PERFILADO DE CLIENTES POR HÁBITO DE PAGO DISCRIMINADO POR REGIONES

Daian Paola Fajardo Becerra
dpfajardob@eafit.edu.co

Juan Carlos Agudelo Acevedo
jcagudeloa@eafit.edu.co

Hernán Sepúlveda Jiménez
hsepulvedj@eafit.edu.co

Juan David Sanz Ramírez
jdsanzr@eafit.edu.co

MAESTRÍA EN CIENCIA DE DATOS Y ANALÍTICA

Proyecto Integrador

Mayo 2020

Contenido

1. Metodología:

1.1. CRISP-DM

La metodología empleada será CRISP-DM *(Cross Industry Standard for Data Mining*) que consiste en una forma de estructurar el trabajo de minería de datos y que consta de seis fases o pasos. Dicha metodología se muestra en el siguiente diagrama (figura 1) donde queda evidente que no es una estructura rígida, permitiendo devolverse para revisar y realizar posibles ajustes; además de permitir avanzar en diferentes frentes.

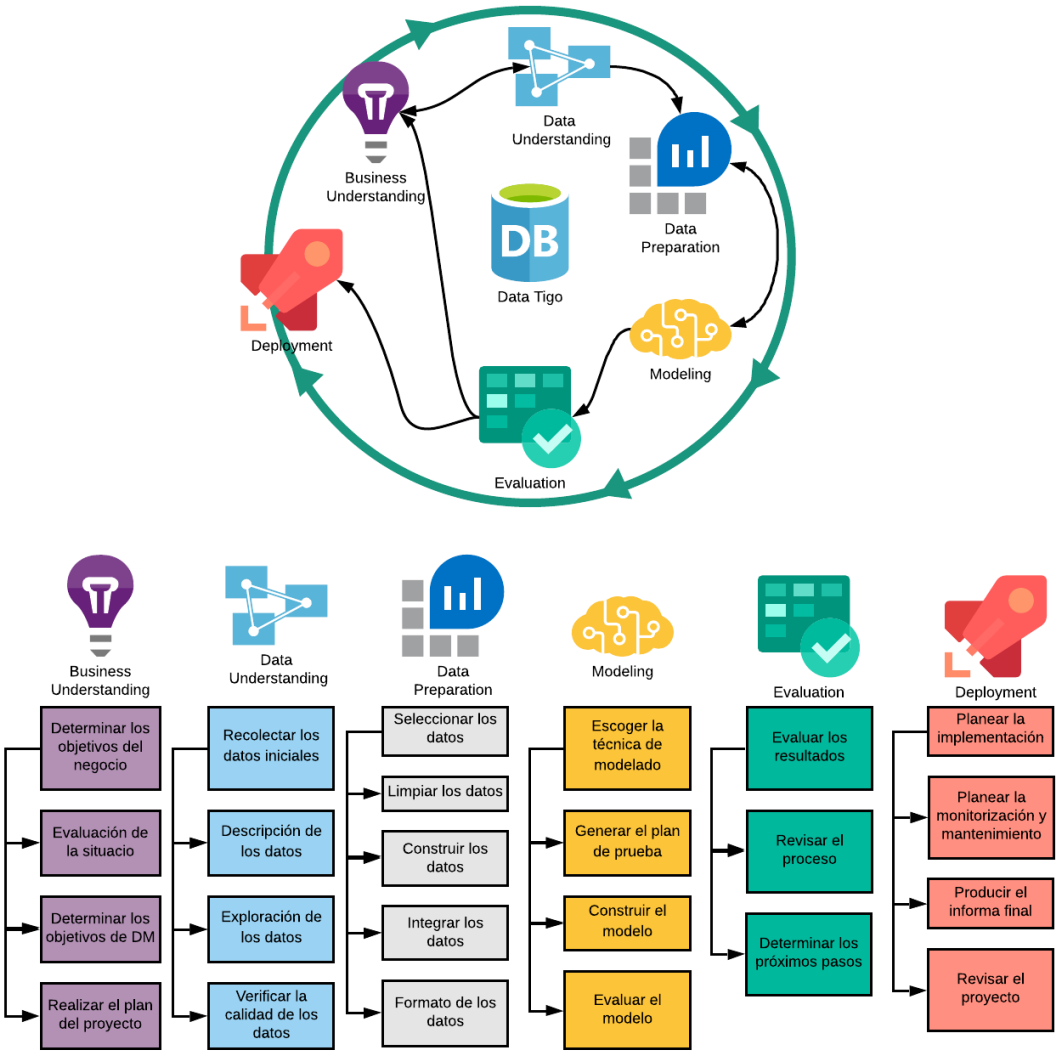


Figure 1: Modelo de CRISP-DM

2. Business understanding:

2.1. Tigo:

Es una empresa de telecomunicaciones colombiana creada en 2006, propiedad del Grupo EPM y de Millicom International Cellular S.A., que ofrece sus servicios en el ámbito nacional e internacional por medio de Colombia Móvil S.A. bajo la marca Tigo y bajo la marca Orbitel en Canadá, Estados Unidos y España.

Tigo presta servicios de internet y telefonía, considerados para las empresas y hogares como algo esencial y que representa el 22,3% del mercado de las telecomunicaciones del país; posicionándolo como el segundo operador más grande.

Por tanto, es importante conocer el comportamiento de pago de los clientes y las razones que pueden tener para justificar dicho comportamiento.

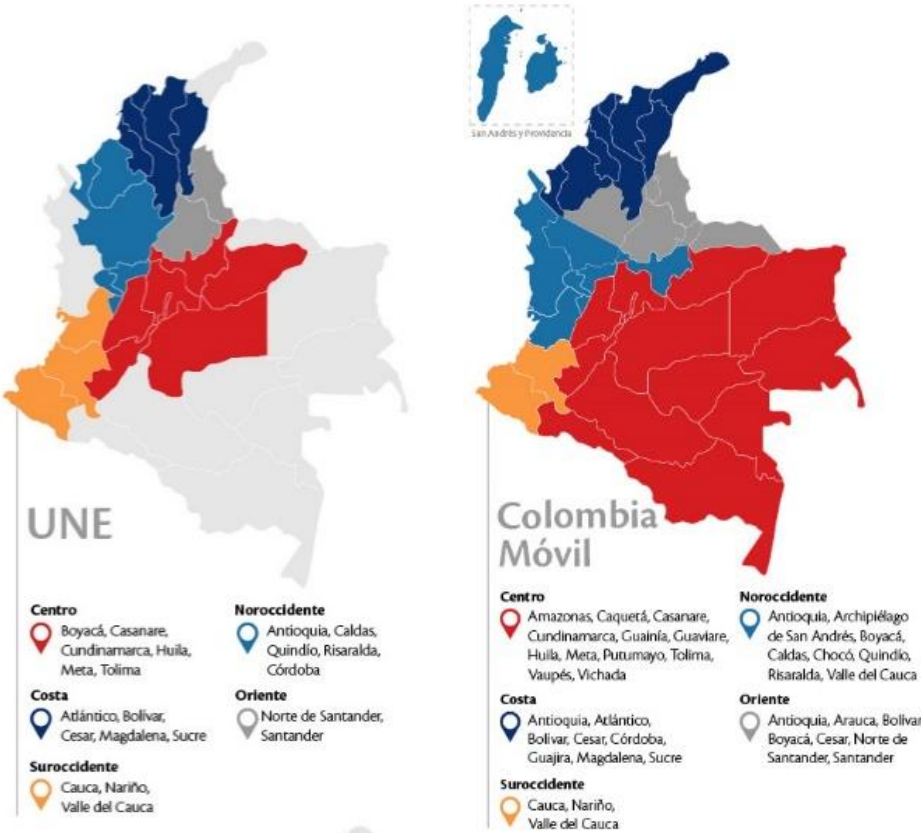


Figure 2: Gráfico de zonas

2.2. Evaluación de la situación:

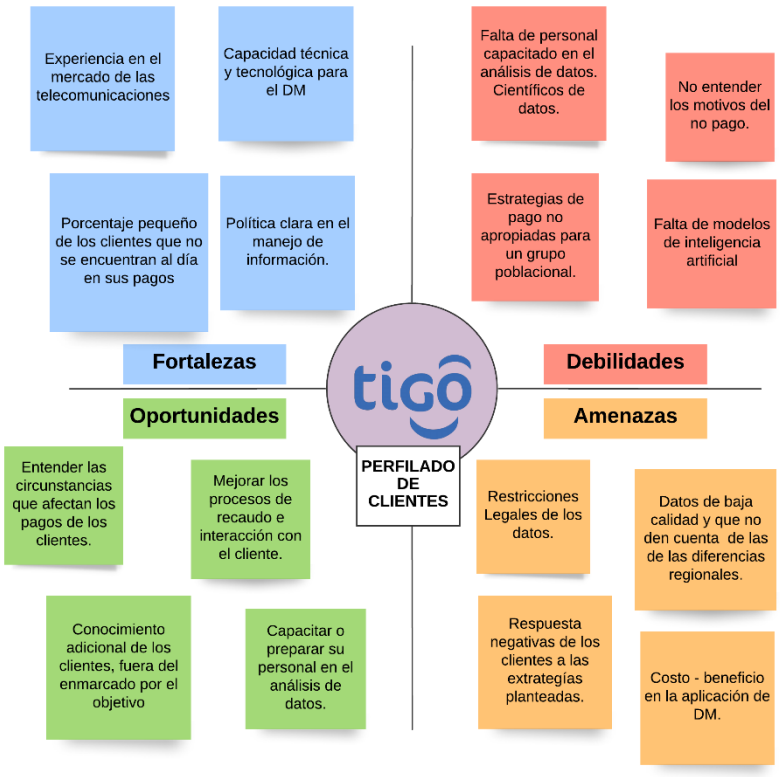


Figure 3: DOFA

2.3. Definición de problema:

Actualmente, Tigo tiene un reconocimiento básico de los usuarios respecto a sus comportamientos de pago, donde este solo se realiza por los días de mora de la cartera, sin embargo, se ha empezado a observar que estos comportamientos van más allá del pago, viéndose implicadas variables como las zonas donde se presta el servicio, variables demográficas de cliente, entre otras.

Por lo que se quiere solucionar las siguientes preguntas dentro de este proyecto:

- ¿Existe alguna relación entre los atributos de cliente con su comportamiento de pago?
- ¿Se puede pronosticar el comportamiento de pago de un cliente nuevo de acuerdo con la data histórica de clientes existentes?
- ¿Cuáles son los temas más relevantes encontrados en las quejas de los clientes de Tigo?

2.4. ¿Cómo lo solucionaremos?

