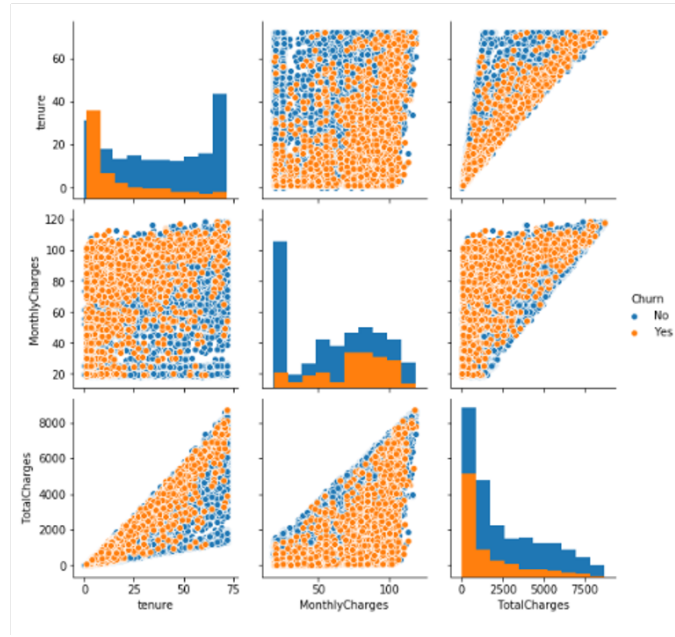
Se detalla en el apartado 5 (Experimentación) de este documento.

#### ■ Evaluación

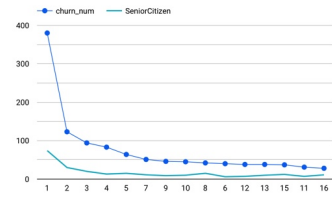
Se detalla en el apartado 5 (Experimentación) de este documento.

#### ■ Despliegue

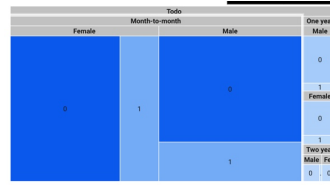
(Se observa en la seccion de limitantes)



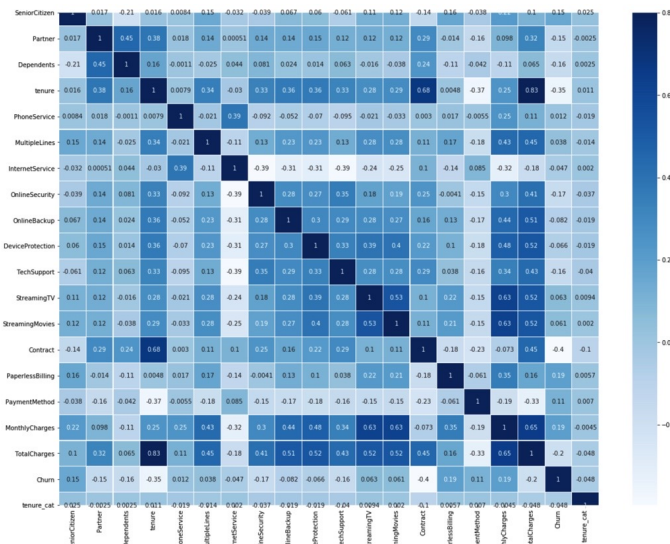
(a) Análisis exploratorio.



(b) Abandono por mes por Seniority.



(c) Relación de abandono por sexo.

