**PERFILADO DE CLIENTES POR HABITO 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)

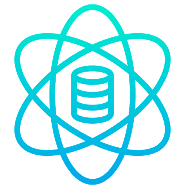
Hernán Sepúlveda Jiménez

[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](#_Toc42456941)

[2. Metodología: 4](#_Toc42456942)

[2.1. CRISP-DM 4](#_Toc42456943)

[3. Fuente de Datos 5](#_Toc42456944)

[3.1. Almacenamiento del Proyecto integrador 6](#_Toc42456945)

[3.2. ETL de los datos del proyecto integrador 6](#_Toc42456946)

[4. Entregables 7](#_Toc42456947)

[4.1. Definición del proyecto. 7](#_Toc42456948)

[4.2. Entendimiento del problema. 7](#_Toc42456949)

[4.3. Entendimiento de los datos. 7](#_Toc42456950)

[4.4. Entregables 7](#_Toc42456951)

[4.5. Preparación de datos 8](#_Toc42456952)

[4.5.1. Preprocesamiento 8](#_Toc42456953)

[4.5.2. Preparación de datos LDA 8](#_Toc42456954)

[4.6. Modelos preliminares. 9](#_Toc42456955)

[4.6.1. Descriptivo 9](#_Toc42456956)

[4.6.2. Modelo LDA sobre las quejas 10](#_Toc42456957)

[4.6.3. Competencia de modelos de clasificación 11](#_Toc42456958)

[4.7. Modelos finales validados. 11](#_Toc42456959)

[5. Requerimientos tecnológicos 12](#_Toc42456960)

[6. Fechas entregas 13](#_Toc42456961)

# Descripción 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.

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?