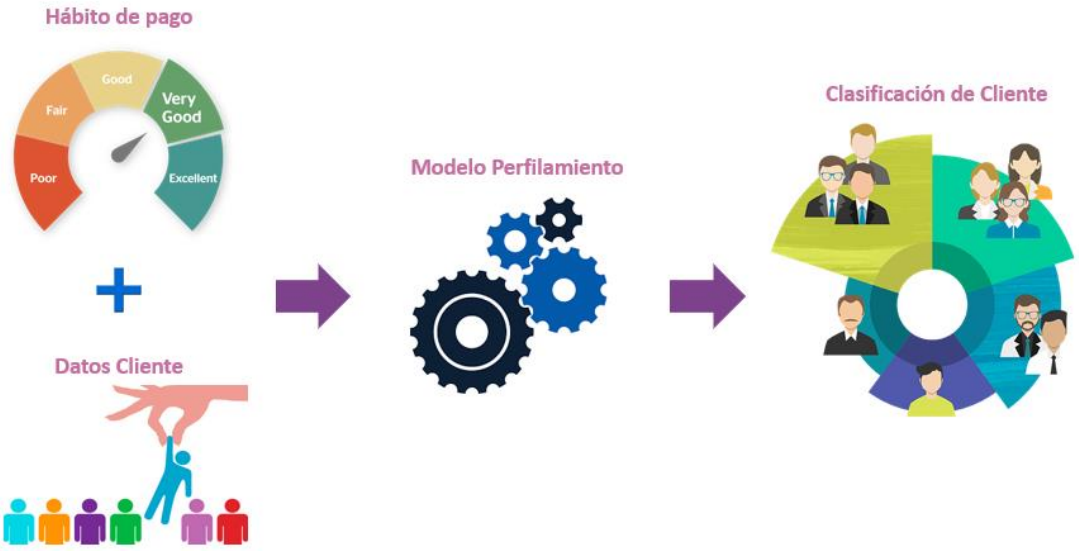


Figure 4:Clasificación de cliente

3. Fuente de Datos

Existen 4 grupos de fuentes de datos, que son:

1. Facturadores:
   Son bases de datos transaccionales bajo tecnología Oracle, que contiene la información de los valores facturados que tienen una frecuencia cíclica, (Un ciclo se compone de 30 a 31 días dependiendo del mes) y los valores pagados que tienen una frecuencia de carga diaria.

2. CRM:
   Contiene la información característica del cliente.

3. Fuentes de datos de analíticos:
   De aquí se toma la información ya procesada del cliente.

4. SOX:
   Se recopila toda la información y se procesa de tal manera que se tiene la información necesaria y en el formato indicado.

   Toda la información es recopilada y se almacena en un solo servidor, con el fin de ser tratada y visualizada por los diferentes departamentos de la compañía, figura 5 y 6.

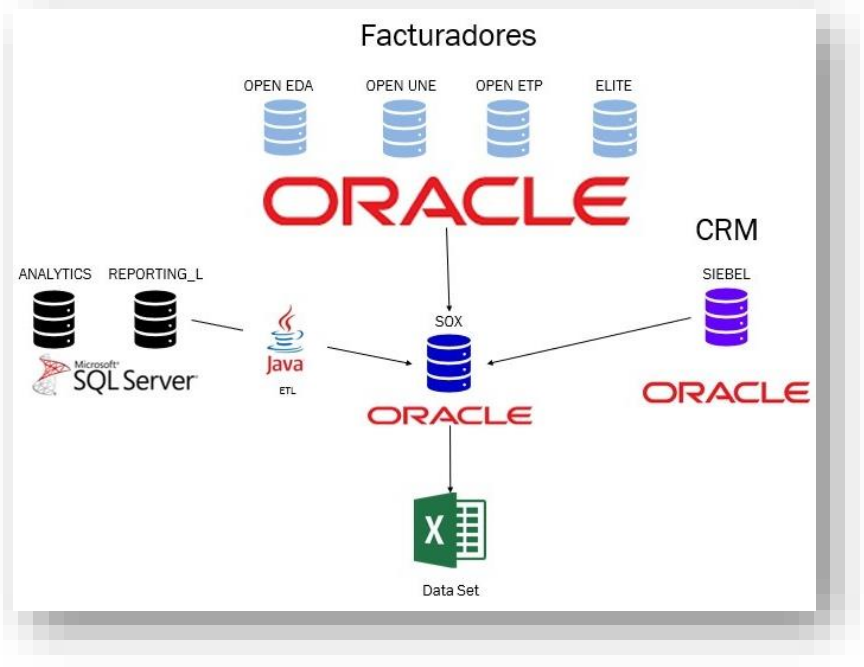


Figure 5: Fuente de datos. Recolección y almacenamiento

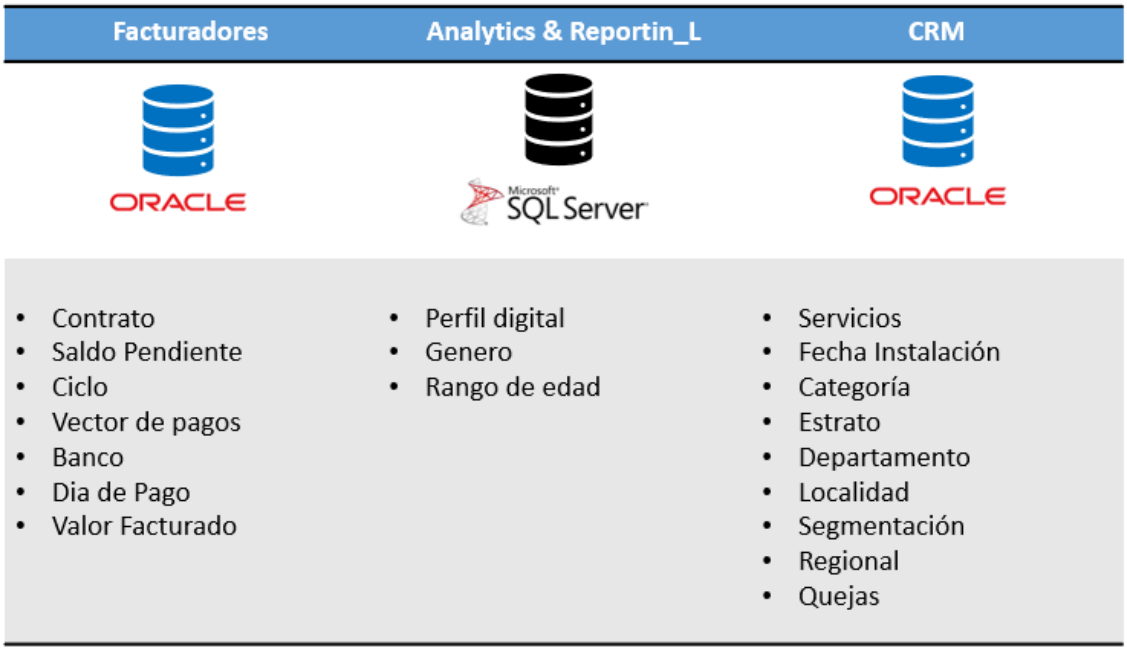


Figure 6: Data Selection

3.1. Almacenamiento del Proyecto integrador

El almacenamiento del proyecto se diseñó en AWS S3, en tres diferentes buckets, figura 7.

1. *Row:* Se almacena todos los Dataset que fueron generados en la base de datos de TIGO para el proyecto de perfilamiento de clientes.
   También, se realizó el análisis descriptivo, donde se realiza un data understanding y exploration de los diferentes Dataset.
2. *Preparation:* Después de un proceso de ETL, se almacena los datos para realizar los modelos de clasificación y LDA. Por lo que se realiza una limpieza básica sobre los datos y se realiza la división por regiones comerciales.
3. *Production:* Después del proceso de ETL se almacenan los Dataset finales para el proceso de analítica.

3.2. ETL de los datos del proyecto integrador

Los ETL presentados en este trabajo son tipo Batch.

3.2.1. Proceso ETL desde la base Oracle TIGO (Oracle – Row)

Para la extracción y basados en la figura 5 los datos se encuentran almacenados en BD relacionales ORACLE y SQLServer, estos datos son recopilados diariamente utilizando db_links de una BD centralizada (SOX) a cada una de las BDs con ETL tipo PL-SQL para las BD ORACLE y un ETL tipo JAVA.jar para la BD SQLServer; estos datos son almacenados en una BD ORACLE final donde se exporta el dataset, un archivo plano con los datos seleccionados en la exploración de datos.

En los ETLs se realizaron transformaciones de los datos, la más importante de ella es tomar lo facturado de cada cliente hasta por 12 meses y validar si el pago de esa factura se realizó oportunamente, pago no oportuno o no pago, asignándole a cada uno de estos pagos una calificación y según esta calificación asignarle la etiqueta inicial.

3.2.2.Proceso ETL del diseño del datalake (Row-Preparation)

Data understanding: En este proceso se realiza la exploración de los datasets, con el fin de entender cada una de las variables y realizar una exploración de los datos.

Se identificaron variables con valores nulos y duplicidad, que fueron eliminados, adicional se identificaron datasets que no cumplen con el alcance, que se mantienen en el bucket Row, ya que se puede realizar otro proyecto con diferentes análisis que incluyan estos datasets.

Se realizó una división de la información de acuerdo con las regiones comerciales definidas por la empresa, ya que los comportamientos comerciales son muy inherentes a la región. También se calcula el promedio de la calificación de servicio y se agrupa por cliente.

3.2.3.Proceso ETL del diseño del datalake (Preparation-Production)

En el ETL de preparation a production del archivo de quejas se debe realizar una transformación de los datos que incluye la eliminación de puntuación, caracteres especiales, adicional:

1. Tokenización por medio de la librería gensim
2. Generación de bigram y trigram con la librería gensim
3. Eliminación de Stopwords y Lematización con la librería gensim

Para la clasificación de clientes por región, se realizó una división de la información de acuerdo con las regiones comerciales definidas por la empresa, ya que los comportamientos comerciales son muy inherentes a la región.