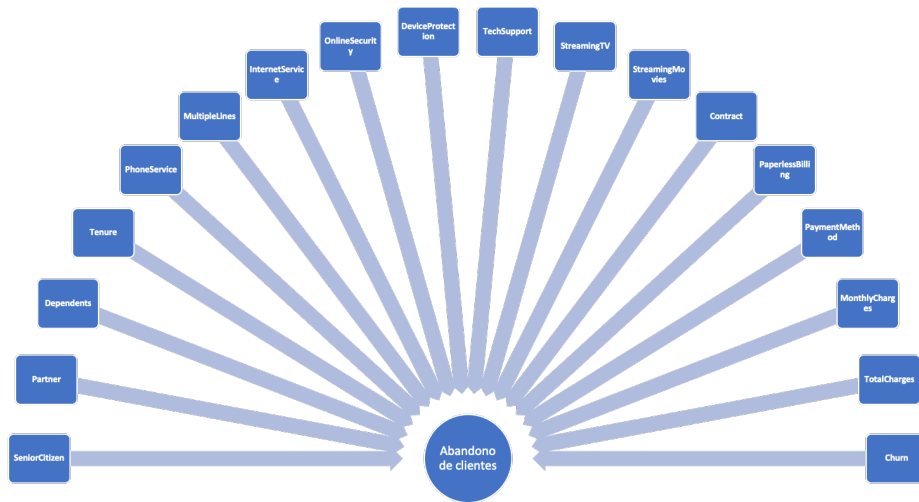


(d) Matriz de correlación.

Figura 3. Análisis Exploratorio

#### 4.1. Descripción General

Se programará la ejecución de cuatro modelos de aprendizaje maquina con el objetivo de seleccionar el modelo que nos proporcione la alternativa mas optima para apoyar a la toma de decisiones, en la figura 4 se puede observar las variables utilizadas para evaluar los cuatro modelos para el caso de estudio de abandono de clientes.



**Figura 4.** Mapa conceptual para un modelo de abandono de clientes.

#### 4.2. Materiales

Para llevar a cabo el caso de estudio utilizamos las siguientes herramientas:

- Python v3
- Google Colaboratory
- Google data studio
- Anaconda
- Microsoft Excel

El resumen del set de datos utilizado se describe en la figura 5:

Artefacto	Descripción
Total de observaciones analizadas	7043
Formato del archivo	csv
Columnas totales	21
Duplicados customerID	0
Columnas homologadas (descripciones)	7
Edición de tipo de dato de columna existente (TotalCharges, String to Int)	1
Variables nulas	11 (TotalCharges)
Entendimiento de columna SeniorCitizen	1=Sr (mayor), 2=Jr (joven)
Entendimiento de columna tenure	Meses de permanencia
Entendimiento de columna PhoneService	Telefonía móvil
Total de variables	21
customerID (string)	7043
Gender (category)	2 tipos
SeniorCitizen (integer)	2 tipos
Partner (category)	2 tipos
Dependents (category)	2 tipos
Tenure (integer)	73 tipos
PhoneService (category)	2 tipos
MultipleLines (category)	3 tipos
InternetService (category)	3 tipos
OnlineSecurity (category)	3 tipos
OnlineBackup (category)	3 tipos
DeviceProtection (category)	3 tipos
TechSupport (category)	3 tipos
StreamingTV (category)	3 tipos
StreamingMovies (category)	3 tipos
Contract (category)	3 tipos
PaperlessBilling (category)	2 tipos
PaymentMethod (category)	4 tipos
MonthlyCharges (integer)	1585 tipos
TotalCharges (integer)	6531 tipos
Churn (category)	2 tipos

Figura 5. Resumen del set de datos.

#### 4.3. Métodos de aprendizaje

Los métodos de aprendizaje máquina utilizados en este trabajo se describen a continuación.

**Regresión Logística** Chitarroni, H. (2002) define la regresión logística como un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas.



- **Ventajas:** Fácil, rápido y simple método de clasificación. Los Parámetros Theta explican la dirección e importancia de la variable independiente sobre la variable dependiente. Puede ser utilizada en clasificaciones multiclase. La función de pérdida siempre es convexa.
- **Desventajas:** No se puede aplicar en problemas de clasificación no lineal. Se requiere una selección adecuada de características. La colinealidad y los valores atípicos alteran la precisión del modelo.

**Máquinas de soporte vectorial** Betancourt, G. A. (2005) explica que una Máquina de Soporte Vectorial (SVM) aprende la superficie de decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento.

