**PERFILADO DE CLIENTES POR HÁBITO DE PAGO DISCRIMINADO POR REGIONES**

Daian Paola Fajardo Becerra

[dpfajardob@eafit.edu.co](mailto:dpfajardob@eafit.edu.co)

Juan Carlos Agudelo Acevedo

[jcagudeloa@eafit.edu.co](mailto:jcagudeloa@eafit.edu.co)

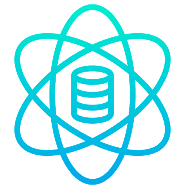
Hernán Sepúlveda Jiménez

[hsepulvedj@eafit.edu.co](mailto:hsepulvedj@eafit.edu.co)

Juan David Sanz Ramírez

[jdsanzr@eafit.edu.co](mailto:jdsanzr@eafit.edu.co)

**MAESTRÍA EN CIENCIA DE DATOS Y ANALÍTICA**



**Proyecto Integrador**

**Mayo 2020**

Contenido

[1. Descripción del problema: 3](#_Toc42954613)

[2. Metodología: 4](#_Toc42954614)

[2.1. CRISP-DM 4](#_Toc42954615)

[3. Fuente de Datos 5](#_Toc42954616)

[3.1. Almacenamiento del Proyecto integrador 6](#_Toc42954617)

[3.2. ETL de los datos del proyecto integrador 6](#_Toc42954618)

[4. Entregables 8](#_Toc42954619)

[4.1. Definición del proyecto. 8](#_Toc42954620)

[4.2. Entendimiento del problema. 8](#_Toc42954621)

[4.3. Entendimiento de los datos. 8](#_Toc42954622)

[4.4. Entregables 8](#_Toc42954623)

[4.5. Preparación de datos 9](#_Toc42954624)

[4.5.1. Preprocesamiento 9](#_Toc42954625)

[4.5.2. Preparación de datos LDA 9](#_Toc42954626)

[4.6. Modelos preliminares. 10](#_Toc42954627)

[4.6.1. Descriptivo 10](#_Toc42954628)

[4.6.2. Modelo LDA sobre las quejas 11](#_Toc42954629)

[4.6.3. Competencia de modelos de clasificación 12](#_Toc42954630)

[4.7. Modelos finales validados. 12](#_Toc42954631)

[5. Requerimientos tecnológicos 13](#_Toc42954632)

[6. Fechas entregas 14](#_Toc42954633)