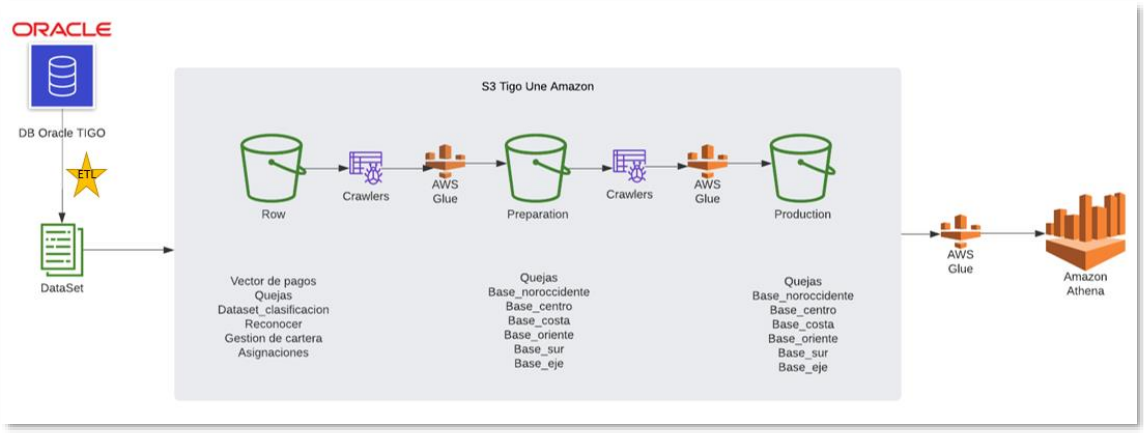


Figure 7: Datalake proyecto integrador

4. Entregables

4.1. Definición del proyecto.

Con el fin de predecir el comportamiento de pago de los clientes en sus facturas, se realizará un modelo de clasificación, el cual nos ayudará a puntualizar los clientes, que servirá para crear estrategias diferenciales para recuperación de cartera.

Para hacer esta clasificación, haremos uso de la implementación de los diferentes algoritmos de clasificación disponibles de manera que podamos establecer una relación, entre los datos del cliente y su comportamiento de pago. Adicional a esto, basado en la información de las quejas, se quiere identificar los tópicos principales y mirar su relación con el "No pago" de las facturas.

4.2. Entendimiento del problema.

Actualmente, Tigo tiene un reconocimiento básico de los usuarios respecto a sus comportamientos de pago, donde este solo se realiza por los días de mora de la cartera, sin embargo, se ha empezado a observar que estos comportamientos van más allá del pago, viéndose implicadas variables como las zonas donde se presta el servicio, variables demográficas de cliente, entre otras.

4.3. Entendimiento de los datos.

A continuación, se listas los Dataset que se evaluarán dentro del proyecto integrador (tabla 1):

Table 1: Entendimiento de los datos

Dataset	Descripción	Relevancia
Vector_pago_fijo	Recopilación de los facturadores y cálculo del vector	Alta
Ctrl_cuencobr	Recopilación de todos los facturados de los últimos 12 meses	Alta
Tbl_reconocer	Información del género y rango de edad	Media
Stg_quejas_siebel	Información de quejas de los últimos 6 meses	Alta
Tbl_gestion_cartera	Gestión de los clientes según su cartera	Baja
Tbl_asignaciones	Asignación de los clientes según la cartera	Baja

Se puede encontrar el detalle de cada una de las tablas en el Excel DATA_DESCRIPTIONS.XLSX

Adicional se realizó un análisis descriptivo con el fin de realizar una exploración total de los datos:

Existe un total de 5.548.249 con un total de 58 atributos cuantitativos y cualitativos; además de observar el número de meses que han tenido servicio en los contratos, donde se puede observar que la mayoría tienen 12 meses (figura 8).

Figure 8: Número de clientes por número de facturas

Dentro de la exploración de los datos del vector fijo se pude determinar, que la región que tiene mayor parcelación es el Noroccidente con un 52,8%, seguido por el eje cafetero con 13,9% (figura 9).

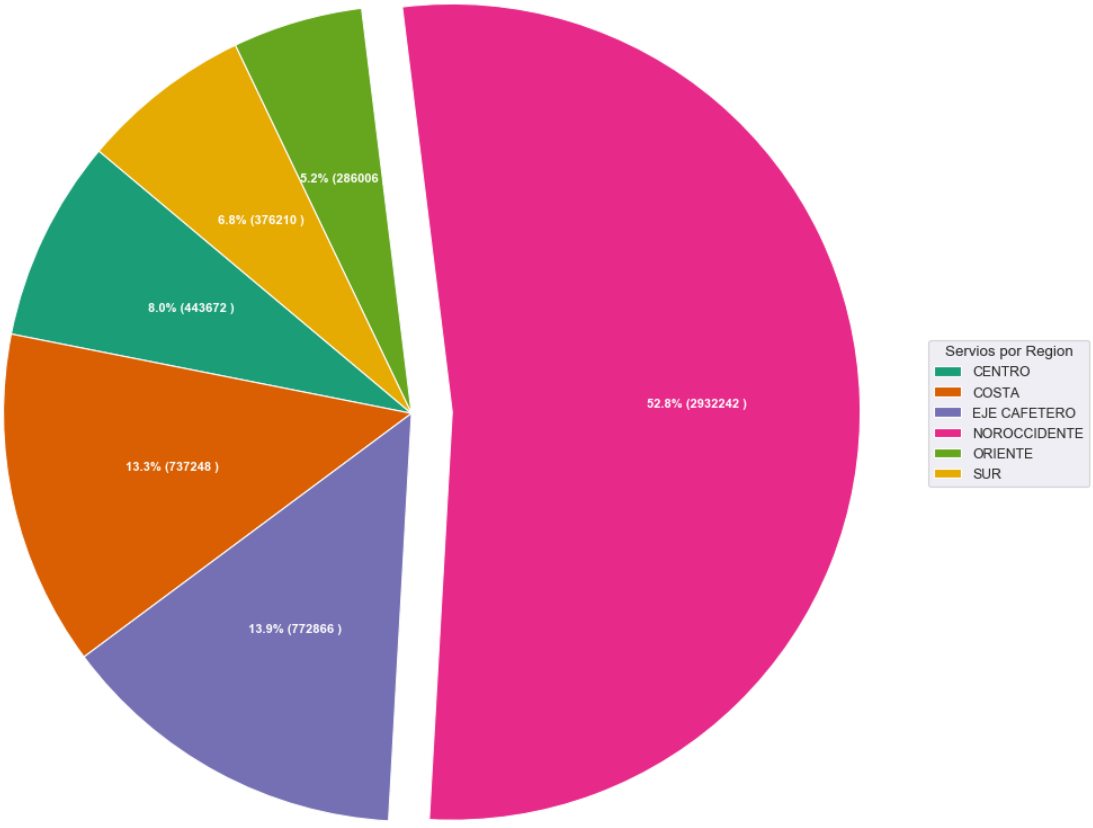


Figure 9: Regiones Tigo

Dentro de los días de pagos más populares se encuentra los principios de mes y el día 15 (figura 10).

Figure 10:Top 1 días de pago

Una de las formas de pago más populares es: PSE 18%, Gana en línea con 17% y seguido de Bancolombia con 9% (figura 11).

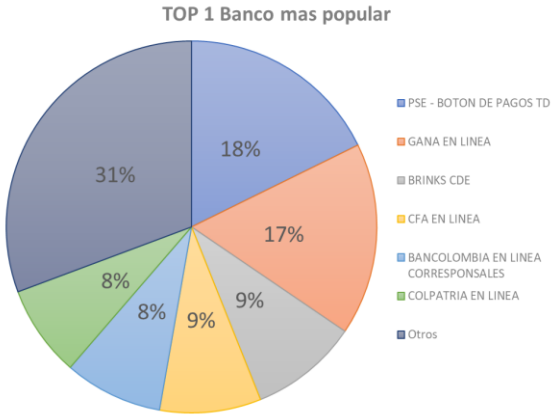


Figure 11: Banco más popular

La calificación por contrato muestra que 36,5% son de pago excelente, y 23,84% son clasificados como buenos (figura 12).

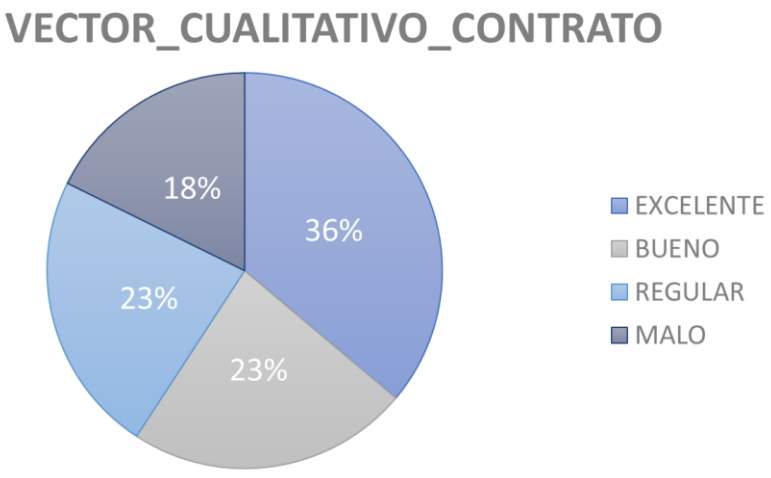


Figure 12: Calificacion por contrato

Dentro de lo servicios más populares que se encuentra en Tigo, se puede ver que El internet es el más popular con un 24% de participación, seguido de telefonía con un 22% y televisión con 18% (figura 13).

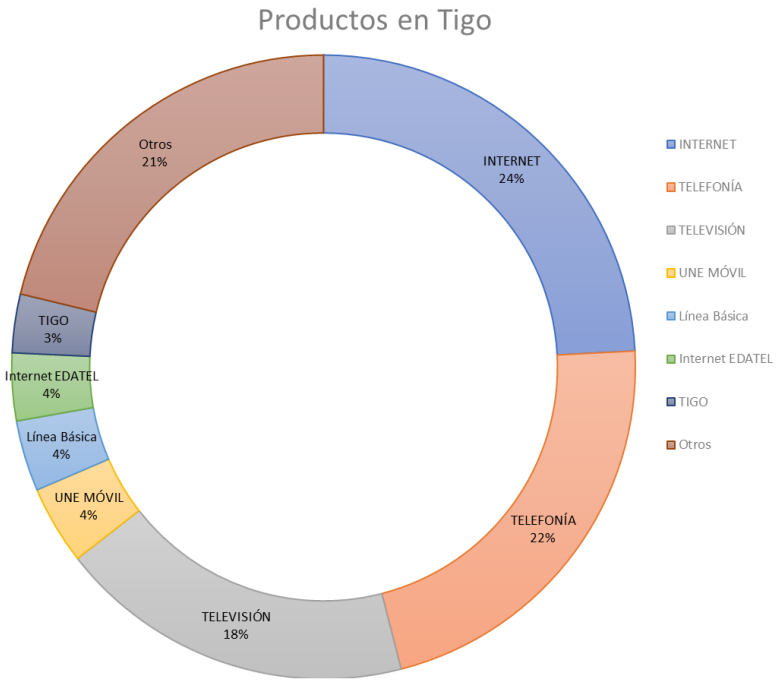


Figure 13: Productos Tigo

Tigo tiene 3 tipos de calificación para cada contrato: por cliente, por contrato y por servicio; como se puede observar su comportamiento es similar (figura 14).