- **Ventajas:** SVM utiliza el truco del kernel para resolver soluciones complejas. SVM utiliza una función de optimización convexa, debido a que los mínimos globales siempre son alcanzables. El concepto de “Hinge loss” proporciona mayor precisión. Los valores atípicos se pueden manejar bien utilizando suavización sobre la constante C.
- **Desventajas:** El concepto de “Hinge loss” conduce a la escasez (matriz poblada principalmente con ceros). Los hiperparámetros y los kernel deben ajustarse cuidadosamente para una precisión suficiente. Mayor tiempo de entrenamiento para conjuntos de datos más grandes.

**Random Forest** Random Forest es una combinación de árboles de decisión, tal que cada árbol depende de los valores de un vector aleatorio, independiente y con la misma distribución para cada uno de estos Friedman, J., Hastie, T., and Tibshirani, R. (2008). Todos los árboles tienen la misma distribución en el bosque (forest), pero son forzados a ser diferentes. Esto reduce la correlación.

Existen diversas técnicas de construcción de conjuntos de clasificadores. En muchas ocasiones, la forma de llevar a cabo estas técnicas es a partir de un conjunto de números aleatorios:

- En bagging Breiman, L. (1996a), cada clasificador se construye a partir de una muestra bootstrap, la cual se genera utilizando tantos números aleatorios como elementos tenga el conjunto de entrenamiento.
- En random subspace Ho, T. K. (1998), cada clasificador base utiliza sólo un subconjunto de atributos seleccionado aleatoriamente de entre el total de las variables. De esta forma, cada clasificador se restringe a un subespacio aleatorio de atributos.

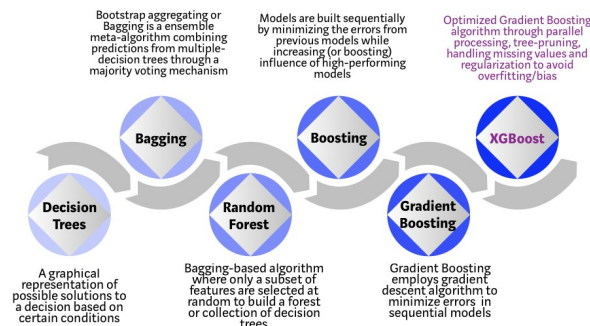
- Las técnicas output flipping y class switching, Breiman, L. (2000); Martínez-Muñoz and Suárez (2005), se basan en la manipulación aleatoria de las etiquetas de clase, por lo que necesitan de números aleatorios para seleccionar los datos cuyas etiquetas serán cambiadas.
- Randomization Dietterich, T. G. (2000b), introduce aleatoriedad en el algoritmo de aprendizaje construyendo conjuntos de clasificadores con árboles de decisión en los que el valor de corte es seleccionado aleatoriamente entre los F mejores cortes posibles.

A continuación se describen las ventajas y desventajas mas representativas del uso de este modelo:

- Ventajas: Modelo preciso y potente. Maneja el sobreajuste de manera eficiente. Admite la selección implícita de funciones y deriva la importancia de las funciones.
- Desventajas Computacionalmente complejo y más lento cuando el bosque se vuelve grande. No es un modelo bien descriptivo sobre la predicción.

Si el clasificador base utilizado es un árbol de decisión, el concepto random forest Breiman, L. (2001) engloba todas estas técnicas.

**XGBoost** Extreme Gradient Boosting (XGBoost), es una técnica de clasificación supervisada que utiliza un conjunto de árboles de decisión. XGBoost es un algoritmo de Machine Learning basado en un árbol de decisiones que utiliza un marco de impulso de gradiente. En los problemas de predicción que involucran datos no estructurados (imágenes, texto, etc.), las redes neuronales artificiales tienden a superar a todos los demás algoritmos o marcos. Sin embargo, cuando se trata de datos estructurados / tabulares de pequeño a mediano, los algoritmos basados en el árbol de decisión se consideran los mejores en su clase en este momento. La figura 6 muestra la evolución de los algoritmos basados en árboles a lo largo de los años.