# Metodología:

## 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.

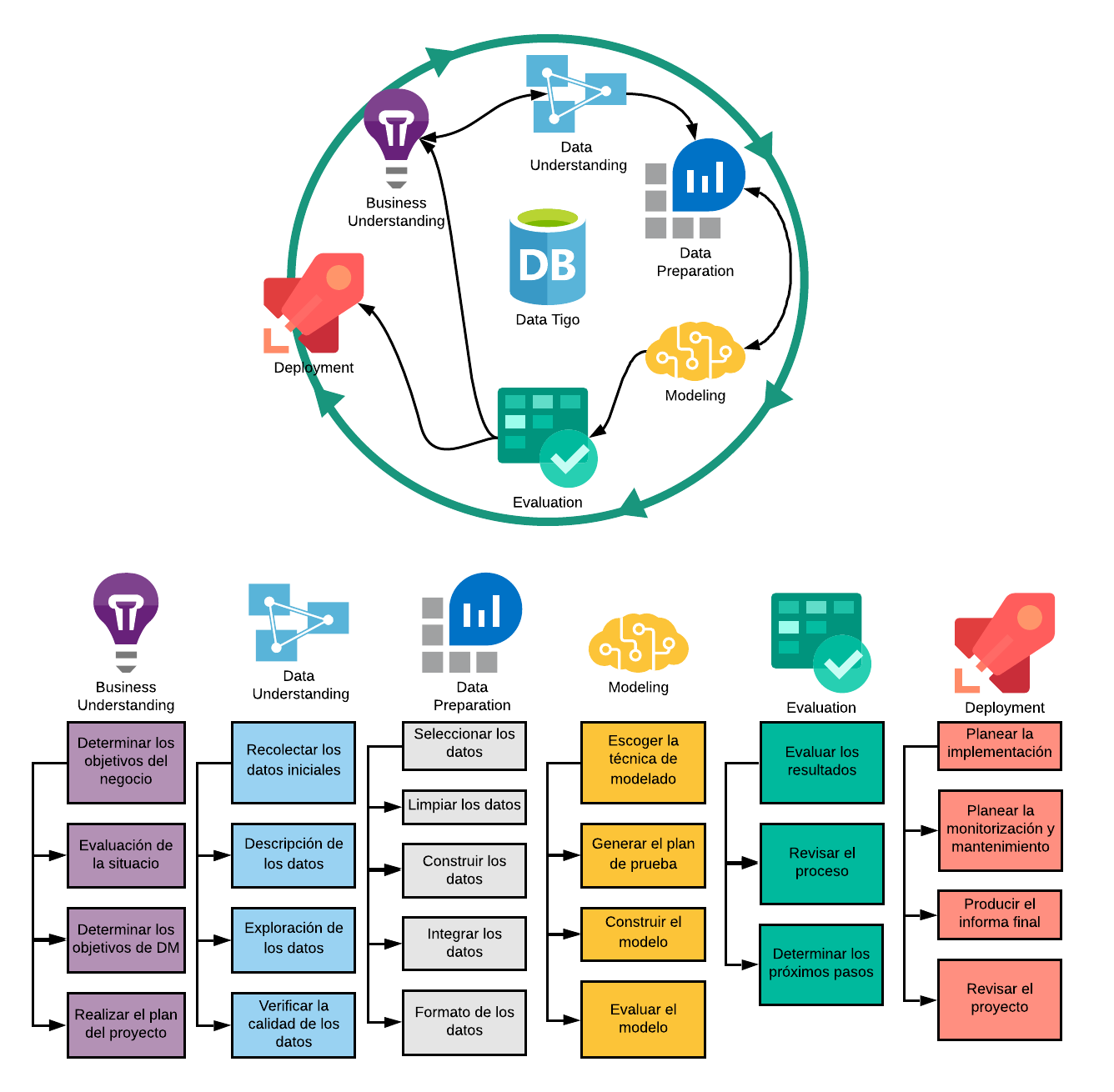


Figure 1: Modelo de CRISP-DM [Rodríguez, 2020]

# Business understanding:

### 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

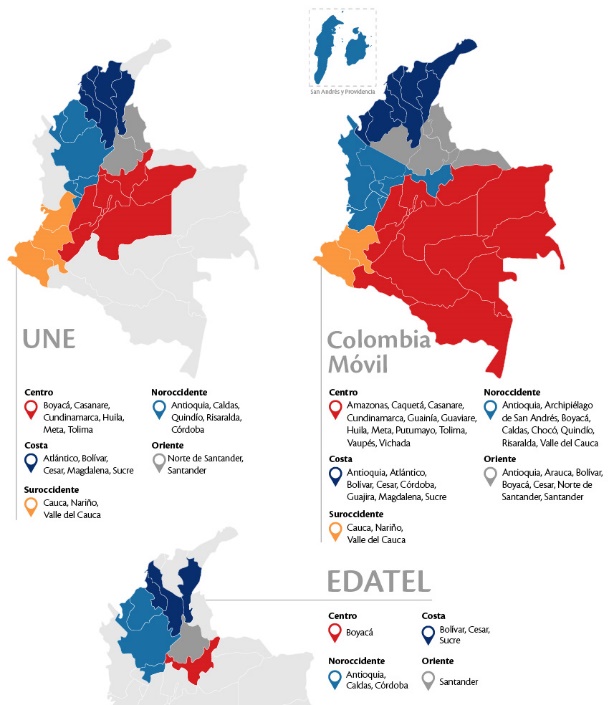
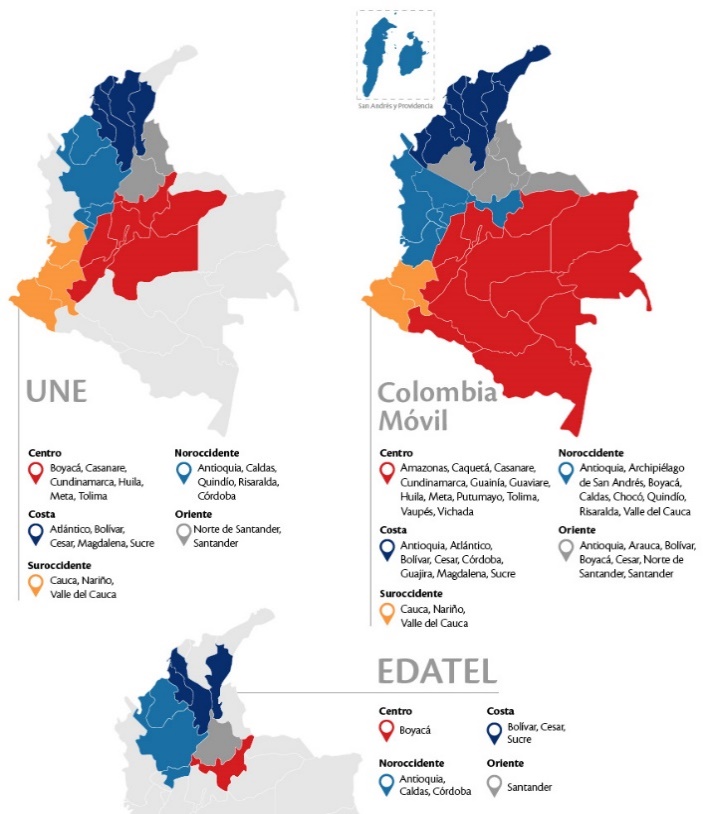
* ¿Existe alguna relación entre los atributos de cliente con su comportamiento de pago?
* ¿Se puede pronosticar el comportamiento de pago de un cliente nuevo de acuerdo con la data histórica de clientes existentes?
* ¿Las quejas y reclamos de un cliente tienen algún efecto en el comportamiento de pago?