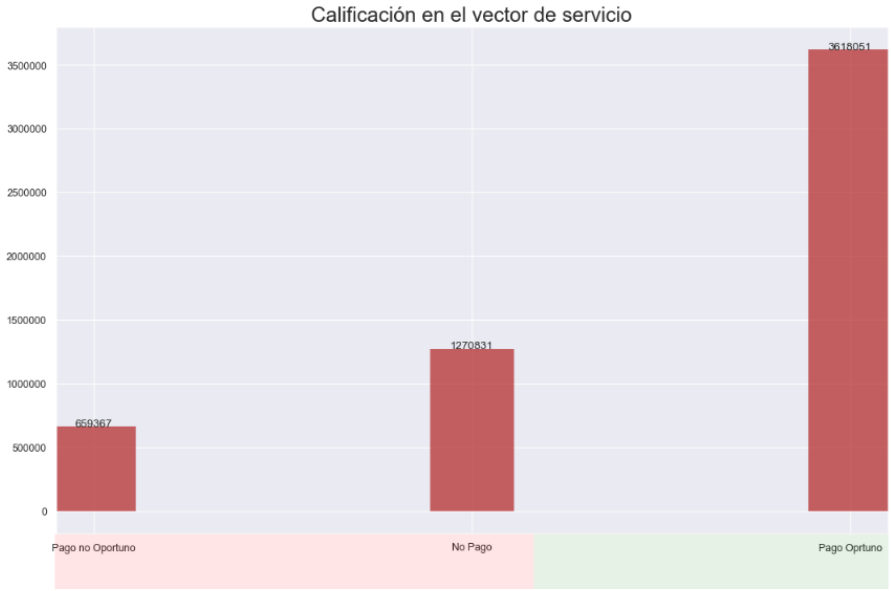


Figure 14: Calificación de cliente

Gracias al algoritmo de Árbol de decisión, se pudo identificar cuál es la variable más importante dentro de nuestro Dataset (figura 15).

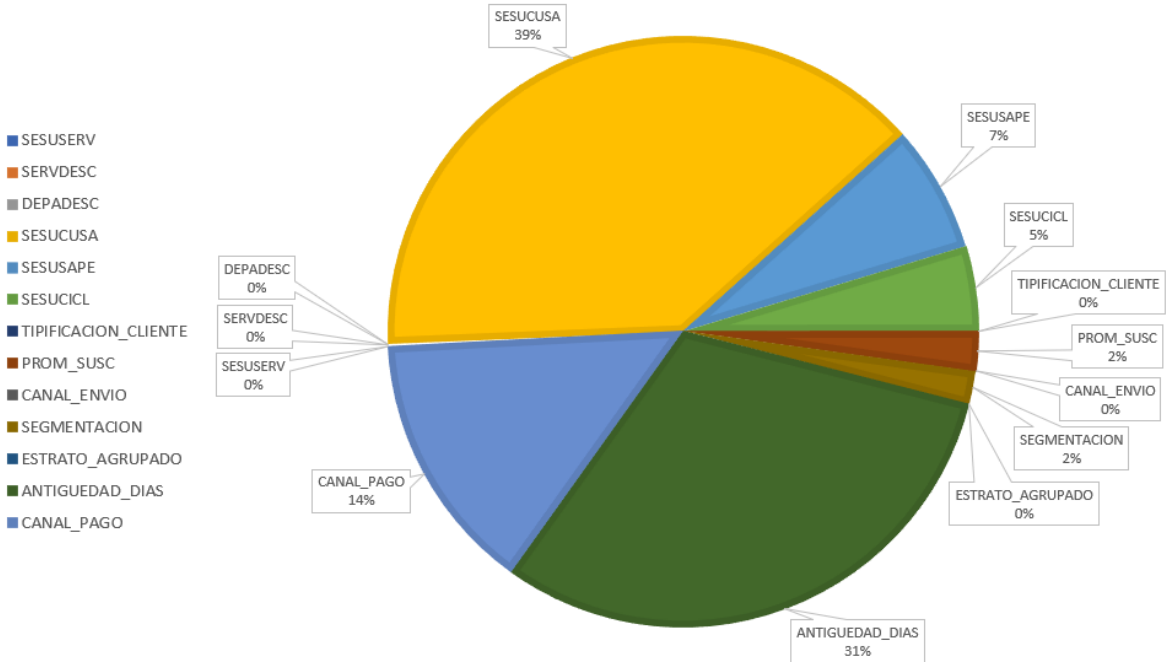


Figure 15: Variable más importante

4.4. Correlación

Con el fin de determinar las variables que están más relacionadas y así mismo poder realizar una reducción dimensional, teniendo en cuenta que las variables seleccionadas representan más del 60%, la correlación de variables es definida por Pearson para una muestra de datos como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Su resultado es el rango [-1, 1], pero ¿qué pasa si los datos son categóricos? ¿Cómo se resta del promedio de la característica? Una solución muy usada es el método one-hot encoding, pero esto hace que cada variable se divida en la cantidad de características disponibles haciendo que las variables aumenten considerablemente, lo que necesitamos es una medida de asociación entre características categóricas, para esto utilizaremos el coeficiente V de Cramer, la cual se basa en una variación nominal de la prueba de Chi-Cuadrado de Pearson, el resultado de la correlación está en el rango [0, 1] donde 0 significa que no hay asociación y 1 es una asociación completa, en el algoritmo construido para calcular la correlación de las variables se tienen en cuenta:

1. La matriz de confusión o tabulación cruzada se calcula con la función *crosstab* de pandas, de las dos variables a comparar.
2. La matriz de contingencia se calcula usando la función chi2_contingency() de la librería scipy, aplicada a la matriz de confusión anterior.
3. Con la matriz de contingencia se calcula el coeficiente phi:

$$\chi^2 = \sum_{ij} \frac{\left(n_{ij} - \frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}}$$

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

4. Con estos tres objetos podemos calcular el coeficiente V de Cramer:

$$V = \sqrt{\frac{\varphi^2}{min(k-1, r-1)}}$$

k: $número\ de\ columnas$
r: $número\ de\ filas$
n: $gran\ total\ de\ observaciones$

Al aplicar este algoritmo a las variables tenemos la siguiente matriz de correlación. Para una fácil visualización graficamos la matriz con un heatmap como se puede ver en la figura 16.

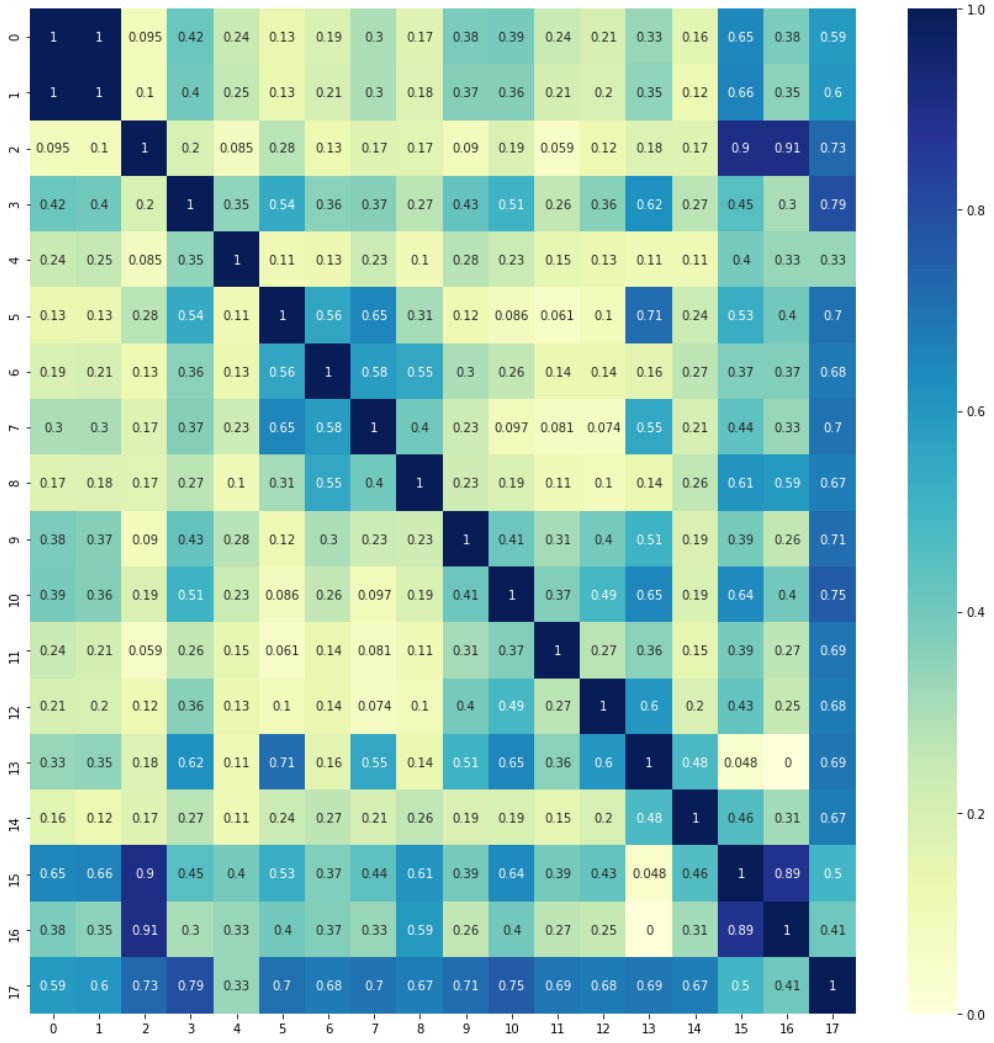


Figure 16: Correlación dataset Vector

Se realizó un análisis de correlación sobre las variables categóricas y continuas que serían utilizadas para los modelos clasificatorios.

Se puede ver que existe una correlación completa entre las variables 1(SESUSERV) y 2(SERVDESC) ya que son variables del tipo de servicio que tiene el cliente.

También se ve una relación entre la variable 3(SESUCUSA) y 15(SESUSAPE) que muestra las cuentas pendientes por pagar del cliente

Y finalmente, se encuentra una relación del 90% de la 3(SESUCUSA) y 16(SESUSAAN) que nos muestra el saldo anterior del cliente.

4.5. Entregables

Los entregables de este proyecto son:

Table 2: Entregables del proyecto integrador

Entregables	Enlace
Presentación pública	Proyecto_integrador.pptx
GitHub del proyecto	https://github.com/hsepulvedaj/proyecto_integrador
Reporte-técnico-y-modelos	ProyectoIntegrador_descriptivo.ipynb Pi_correlacion.ipynb LDA.ipynb lda_Quejas_2_v2.html class_report.py Mdl_preliminar_costa.ipynb Mdl_preliminar_eje.ipynb Mdl_preliminar_noroccidente.ipynb Mdl_preliminar_oriente.ipynb Mdl_preliminar_sut.ipynb Mdl_preliminar_centro.ipynb Mdl_outlierdetection_costa.ipynb Mdl_outlierdetection_eje.ipynb Mdl_outlierdetection_noroccidente.ipynb Mdl_outlierdetection_oriente.ipynb Mdl_outlierdetection_sut.ipynb Mdl_outlierdetection_centro.ipynb
Planeación Proyecto	Planeacion_Proyecto.xlsx
Descripción/contexto del proyecto	Proyecto_integrador.pdf

4.6. Preparación de datos

4.6.1. Preprocesamiento

A partir de los procesos de ETL anteriormente establecidos, obtenemos la información de forma semiadecuada pues nuestra solución va enfocada al perfilado de clientes y no de productos, por ende, al tener registros históricos de los productos y comportamiento de pago, existen escenarios favorables para ciertos clientes con un producto, pero desfavorable (moras) para otros productos, luego es necesario tener una calificación única por cliente a partir del número de productos que haya adquirido.