**Figura 6.** Evolución del algoritmo XGBoost.

- **Ventajas:** Dado que los árboles potenciados se derivan de la optimización de una función objetivo, básicamente los “Gradient Boosting Model” se pueden utilizar para resolver casi todas las funciones objetivas que podemos escribir en gradientes.
- **Desventajas:** Los “Gradient Boosting Model” son más sensibles al sobreajuste si los datos son “ruidosos”. La capacitación generalmente lleva más tiempo debido al hecho de que los árboles se construyen secuencialmente. Los “Gradient Boosting Model” son más difíciles de afinar que los “Random Forest”. Generalmente hay tres parámetros: número de árboles, profundidad de árboles y tasa de aprendizaje, y cada árbol construido es generalmente poco profundo.

## 5. Experimentación

En esta sección contrastaremos el desempeño de los modelos:

- Regresión Logística
- Random Forest
- XGBoost
- Máquinas de Soporte Vectorial

Para generar los modelos realizamos una división del set de datos en el cual 70 por ciento de las observaciones se usaron para entrenar el modelo y el 30 por ciento restante para probar el modelo, adicional se hizo uso de la técnica de validación por cruce de k-pliegue para mejorar el desempeño de los modelos, el valor de K se definió en 10 (pliegues).

Utilizamos los indicadores de desempeño descritos en la figura 7 para evaluar y contrastar el desempeño entre los modelos y de esta manera definir nuestra recomendación del modelo mas optimo.

**Accuracy:** Mide el porcentaje de casos que el modelo ha acertado.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

**Precision:** Mide la calidad del modelo.

$$Precision = \frac{TP}{(TP + FP)}$$

**Recall:** Informa sobre la cantidad que el modelo es capaz de identificar.

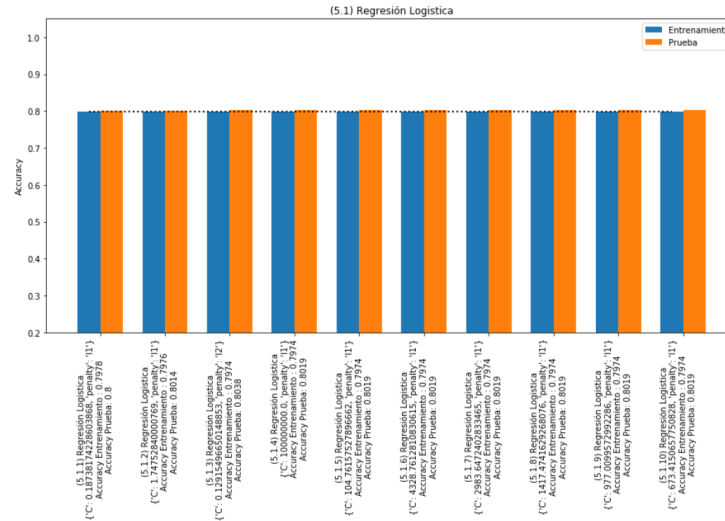
$$Recall = \frac{TP}{(TP + FN)}$$

**F1-Score:** Compara el rendimiento combinado de la precisión y la exhaustividad.

$$F1Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

**Figura 7.** Interpretación de indicadores.

Se realizó un ranking de desempeño de los hiperparámetros utilizados por modelo con la finalidad de definir los valores óptimos de los hiperparámetros a fijar por modelos, en la figura 8 se muestra un ejemplo de la evaluación de hiperparámetros realizada a la regresión logística de este estudio (las figuras de los modelos restantes se pueden observar en el libro de trabajo que contiene el código que soporta este artículo).



**Figura 8.** Ejemplo de comparativa de desempeño de hiperparámetro (RL).

En la figura 9 se aprecia el listado de hiperparámetros seleccionados para utilizar por cada uno de los modelos.