# 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.

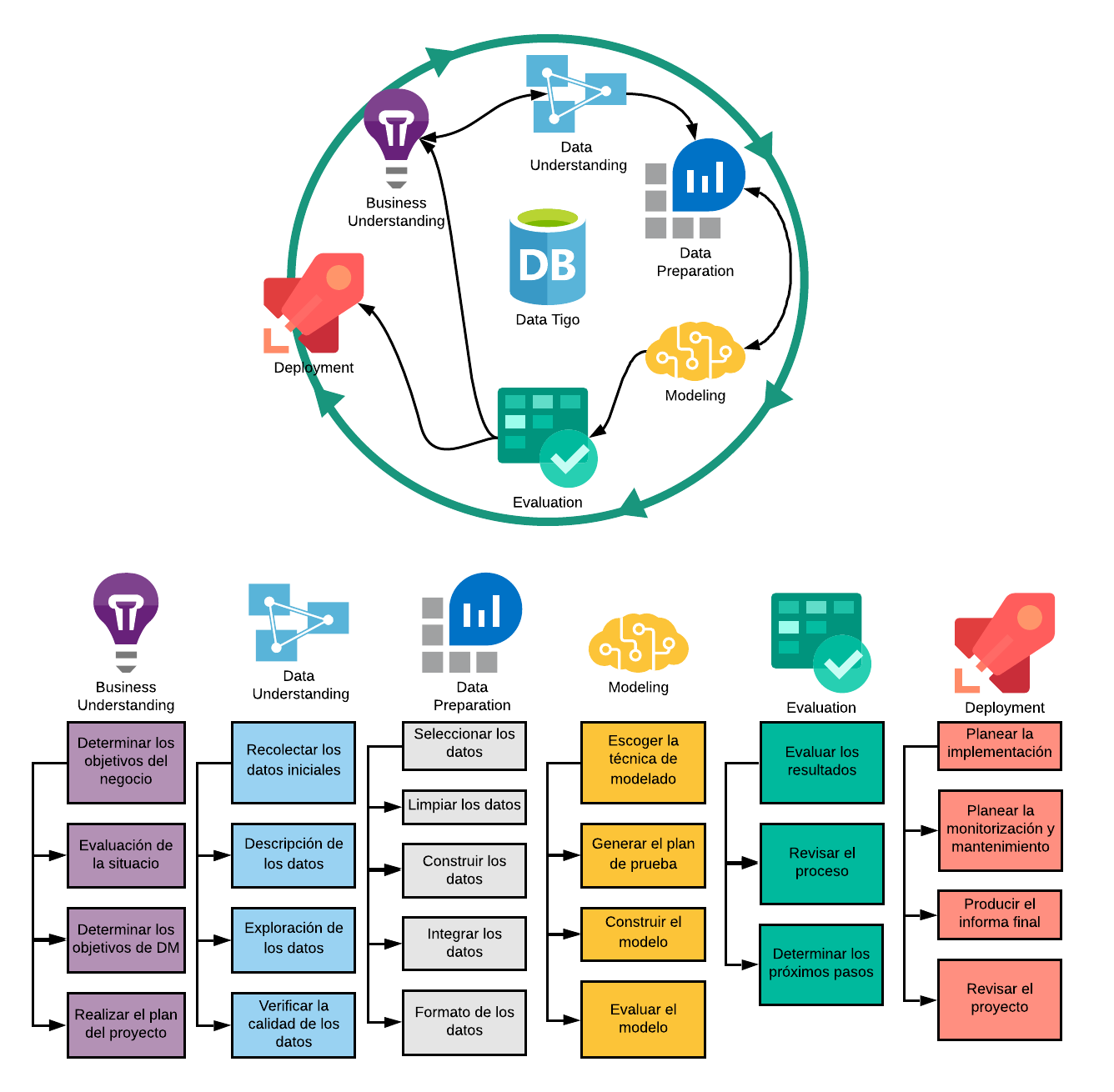


Figure 1: Modelo de CRISP-DM [Rodríguez, 2020]

# 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.

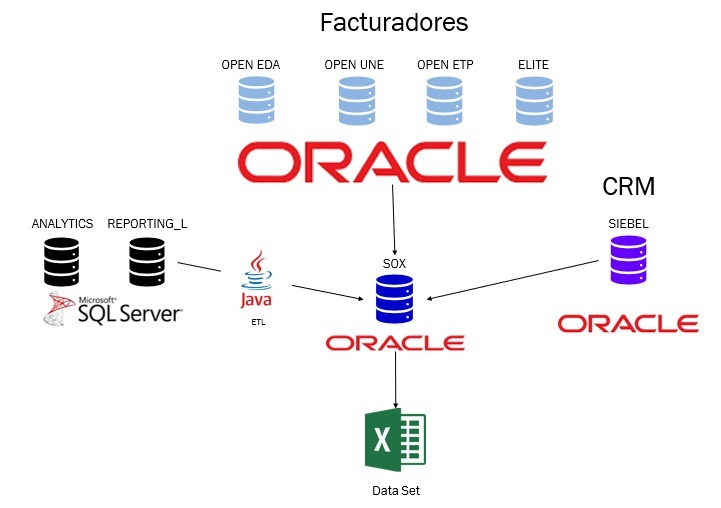


Figure 2: Fuente de datos. Recolección y almacenamiento

## Almacenamiento del Proyecto integrador

## ETL de los datos del proyecto integrador

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.

# 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:

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)

## 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 | LDA\_v1.py  Mdl\_perfilado\_clientes\_tigoune.py  ProyectoIntegrador\_descriptivo.ipynb |
| 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 semi-adecuada 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. Tokenizacion 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.

## Modelos preliminares.

### Descriptivo

Con el fin de conocer los datos con los cuales se van a realizar los trabajos, se realizó un análisis descriptivo de lo data sets de Vectores y de quejas del cliente.