Se establece una medida de centralidad para los *n* productos que contienen los clientes y a partir de esta medida recrear la variable de interés de la siguiente manera:

Calificación del producto: valores entre [0 - 100]

Nueva variable de interés:

- No pago: valores entre [0 – 50]
- Pago inoportuno: valores entre [51 – 75]
- Pago: valores entre [76 – 100]

Teniendo así, una recategorización apta para un posible modelo multinomial el cual se ajusta a las reglas de negocio de la empresa y nos permite tener grupos coherentes basado en el antiguo vector de pagos.

4.6.2. Preparación de datos LDA

Con el fin de poder generar un modelo en el cual se pueda identificar los tópicos que son prominentes en la base de datos de quejas de los clientes de Tigo.

Por esta razón, se debe realizar una preparación de los datos que incluye la eliminación de puntuación, caracteres especiales, adicional:

4. Tokenización por medio de la librería gensim
5. Generación de bigram y Trigram con la librería gensim
6. Eliminación de Stopwords y Lematización con la librería gensim

El modelo LDA fue construido con la librería gensim Y NLTK para el preprocesamiento.

Figure 17: Modelo LDA

4.6.3. Outliers detections

Table 3: Características del modelo AVF

Método:	**AVF (Attribute Value Frequency)**
Descripción del método:	Estrategia de detección de valores atípicos, escalable para datos categóricos.
Objetivo:	Detección de valores atípicos, basándose en la frecuencia de cada uno de sus atributos.
Pasos:	1. Calcular la frecuencia de cada valor posible en cada uno de sus atributos. 2. Calcular la frecuencia media para cada individuo. 3. Los puntos con puntuaciones más bajas tienen más probabilidades de ser valores atípicos ya que tendrán valores poco frecuentes en promedio.
Ecuación:	$$AVFScore(X_i) = \frac{1}{m}\sum_{j=1}^{m} Freq(x_{ij})$$
Complejidad:	O(nm)

4.7. Modelos preliminares.

4.7.1. Modelo LDA sobre las quejas

Buscando innovar y aprovechar al máximo todas las bondades que obtenemos en la correcta aplicación estadística/informática de algoritmos, nos permitimos ver más allá de los millones de datos que contienen las empresas.

Quisimos representar en Tigo una solución basada en 2 fases que le brinde herramientas a las áreas de cartera y cobranza en la hora de ser oportunos y prever los posibles estados de pago en los clientes.

- Fase 1: LDA, como herramienta intuitiva de análisis en los temas de PQR que se generan día a día obteniendo así la capacidad de mejorar el servicio y/o crear campañas preventivas ante inconformidades recurrentes.
- Fase 2: Modelo de clasificación en el cual Tigo podrá apoyarse en la estimación de la probabilidad de pago que tienen los clientes según sus hábitos de pago y comportamiento histórico para prever posibles escenarios con grupos de clientes puntuales y actuar de manera oportuna.

En la figura 18 se puede visualizar las burbujas que representa un tema. Cuanto más grande es la burbuja, más frecuente es ese tema. Este es un buen modelo, ya que las burbujas son grandes y no se encuentran superpuestas y no se encuentran agrupadas en un solo cuadrante.

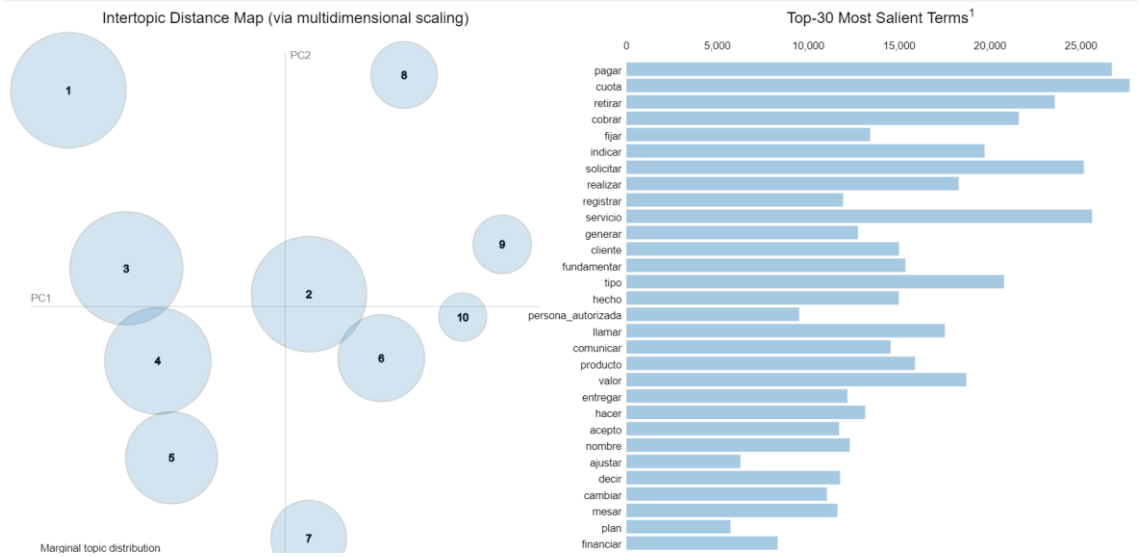


Figure 18: Modelo LDA

4.7.2. Competencia de modelos de clasificación

Buscando la resolución a partir de la información histórica de los clientes, en donde se comprende información de todos los productos adquiridos, formas habituales de pago y comportamiento oportuno de pagos e información demográfica, se realiza una serie de corridas de los siguientes algoritmos (figura 19):

- KNN (K-Neighboors Classifier)
- Árbol de decisión (DecisionTreeClassifier)
- Bosque Aleatorio (RandomForestClassifier)
- Red Neuronal Multicapa con Perceptrón (MLPClassifier),
- Regresión Logística (LogisticRegression)
- Nayve Bayes (GaussianNB)

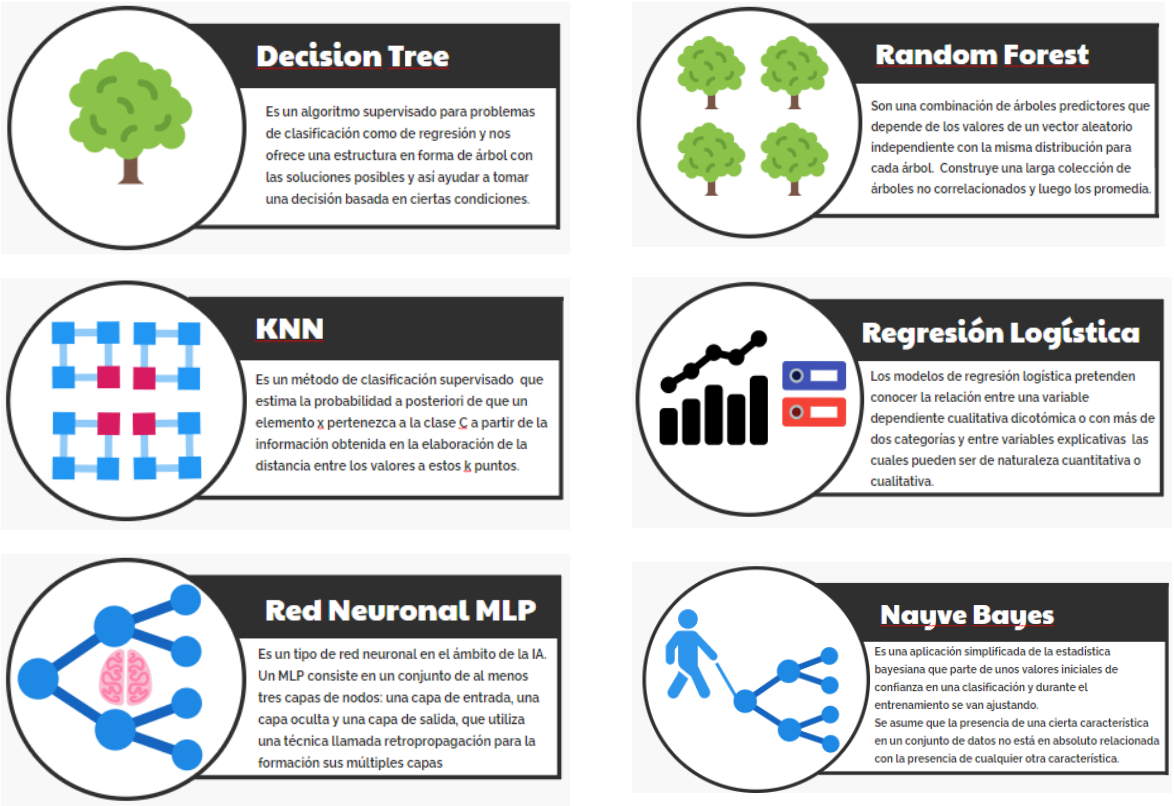


Figure 19: Resumen de los Clasificadores

El procedimiento que se realizó para evaluar los diferentes algoritmos fue el siguiente:

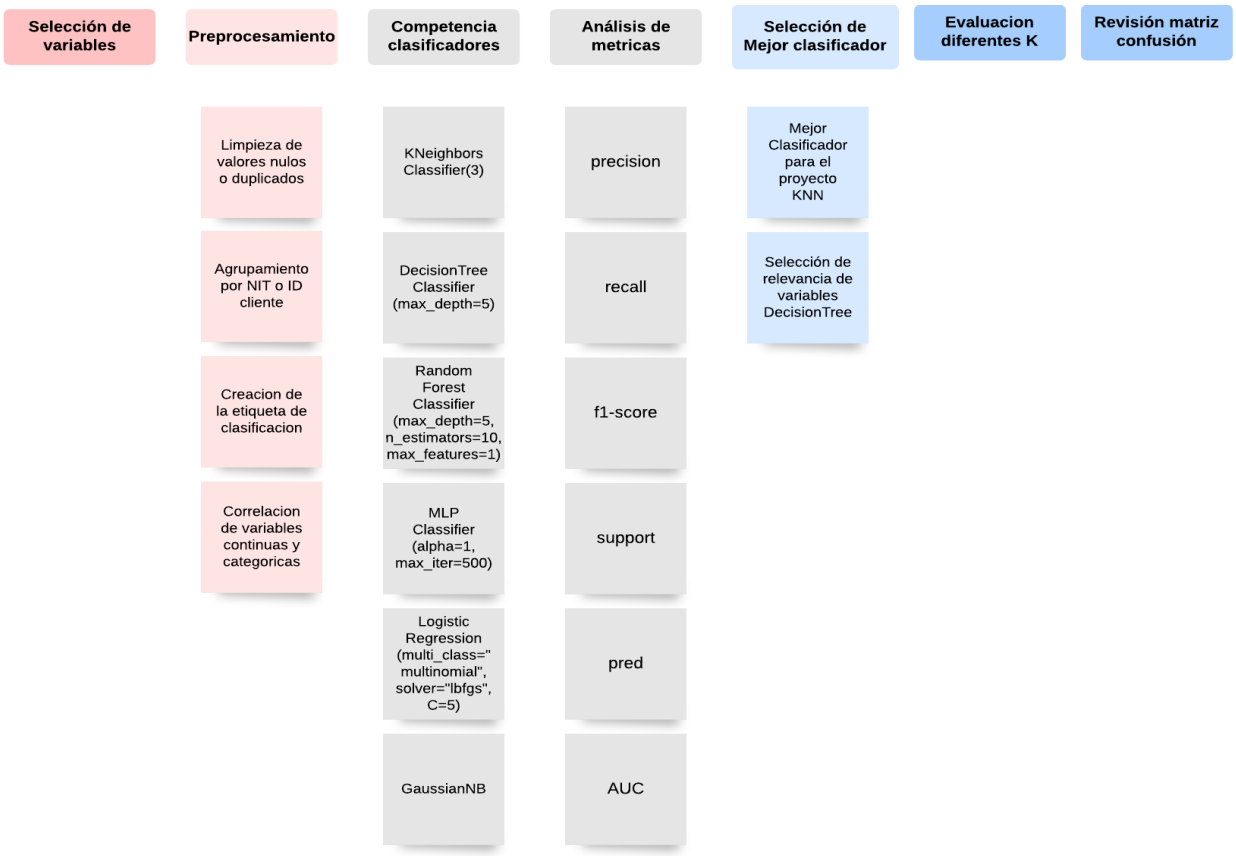


Figure 20: Procedimiento clasificadores

4.8. Modelos finales validados.

Con el fin de evaluar los modelos, se tomaron en cuenta las siguientes métricas:

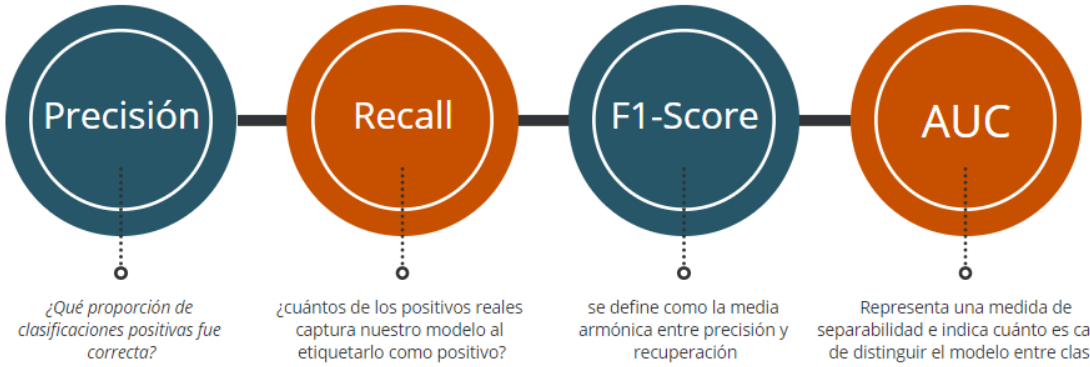


Figure 21: Medidas de Desempeño

Y para validar cualquier modelo, es necesario revisar la matriz de confusión figura 22, donde:

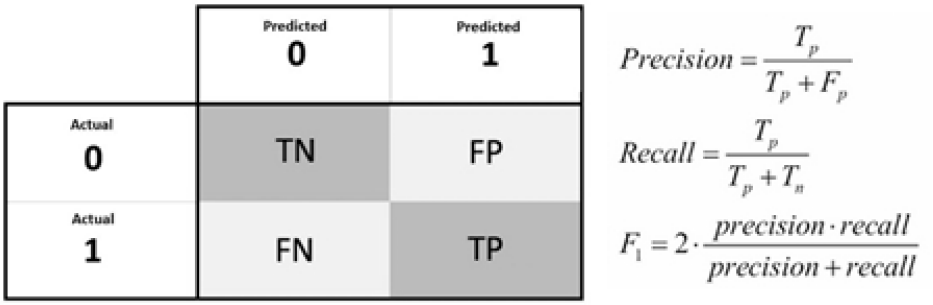


Figure 22: Matriz de confusión

El concepto de ROC y AUC se basa en el conocimiento de la matriz de confusión, especificidad y sensibilidad.

La sensibilidad nos dice qué porcentaje de clientes con buen hábito de pagos fueron identificadas correctamente; mientras la especificidad nos dice qué porcentaje de clientes con mal hábito de pago se identificaron correctamente, figura23.

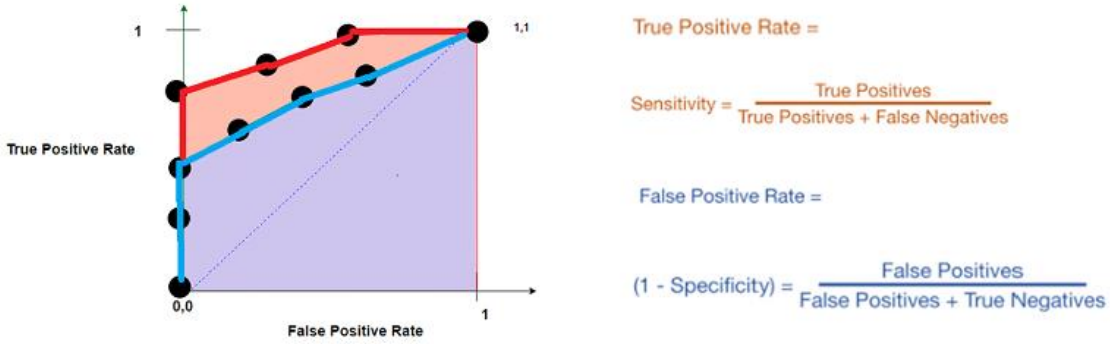
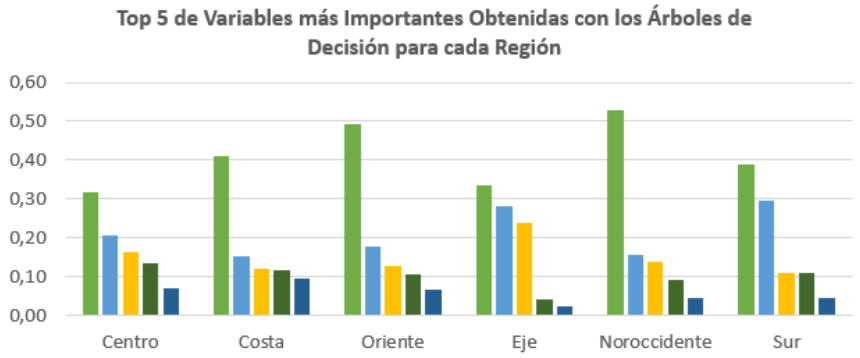


Figure 23: Sensibilidad vs Especificidad

Tomando como modelo final el algoritmo de KNN bajo la doble validación de la competencia de modelos y posterior validación bajo la metodología Cross Validation, la cual a partir de la definición de un numero de K-Folds, recrea particiones en las que se entrena K veces el modelo con mejores métricas, se observa explícitamente el buen ajuste para los niveles de pago inoportuno, pago y no pago a partir del AUC, F1-Score, recall y precisión.

4.8.1. Resultados de los clasificadores

En esta parte se hace una recopilación de los resultados obtenidos en la competencia de los clasificadores, el árbol de decisión, el KNN con diferentes k y el óptimo; entre otros resultados de importancia, figuras de la 24 a la 29.



Centro	Costa	Oriente	Eje	Noroccidente	Sur
Saldo Anterior	Cuentas con Saldo	Cuentas con Saldo	Saldo Pendiente	Cuentas con Saldo	Ciclo
Días Antigüedad	Tipif. Cliente	Saldo Anterior	Código Localidad	Saldo Anterior	Saldo Pendiente
Cuentas con Saldo	Canal Pago	Días Antigüedad	Desc. Serv.	Días Antigüedad	Canal Pago
Canal Pago	Saldo Anterior	Canal Pago	Días Antigüedad	Saldo Pendiente	Saldo Anterior
Saldo Pendiente	Saldo Pendiente	Prom. Suscrip.	Saldo Anterior	Canal Pago	Cuentas con Saldo

Figure 24: Top 5 de Variables más importantes obtenidas con los Árboles de Decisión para da Región.

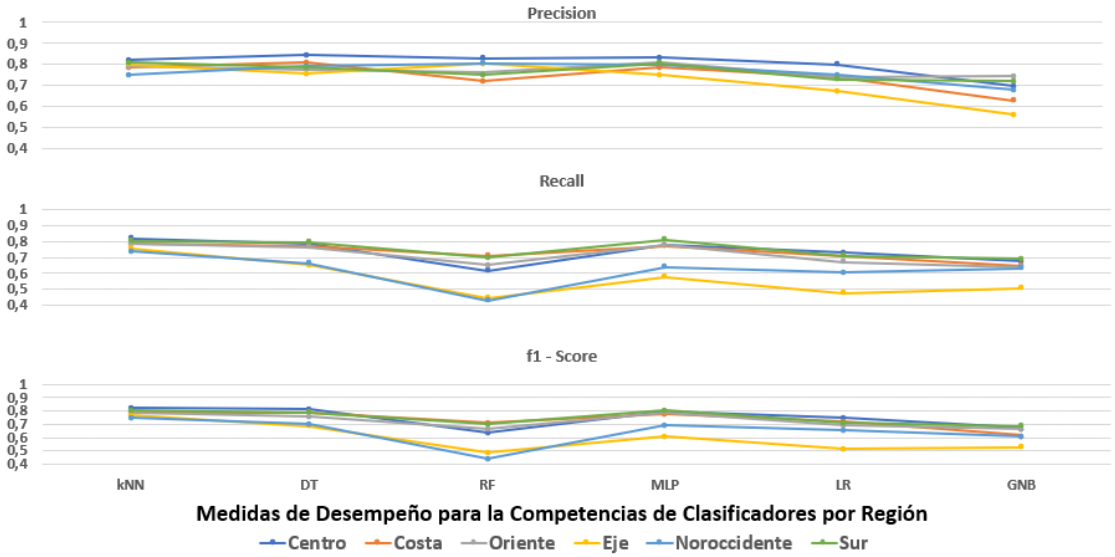


Figure 25: Medidas de Desempeño para la Competencia de Clasificadores por Región.