Modelo	Hiperparámetro
Maquinas de Soporte Vectorial - SVM	{'C': 4, 'gamma': 0.001, 'kernel': 'rbf'}
Regresión Logística	{'C': 0.12915496650148853, 'penalty': 'l2'}
Random Forest	'bootstrap': [True], 'max_depth': [100, 110], 'max_features': ['auto'], 'min_samples_leaf': [3, 4], 'min_samples_split': [8, 10], 'n_estimators': [100, 500, 1000]
XGBoost	'learning_rate': [0.15, 0.20], 'max_depth': [5, 6, 7, 8], 'gamma': [0.4, 0.6], 'colsample_bytree': [0.4], 'n_estimators': [200], 'objective': ['binary:logistic']

**Figura 9.** Hiperparámetro por modelo.

## 6. Resultados y discusión

**Resultados** En las figuras 10 se observan los resultados obtenidos tras haber evaluado los modelos enunciados en este estudio y el ranking que presento cada uno de ellos (evaluados con distintos valores en hiperparametros).

Ranking	Modelo	params_cv	Accuracy_Train	Accuracy_test	Recall_test	Precision_test	F1_test
1	(5.4.1) Maquinas de Soporte Vectorial - SVM	{'C': 4, 'gamma': 0.001, 'kernel': 'rbf'}	0.7972	0.8043	0.4957	0.7034	0.5816
2	(5.1.3) Regresión Logística	{'C': 0.12915496650148853, 'penalty': 'l2'}	0.7974	0.8038	0.5354	0.6813	0.5996
3	(5.4.6) Maquinas de Soporte Vectorial - SVM	{'C': 8, 'gamma': 0.001, 'kernel': 'sigmoid'}	0.7962	0.8033	0.4974	0.6990	0.5812
4	(5.4.3) Maquinas de Soporte Vectorial - SVM	{'C': 7, 'gamma': 0.001, 'kernel': 'sigmoid'}	0.7966	0.8024	0.4853	0.7025	0.5741
5	(5.1.4) Regresión Logística	{'C': 100000000.0, 'penalty': 'l1'}	0.7974	0.8019	0.5354	0.6754	0.5973
6	(5.1.5) Regresión Logística	{'C': 104.76157527896662, 'penalty': 'l1'}	0.7974	0.8019	0.5354	0.6754	0.5973
7	(5.2.9) Random Forest	{'bootstrap': True, 'max_depth': 110, 'max_fea...	0.8003	0.8000	0.5095	0.6813	0.5830
8	(5.2.10) Random Forest	{'bootstrap': True, 'max_depth': 100, 'max_fea...	0.8003	0.8000	0.5095	0.6813	0.5830
9	(5.2.1) Random Forest	{'bootstrap': True, 'max_depth': 110, 'max_fea...	0.8009	0.7995	0.5130	0.6781	0.5841
10	(5.3.3) XGBoost	{'colsample_bytree': 0.4, 'gamma': 0.4, 'learn...	0.7926	0.7976	0.5181	0.6696	0.5842
11	(5.3.7) XGBoost	{'colsample_bytree': 0.4, 'gamma': 0.6, 'learn...	0.7899	0.7943	0.5250	0.6566	0.5835
12	(5.3.1) XGBoost	{'colsample_bytree': 0.4, 'gamma': 0.4, 'learn...	0.7936	0.7910	0.5147	0.6507	0.5747

**Figura 10.** Ranking de evaluacion de performance por modelo .

Adicional cabe destacar que utilizamos RFE para verificar si el desempeño de los modelos mejoraba sin embargo no notamos diferencia por lo que la ejecución final no concidera RFE (en el libro de trabajo del documento se aprecia el código ejecutado a para esta puntualización).