Son de se pudo evidenciar que:

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%:

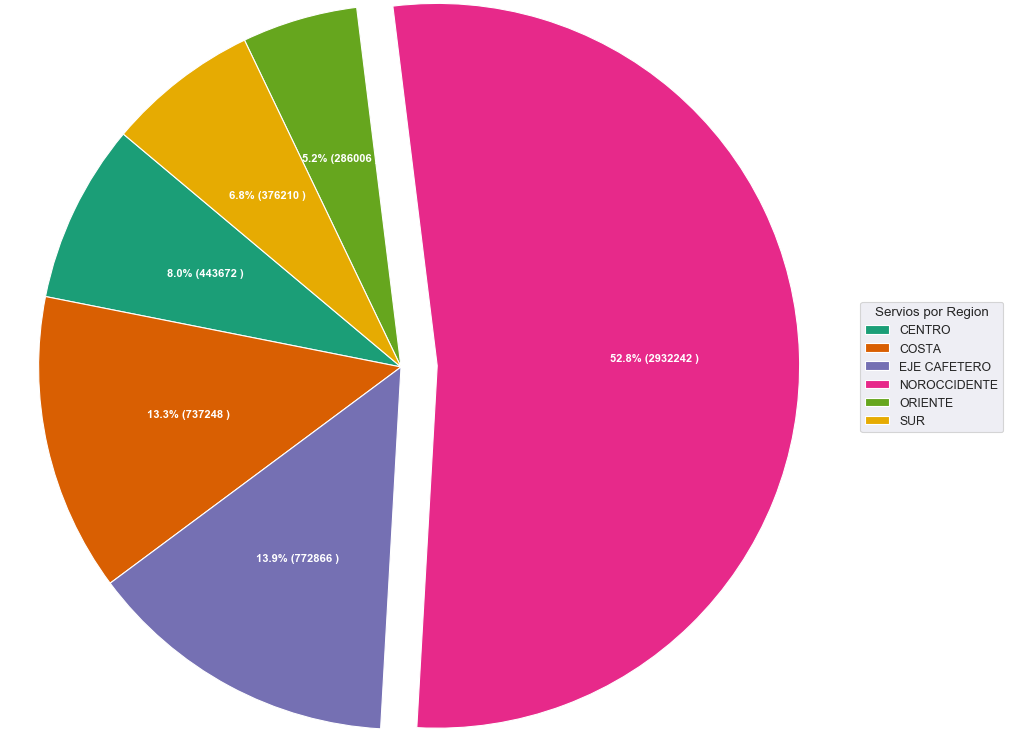


Figure 3: Regiones Tigo

Y según una clasificación del preprocesamiento, se determino el porcentaje de clientes con las variables de interés.