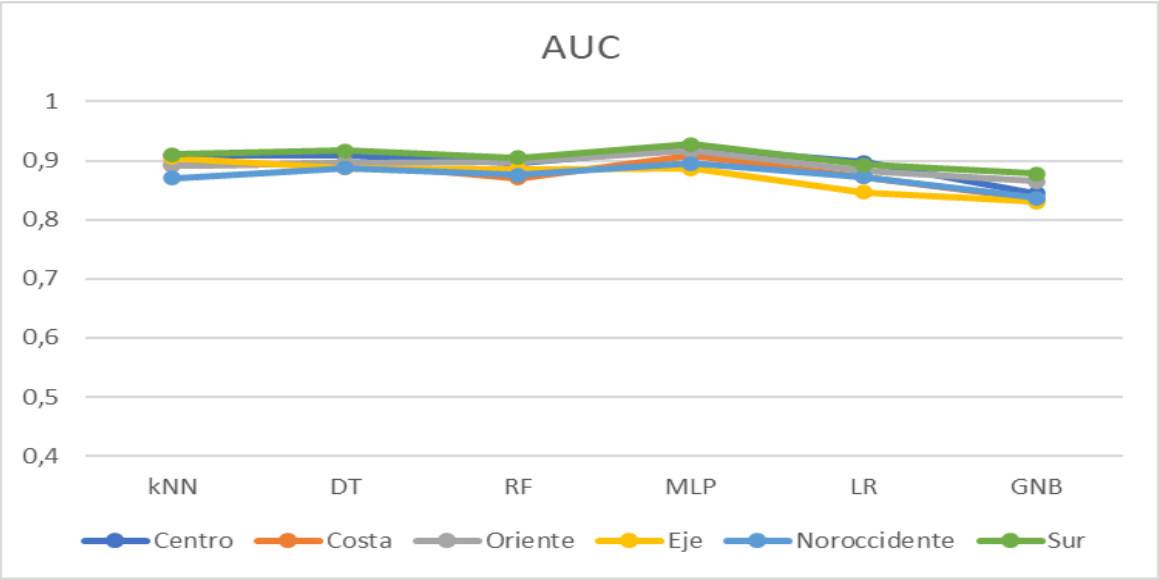


Figure 26: Medida de Desempeño AUC obtenida en la Competencia de Modelos por Región

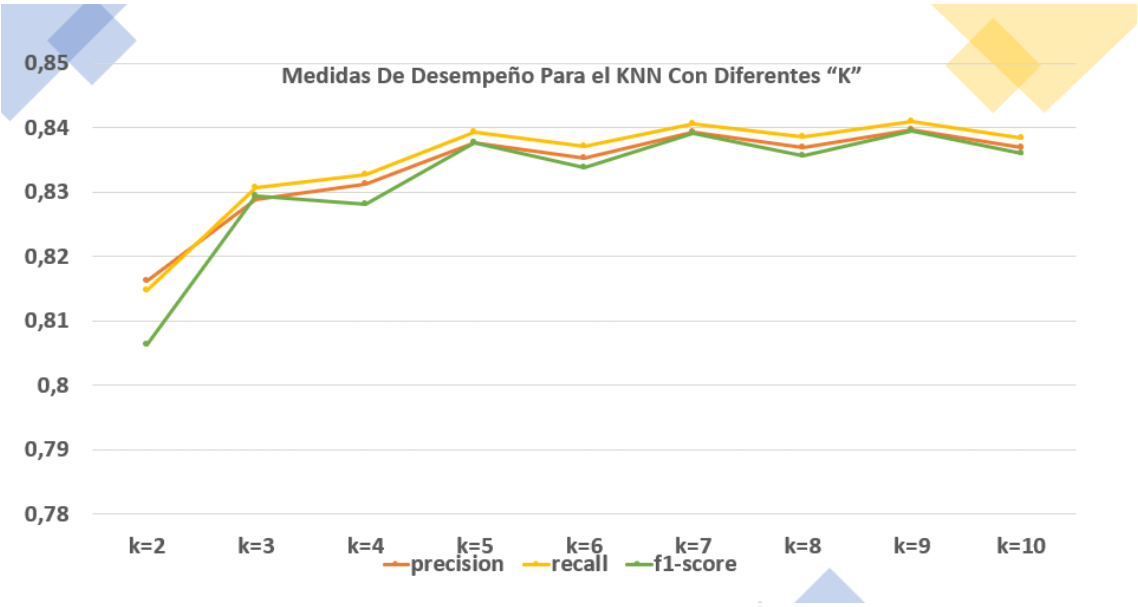


Figure 27: Medidas de Desempeño para el KNN con diferentes "K"

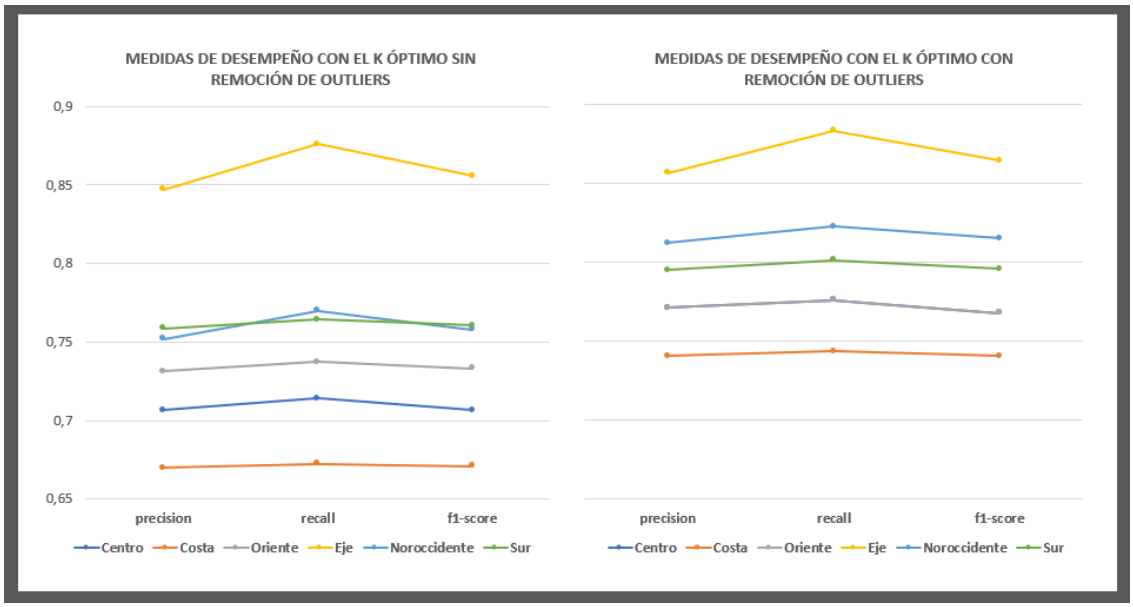


Figure 28: Medidas de Desempeño Con y Sin Outliers

Predicted

Actual	Centro			Costa			Oriente		
	NP	PI	P	NP	PI	P	NP	PI	P
NP	9904	1975	1821	24086	1542	5657	9904	1975	1821
PI	458	42531	3771	1241	52466	8997	458	42531	3771
P	2291	9032	16420	4897	13157	34656	2291	9032	16420

Actual	Eje			Noroccidente			Sur		
	NP	PI	P	NP	PI	P	NP	PI	P
NP	93640	1259	1336	20664	5825	9843	18899	839	1524
PI	2248	4530	572	6204	372583	27568	744	25606	2672
P	5400	2142	1241	7230	48892	62867	2278	4652	9392

MATRICES DE CONFUSIÓN OBTENIDAS PARA CADA REGIÓN APLICANDO EL MODELO KNN REMOVIENDO OUTLIERS

V. Mín. V. Máx.

Actual	Centro			Costa			Oriente		
	NP	PI	P	NP	PI	P	NP	PI	P
NP	9008	1686	3013	22388	2457	6463	6115	798	1460
PI	976	39979	5828	2795	46893	13043	511	25281	3985
P	2847	11049	13861	7917	15124	29692	1454	5295	8509

Actual	Eje			Noroccidente			Sur		
	NP	PI	P	NP	PI	P	NP	PI	P
NP	93434	1532	1338	15096	7379	8375	18358	1117	1794
PI	2981	3872	513	3918	312242	30221	864	24113	4047
P	5493	2334	1009	7785	53424	40297	2473	4901	8947

MATRICES DE CONFUSIÓN OBTENIDAS PARA CADA REGIÓN APLICANDO EL MODELO KNN SIN REMOVER OUTLIERS

Figure 29: Matrices de Confusión Con y Sin Outliers

4.9. Deployment

4.9.1. Plantear la implementación

Para esta primera parte se tiene que el modelo que mejor se ajusta para la clasificación es el kNN (k-Nearest Neighbor); además de poder usar el LDA (Linear Discriminant Analysis) para entender la mecánica de los pagos, usando los patrones de las quejas para dicho objetivo.

Sin embargo, para iniciar se dará a conocer los resultados logrados y su aplicabilidad en los procesos de entender y mejorar la recuperación de cartera; explicando en que información se enfocará el análisis de datos, que respuestas preliminares se arrojan y que variables serían las más apropiadas; además de los recursos que se puedan requerir.

Si hay un visto bueno por parte de gerencia, se plantearía un cronograma para su aplicación, que sería inicialmente como una estrategia piloto aplicada a algunos clientes para medir la recuperación de cartera y el entendimiento de cuál es la mejor estrategia para que el usuario pague, metodología que mediría la efectividad de la implementación basada en la información del modelo.

Esta prueba piloto dependerá en gran medida de su correcta aplicación y del trabajo conjunto con cartera; en su entendimiento y correcto análisis que permitirán llevarla a otro nivel dentro de la empresa.

4.9.2. Planear la monitorización y mantenimiento

El monitoreo se puede hacer con los mismos índices de recuperación de cartera que tiene Tigo y su mejoría, en los mismos ciclos de facturación o por periodos trimestrales o semestrales de pagos.

La efectividad y alcance de la herramienta aún no es clara y su uso es quien indicará cuánto mejorará los procesos de facturación, teniendo en cuenta que ninguna herramienta es 100% efectiva, lo cual dará la finalización del proyecto o la búsqueda de su aplicación en las diferentes zonas donde tenga presencia la compañía; además de revisar en que otras áreas podría ser útil para responder otras preguntas y plantear estrategias de marketing.