### 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.



# 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.

1. CRM:

Contiene la información característica del cliente.

1. Fuentes de datos de analíticos:

De aquí se toma la información ya procesada del cliente.

1. 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.

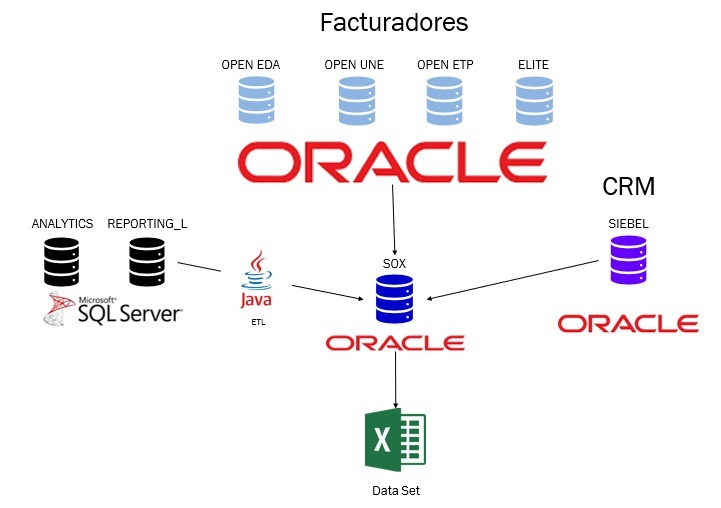


Figure 2: Fuente de datos. Recolección y almacenamiento

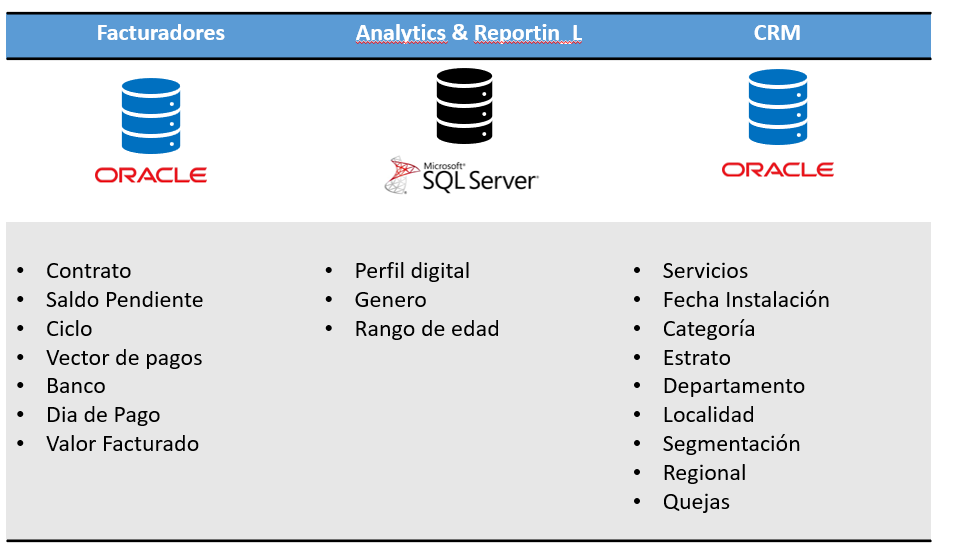


Figure : Data Selection

## Almacenamiento del Proyecto integrador

El almacenamiento del proyecto se diseño en AWS S3, en tres diferentes buckets.

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.

1. *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.
2. *Production:* Después del proceso de ETL se almacenan los Dataset finales para el proceso de analítica.

## ETL de los datos del proyecto integrador

Los ETL presentados en este trabajo son tipo Batch.

* + 1. *Proceso ETL desde la base Oracle TIGO (Oracle – Row)*

Para la extracción y basados en la *figura 2* 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.

* + 1. *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 realizo 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.

* + 1. *Proceso ETL del diseño del datalake (Preparation-Production)*

En el ETL de preparation a producción 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. JUAN CARLOS

A screenshot of a cell phone

Description automatically generated

Figure : Datalake proyecto integrador

# Entregables

## 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.

## 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.

## Entendimiento de los datos.

A continuación, se listas los Dataset que se evaluaran 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](https://github.com/hsepulvedaj/proyecto_integrador)

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 5).

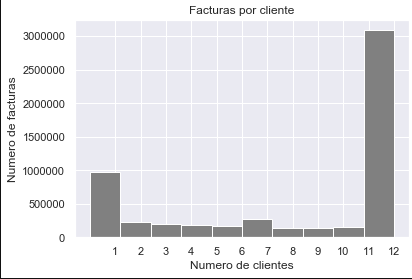


Figure 5: Número de clientes por número de facturas

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 6).

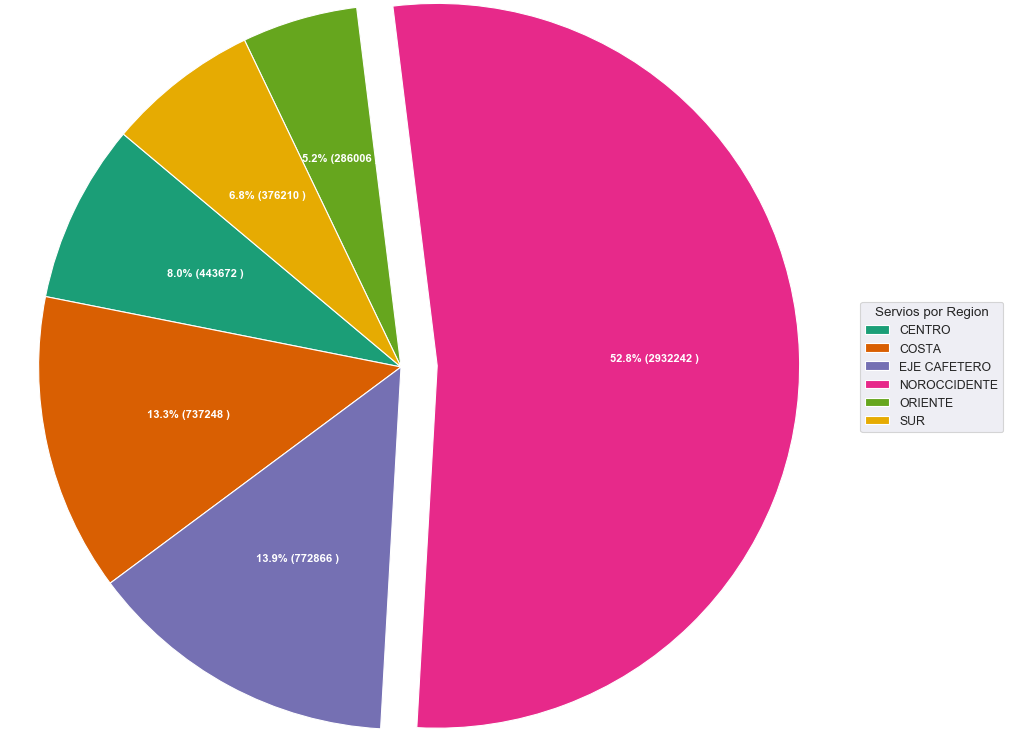


Figure 6: Regiones Tigo

Dentro de los días de pagos más populares se encuentra los principios de mes y el día 15 (figura 7).