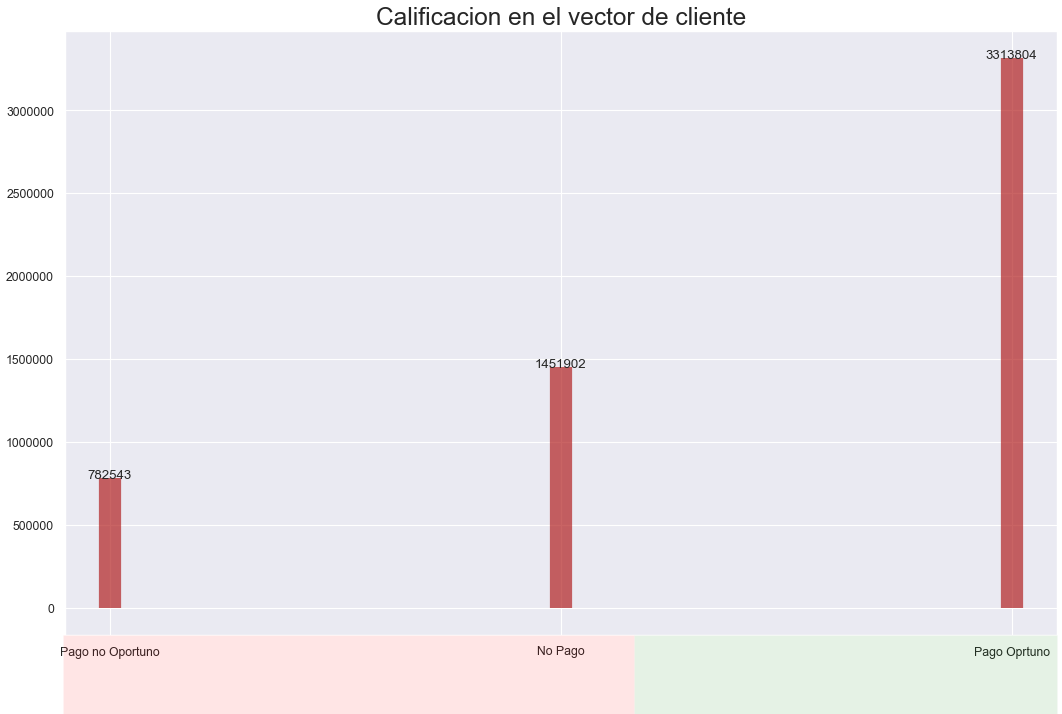


Figure : Comportamientos de pago q

### Modelo LDA sobre las quejas

En el siguiente grafico 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, no se encuentran superpuestas y no se están agrupadas en un solo cuadrante.