Respecto al mantenimiento, se puede pensar en la sincronización del modelo con las bases de datos y su actualización en tiempo real con el uso de APIs, que claramente también reduciría los tiempos de procesamiento.

Todo esto generará desde el área de cartera de Medellín la posibilidad de tener un informe final del modelo con resultados y desempeño; para revisar que aspectos han mejorado y si se ha llegado al límite de su alcance y plantear que otros proyectos pueden verse influenciados por las técnicas de minería de datos.

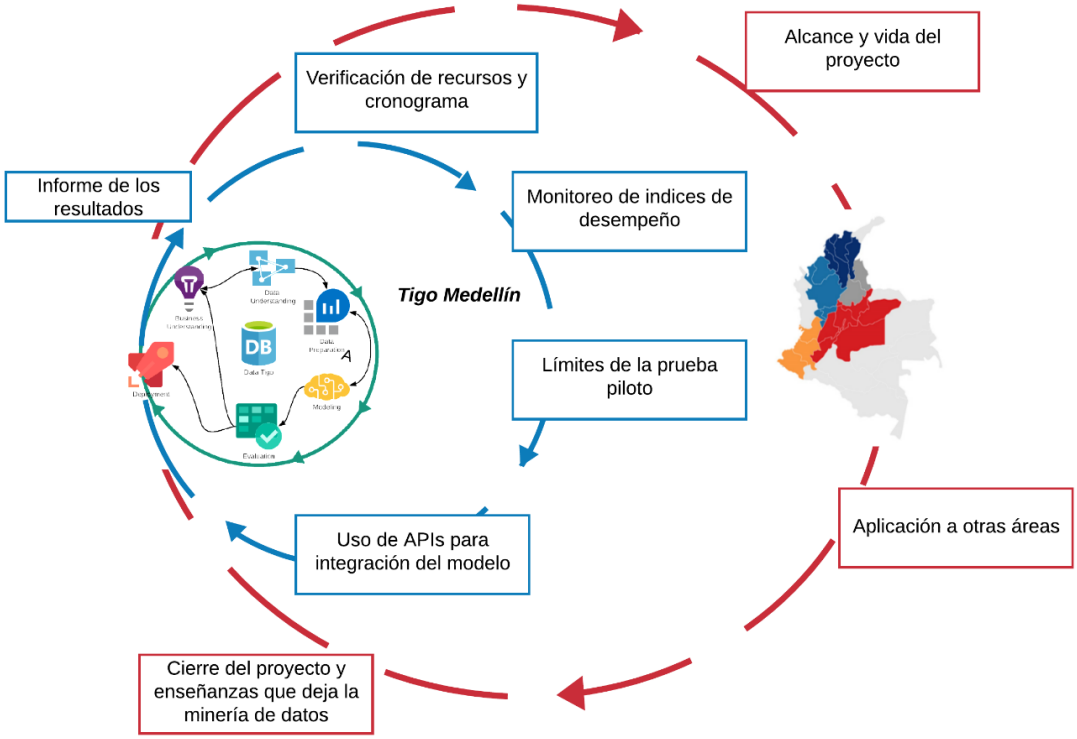


Figure 30: Despliegue de la propuesta de DM

5. Fechas entregas

Estructura	Task	Start Date	Finish Date
Avance 1	Definición del proyecto	**5/10/2020**	5/25/2020
Avance 1	Entendimiento del problema	**5/10/2020**	5/25/2020
Avance 1	Entendimiento de los datos	**5/10/2020**	5/25/2020
Avance 1	Creación de presentación Publica	**5/10/2020**	5/25/2020
Avance 1	Creación proyecto en GitHub	**5/10/2020**	5/25/2020
Avance 1	Reporte técnico y modelos	**5/10/2020**	5/25/2020
Avance 2	Creación de presentación Publica	**5/25/2020**	6/2/2020
Avance 2	Creación proyecto en GitHub	**5/25/2020**	6/2/2020
Avance 2	Reporte técnico y modelos	**5/25/2020**	6/2/2020
Avance 2	Disposición tecnologías para el proyecto	**5/25/2020**	6/2/2020
Avance 2	Preparación de los datos	**5/25/2020**	6/2/2020
Avance 2	Modelos preliminares	**5/25/2020**	6/2/2020
Avance 3	Creación de presentación Publica	**6/2/2020**	6/14/2020
Avance 3	Creación proyecto en GitHub	**6/2/2020**	6/14/2020
Avance 3	Reporte técnico y modelos	**6/2/2020**	6/14/2020
Avance 3	Modelos preliminares validos	**6/2/2020**	6/14/2020
Avance 3	producto-desplegado	**6/2/2020**	6/14/2020
Definición	Definición del proyecto a presentar	**5/1/2020**	5/2/2020
Definición	Envío del correo al PI	**5/1/2020**	5/3/2020
Definición	Propuesta - Documento	**5/10/2020**	5/21/2020
Evaluación	Creación de presentación Publica	**6/14/2020**	6/20/2020
Evaluación	Creación proyecto en GitHub	**6/14/2020**	6/20/2020
Evaluación	Reporte técnico y modelos	**6/14/2020**	6/20/2020
Evaluación	Modelos finales validos	**6/14/2020**	6/20/2020

6. Conclusiones

- A partir de la competencia de modelos de clasificación para cada región de la operación a nivel nacional de Tigo, pudimos obtener un clasificador estable y con métricas bastante aceptables obteniendo un buen número de clasificaciones correctas dentro de los márgenes con tolerancias ante los errores de tipo 1 y 2 para el área de cartera.

- Al resaltar los diferentes parámetros para los cuales el algoritmo K-Nearest Neighbors fue puesto a prueba, obtuvimos métricas de aprobación bastante similares para las diferentes clasificaciones en regiones y los hábitos de pago de los clientes, dándonos así un campo más amplio de seguir generando pruebas en caso de que se quieran hacer campañas preventivas a nivel nacional y no regional.

- Bajo el análisis y recuperación de textos, obtuvimos un claro caso de estudio a partir de las Peticiones, Quejas y Reclamos que se generan día a día en Tigo, obteniendo así un modelo de tópicos donde se resaltan posibles focos de atención para los clientes y que permite la generación de posibles sinergias entre áreas y desarrollos por parte de TI en donde se integren los resultados del análisis de tópicos y las búsquedas puntuales de clientes según su inquietud.

- Destacamos y comprobamos que la correcta aplicación de las metodologías para datos continuos y categóricos, pueden resaltar mejorías en los modelos de clasificación, pues una adecuada separación nos permite desde el análisis descriptivo poder entender que requieren los modelos y si estos se podrán acoplar a las expectativas en las empresas.

- Generamos una solución para el área de cartera que permitirá acoplarse con otras metodologías no tan técnicas (como las de riesgo de crédito/pricing) que se apalanquen bajo las características que permiten obtener la probabilidad de que un cliente tenga un pago recurrente y estable o no estable, a partir de su historia.

7. Referencia

1. IBM. IBM Knowledge Center - Ορισμός υπηρεσίας. Published online 2013. Accessed June 26, 2020. https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html

2. Tigo. Mapa de Cobertura | Tigo | Planes de Internet, Televisión y Móvil para tí y tu hogar. Published 2020. Accessed June 26, 2020. https://www.tigo.com.co/nuestra-compania/corporativo/quienes-somos

3. dblp: Intelligent Systems Reference Library. Accessed June 26, 2020. https://dblp.org/db/series/isrl/index

4. Suri NNRR, Murty N, Athithan MG. *Outlier Detection: Techniques and Applications A Data Mining Perspective*.; 2019. Accessed June 26, 2020. https://b-ok.lat/book/4981230/fabc16?regionChanged

5. Plot C, Histogram M, Boxplot M, Deviation PP, Bars D. Top 50 matplotlib Visualizations – The Master Plots (with full python code). Published online 2019:1-71. Accessed June 26, 2020. https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

6. Zychlinski S. The Search for Categorical Correlation - Towards Data Science. Published 2018. Accessed June 26, 2020. https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c

7. Prabhakaran S. Gensim Topic Modeling - A Guide to Building Best LDA models. Machine Learning plus. Published 2018. Accessed June 26, 2020. https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

8. Wikipedia. F1 score. Wikipedia. Published 2020. Accessed June 26, 2020. https://en.wikipedia.org/wiki/F1_score

9. Schott M. K-Nearest Neighbors (KNN) Algorithm for Machine Learning. Capital One Tech. Published 2019. Accessed June 26, 2020. https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26

10. Roman V. Algoritmos Naive Bayes: Fundamentos e Implementación. medium. Published 2019. Accessed June 26, 2020. https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementación-4bcb24b307f

11. perceptrón multicapa - Multilayer perceptron - qwe.wiki. Accessed June 26, 2020. https://es.qwe.wiki/wiki/Multilayer_perceptron

12. Khandelwal R. Decision Tree and Random Forest - Data Driven Investor - Medium. Published 2018. Accessed June 26, 2020. https://medium.com/datadriveninvestor/decision-tree-and-random-forest-e174686dd9eb

13. Breiman L, Cutler A. Random forests - classification description. Published online 2004. Accessed June 26, 2020. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#overview

14. sklearn.metrics.precision_recall_fscore_support — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

15. sklearn.metrics.auc — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html?highlight=auc

16. sklearn.model_selection.RepeatedStratifiedKFold — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKFold.html?highlight=stratifiedkfold#sklearn.model_selection.RepeatedStratifiedKFold

17. Scikit-learn.org. sklearn.linear_model.LogisticRegression — scikit-learn 0.23.1 documentation. Published 2020. Accessed June 26, 2020. https://scikit-

learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logistic#sklearn.linear_model.LogisticRegression

18. sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html?highlight=kneighborsclassifier#sklearn.neighbors.KNeighborsClassifier

19. sklearn.tree.DecisionTreeClassifier — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html?highlight=decisiontree#sklearn.tree.DecisionTreeClassifier

20. sklearn.neural_network.MLPClassifier — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html?highligh=mlp#sklearn.neural_network.MLPClassifier

21. 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=randomforest#sklearn.ensemble.RandomForestClassifier

22. sklearn.naive_bayes.GaussianNB — scikit-learn 0.23.1 documentation. Accessed June 26, 2020. https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html?highlight=gaussiannb#sklearn.naive_bayes.GaussianNB

23. Sasaki Y. The truth of the F-measure The truth of the F-measure. 2015;(January 2007):1-6.

24. Tema. *Perceptron Multicapa OPENCOURSEWARE REDES DE NEURONAS ARTIFICIALES Inés M. Galván-José M. Valls*. Vol 1.

25. Haykin S. *Neural Networks and Learning Third Edition*. Vol 127.; 2009.

26. Frank E, Bouckaert RR. Naive bayes for text classification with unbalanced classes. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 4213 LNAI. ; 2006:503-510. doi:10.1007/11871637_49