Figure 7:Top 1 dias de pago

La calificación por contrato muestra que 36,5% son de pago excelente, y 23,84% son clasificados como buenos (figura 8).

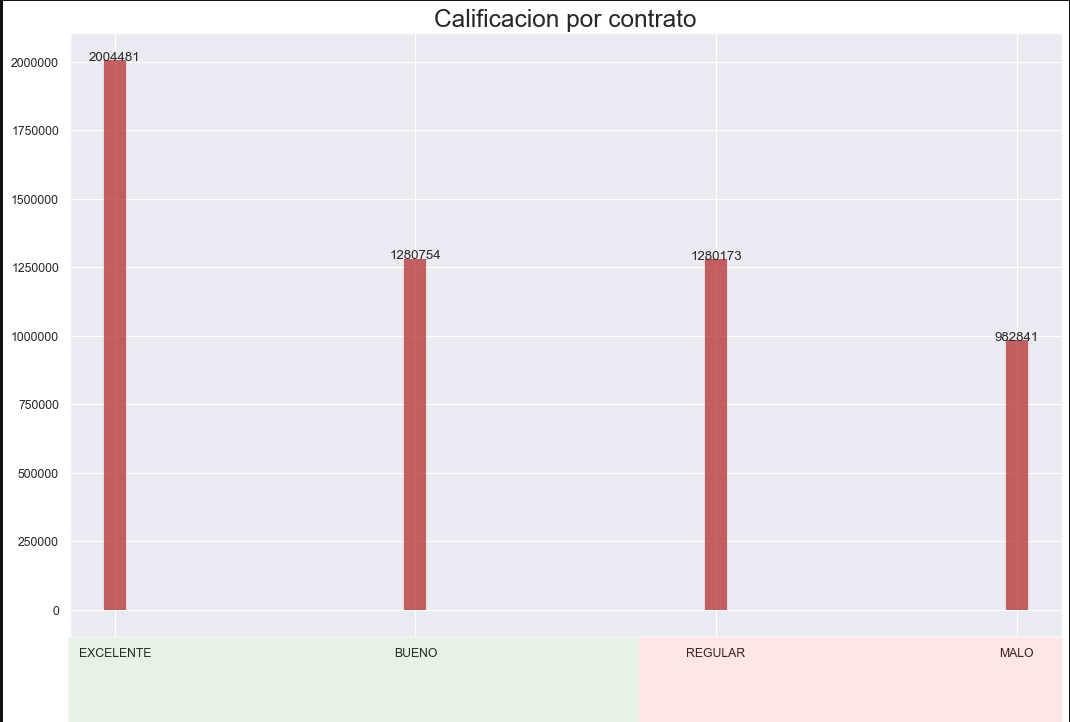


Figure 8: Calificacion por contrato

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 9).