**Discusión** Con base a la figura 10, podemos argumentar las siguientes observaciones:

El mejor indicador de accuracy lo tiene el modelo (5.4.1) de Maquinas de Soporte Vectorial - SVM, y en un segundo lugar encontramos que el modelo (5.1.3) de regresión logística, on base a lo anterior podríamos concluir fría y numéricamente que la opción a utilizar seria la implementación de un modelo de maquina de soporte sectorial SVM, sin embargo si consideramos el uso computacional que conlleva

**evaluar los modelos, la complejidad de los mismos y la diferencia marginal de resultados (accuracy), nuestra recomendación para este caso de estudio es utilizar el modelo de regresión logística (considerando las limitantes y el uso de hiperparametros recomendado).**

Con base a las condicionantes del set de datos, limitantes del entorno y el tipo de modelos seleccionados al iniciar este estudio es necesario considerar que para determinar un modelo optimo no solo se debe contemplar la precisión demostrada por los modelos, dicha decisión debería considerar algunos de los siguientes puntos:

- Tiempo de implementación
- Infraestructura con la que cuenta la compañía para evaluar un modelo
- Infraestructura con la que cuenta la compañía para desplegar y operar el modelo
- Conocimiento del caso de negocio y el objetivo del mismo
- Conocimiento de interpretación de los datos
- Entorno
- Madurez del equipo que desarrolla los modelos

## 7. Conclusiones

Para este caso de estudio seguimos los pasos de la metodología CRISP para evaluar el abandono de clientes de una empresa de telecomunicaciones, teniendo como objetivo realizar un análisis comparativo entre distintos modelos de aprendizaje maquina que nos permita determinar el modelo mas optimo para el set de datos utilizado.

Si bien se siguió la metodología CRISP es necesario puntualizar nuevamente que no se logro implementar el paso final de despliegue de los modelos, debido a las limitantes ya descritas en este documento, lo que implica que no tenemos parámetros tangibles para observar el desempeño productivo de dichos modelos y por consiguiente dar el seguimiento correspondiente para realizar los ajustes necesarios a los modelos, set de datos, condiciones de negocio, modificación de hiperparametros, etc.

## Referencias

1. AbbasKeramati, HajarGhaneei and SeyedMohammadMirmohammadi (2016). Developing a prediction model for customer churn from electronic banking services using data mining <https://link.springer.com/content/pdf/10.1186>
  2. Betancourt, G. A. (2005). Las máquinas de soporte vectorial (SVMs). *Scientia et technica*, 1(27). <http://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895/4139>
  3. Breiman, L. (1996a). Bagging Predictors. *Machine Learning*, 24(2):123–140
  4. Breiman, L. (2000). Randomizing Outputs to Increase Prediction Accuracy. *Machine Learning*, 40(3):229–242.
  5. Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
  6. Carranza, J. P., Salomón, M. J., Piumetto, M. A., Monzani, F., Montenegro, M. G., and Córdoba, M. A. (2018). RANDOM FOREST COMO TÉCNICA DE VALUACIÓN MASIVA DEL VALOR DEL SUELO URBANO: UNA APLICACIÓN PARA LA CIUDAD DE RÍO CUARTO, CÓRDOBA, ARGENTINA.
  7. Chitarroni, H. (2002). La regresión logística. <https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf>
  8. Dietterich, T. G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157.
  9. Friedman, J., Hastie, T., and Tibshirani, R. (2008). Response to Mease and Wyner, Evidence Contrary to the Statistical View of Boosting, *JMLR* 9: 131–156, 2008. *Journal of Machine Learning Research*, 9, 175–180.
  10. Guo-en X, Wei-dong J (2008). Model of customer churn prediction on support vector machine. *Syst Eng Theory Pract* 28(1):71–77
  11. Ho, T. K. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844.
  12. Martínez-Muñoz, G. and Suárez, A. (2005). Switching class labels to generate classification ensembles. *Pattern Recognition*, 38(10):1483–1494.
  13. Pavan Raj (2018). Telecom Customer Churn Prediction <https://www.kaggle.com/pavanraj159/telecom-customer-churn-prediction>
  14. Verbeke W, Martens D, Mues C, Baesens B (2011) Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Syst Appl* 38:2354–2364
  15. Vishal Morde (2019). XGBoost Algorithm: Long May She Reign!
  16. Código fuente del estudio realizado:
- Ingresar-> Notebook Python**
17. Set de datos
- Ingresar-> Archivo csv**
18. Exploratorio del set de datos
- Ingresar-> Análisis numero dos realizado con la herramienta google studio**