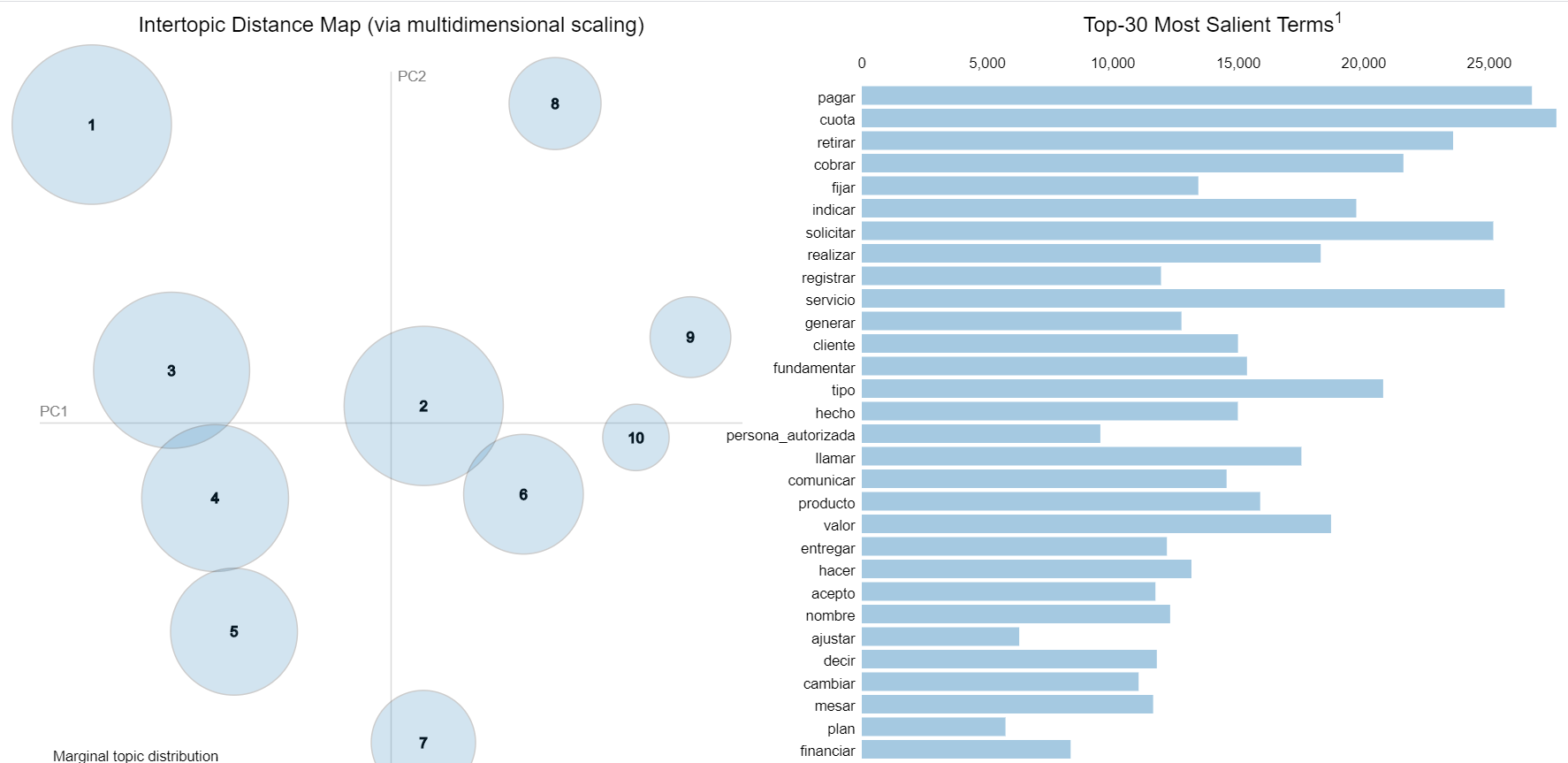


Figure : Modelo LDA

### 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 | Tarea | Descripción detallada | Fecha Inicial | Duración | Fecha Final |
| Definición | Definición del proyecto a presentar | Definición del proyecto a presentar | 2020-05-01 | 1 | 2020-05-02 |
| Definición | Envío del correo al PI | En el momento de definir su proyecto, deben anunciarlo en este medio (en la pestaña publicaciones nombrándome). Juan Carlos Monroy Osorio Enviar mensaje a Lalinde con copia a Juan Carlos Monroy, explicar el tema y los participantes del equipo | 2020-05-01 | 2 | 2020-05-03 |
| Definición | Propuesta - Documento | • <titulo> • <fuentes-datos> • <integrantes, emails> • <entregables> • <descripción-problema, contexto> • <fechas-entregas>  • <requerimientos-tecnológicos-o- definición-tecnología>  • <utilización-y-uso-de-una-metodología>: CRISP-DM o Microsoft-TDSP | 2020-05-10 | 11 | 2020-05-21 |
| Avance 1 | Definición del proyecto |  | 2020-05-10 | 15 | 2020-05-25 |
| Avance 1 | Entendimiento del problema |  | 2020-05-10 | 15 | 2020-05-25 |
| Avance 1 | Entendimiento de los datos |  | 2020-05-10 | 15 | 2020-05-25 |
| Avance 1 | Creación de presentación Publica | Entregable | 2020-05-10 | 15 | 2020-05-25 |
| Avance 1 | Creación proyecto en GitHub | Entregable | 2020-05-10 | 15 | 2020-05-25 |
| Avance 1 | Reporte técnico y modelos | Entregable | 2020-05-25 | 8 | 2020-06-02 |
| Avance 2 | Disposición tecnologías para el proyecto | Definir que método de almacenamiento, transferencia y seguridad utilizaremos | 2020-05-25 | 8 | 2020-06-02 |
| Avance 2 | Preparación de los datos | Transformaciones necesarias | 2020-05-25 | 8 | 2020-06-02 |
| Avance 2 | Modelos preliminares |  | 2020-05-25 | 8 | 2020-06-02 |
| Avance 2 | Creación de presentación Publica | Entregable | 2020-05-25 | 8 | 2020-06-02 |
| Avance 2 | Creación proyecto en GitHub | Entregable | 2020-05-25 | 8 | 2020-06-02 |
| Avance 2 | Reporte técnico y modelos | Entregable | 2020-05-25 | 8 | 2020-06-02 |
| Avance 3 | Modelos preliminares validos |  | 2020-06-02 | 12 | 2020-06-14 |
| Avance 3 | Creación de presentación Publica | Entregable | 2020-06-02 | 12 | 2020-06-14 |
| Avance 3 | Creación proyecto en GitHub | Entregable | 2020-06-02 | 12 | 2020-06-14 |
| Avance 3 | Reporte técnico y modelos | Entregable | 2020-06-02 | 12 | 2020-06-14 |
| Avance 3 | **producto-desplegado** | Entregable | 2020-06-02 | 12 | 2020-06-14 |
| Evaluación | Modelos finales validos |  | 2020-06-14 | 6 | 2020-06-20 |
| Evaluación | Creación de presentación Publica | Entregable | 2020-06-14 | 6 | 2020-06-20 |
| Evaluación | Creación proyecto en GitHub | Entregable | 2020-06-14 | 6 | 2020-06-20 |
| Evaluación | Reporte técnico y modelos | Entregable | 2020-06-14 | 6 | 2020-06-20 |