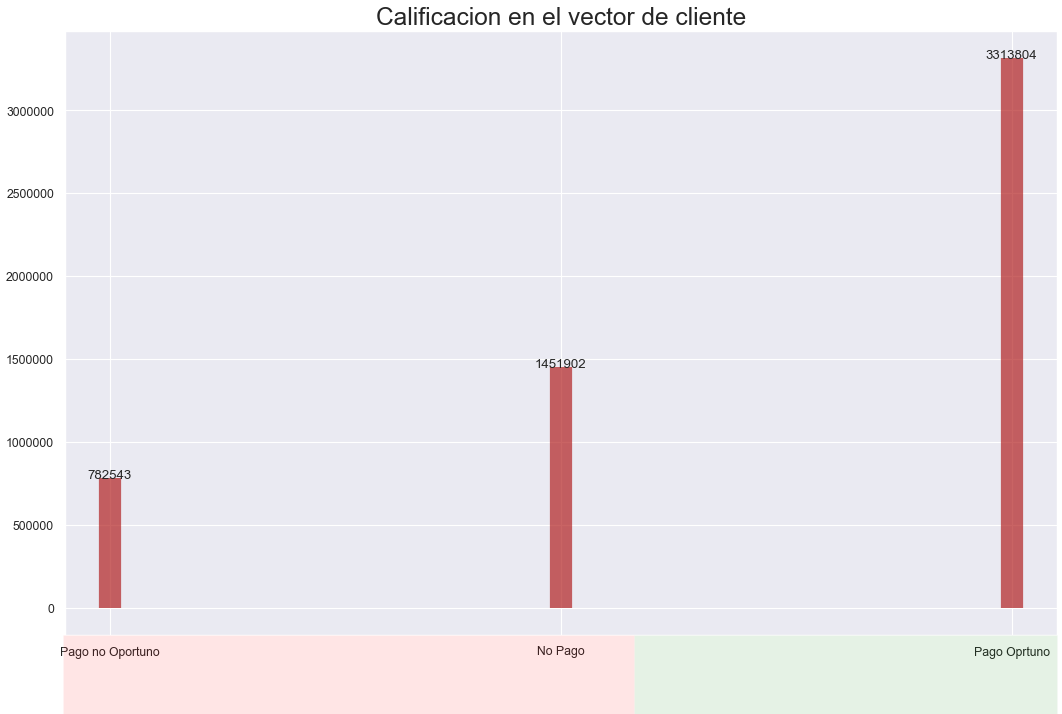


Figure 9: Calificación de cliente

## Entregables

Los entregables de este proyecto son:

Table 2: Entregables del proyecto integrador

|  |  |
| --- | --- |
| Entregables | Enlace |
| Presentación pública | Presentacion1.pptx |
| GitHub del proyecto | <https://github.com/hsepulvedaj/proyecto_integrador> |
| Reporte-técnico-y-modelos | ProyectoIntegrador\_descriptivo.ipynb  LDA.py  lda\_Quejas\_2\_v2.html  Mdl\_perfilado\_clientes\_tigoune.py  Mdl\_preliminar\_costa.ipynb  Mdl\_preliminar\_eje.ipynb  Mdl\_preliminar\_noroccidente.ipynb  Mdl\_preliminar\_oriente.ipynb  Mdl\_preliminar\_sut.ipynb  Falta centro |
| Planeación Proyecto | Proyecto.xlsx |
| Descripción/contexto del proyecto | Perfilado\_clientes.pdf |
| Producto Desplegado |  |

## Preparación de datos

### 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.

### 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:

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

Una vez se tienen preparados los datos, se puede generar el modelo del LDA. HERNAN

## Modelos preliminares.

### Modelo LDA sobre las quejas

En el siguiente gráfico 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.

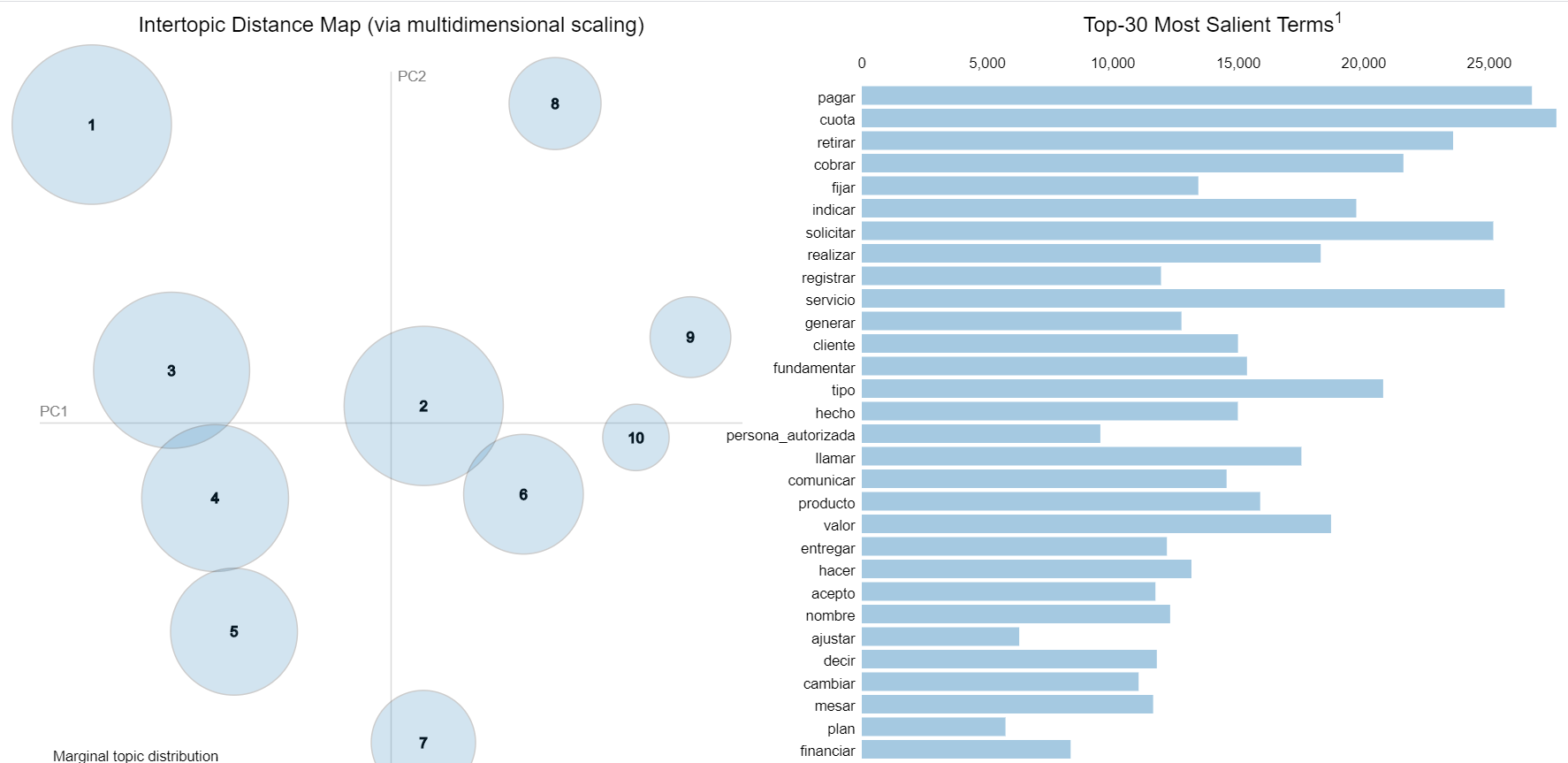
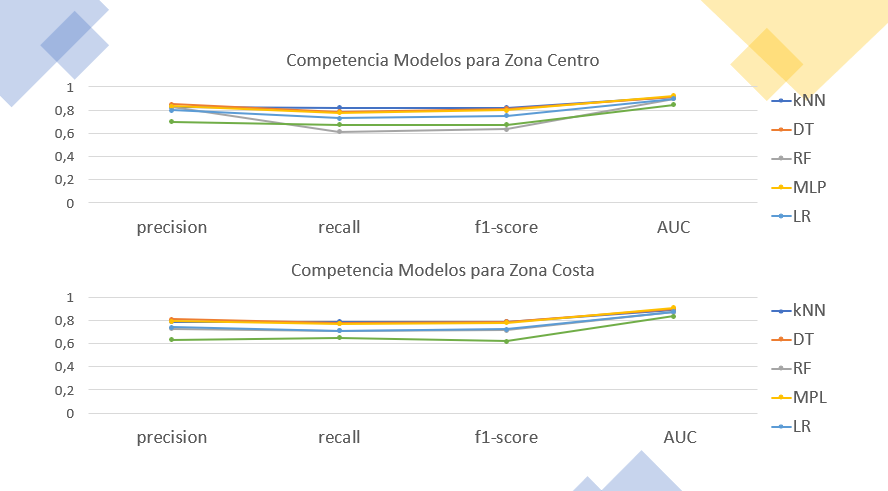


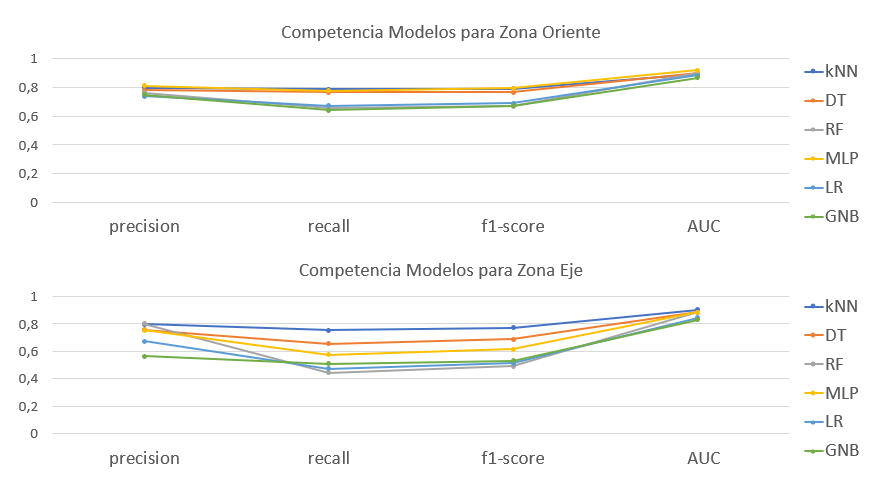
Figure : Modelo LDA

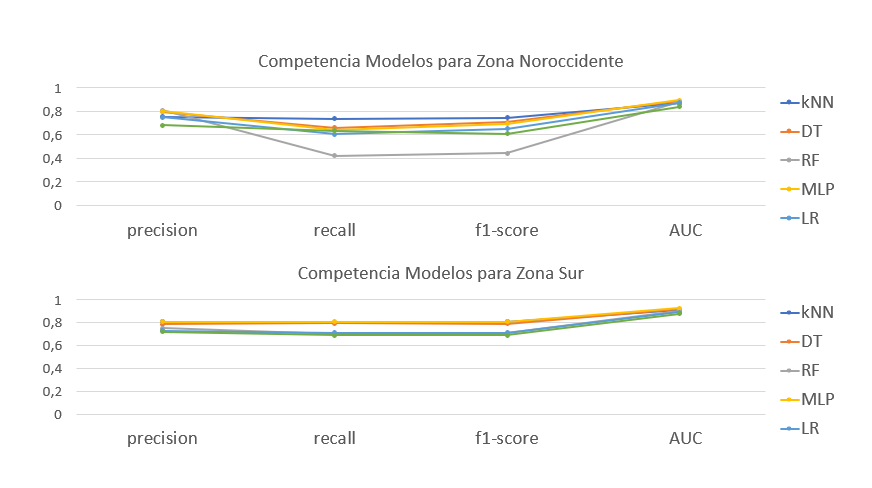
### Competencia de modelos de clasificación

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:

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

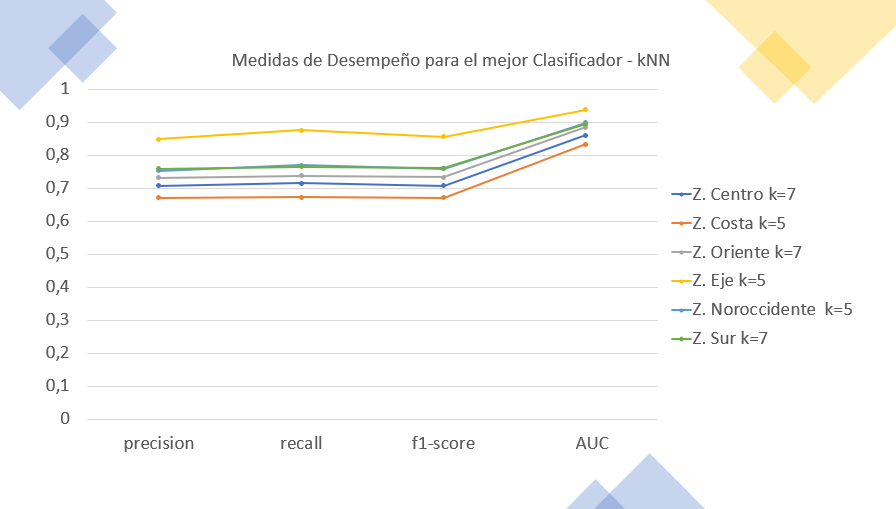






## Modelos finales validados.

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.



# Requerimientos tecnológicos

El proyecto se desarrollará localmente hasta construir unos modelos robustos que puedan ser migrados a la nube.

# 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 |

# Bibliografía