

Entrega Primer Sprint Proyecto

Ciencia de Datos Aplicada 2-2021

Integrantes:

Sara Fernanda Sánchez Sánchez sf.sanchez10@uniandes.edu.co

Rafael de Jesús Guzmán Martínez rguzmanm@uniandes.edu.co

Yamile Andrea Cepeda Chaparro ay.cepeda@uniandes.edu.co

Tabla de Contenido:

<i>Presentación de proyecto:</i>	1
<i>Entender el problema:</i>	2
<i>Ideación:</i>	2
<i>Implicaciones Éticas</i>	5
<i>Enfoque analítico:</i>	5
<i>Entender los datos</i>	5
<i>Entender los datos/producto</i>	5
<i>Conclusiones y primeras acciones.</i>	8

Presentación de proyecto:

Chiper es un startup colombiano, nació como proyecto del semillero de proyectos Ideamos, en el año 2017 y en el 2018 se materializa e inicia operación en Bogotá, hoy en día se encuentra en México, Colombia, y próximamente en Brasil.

Chiper es un servicio de abastecimiento para los negocios y tiendas de barrio, se encuentran todos los productos de la canasta básica familiar, licores, productos de aseo personal y del hogar y alimentos para mascotas, en Chiper se pueden hacer pedidos mínimos de 90.000 y el envío gratis, todo a través de la app todos los días en cualquier horario. Queremos, a través de este proyecto ayudarles a identificar sus dolores y entregarles una solución que les permita seguir con su crecimiento y expansión.

Definición de roles:

Andrea Cepeda, Líder de proyecto: Está a cargo de la gestión del proyecto. Define las fechas de reuniones, pre-entregables del grupo y verifica las asignaciones de tareas para que la carga sea equitativa. Se encarga de subir la entrega del grupo, si no hay consenso sobre algunas decisiones, tiene la última palabra.

Sara Sánchez, Líder de datos: Se encarga de gestionar los datos que se van a usar en el proyecto y de las asignaciones de tareas sobre datos. Debe dejarlos disponibles para todo el grupo, y puede definir los datos que se usarán para cada iteración. Adicional se encarga de la extracción y preprocesamiento de los datos.

Rafael Guzmán, Líder de negocio y líder de Analítica: Es responsable de velar por resolver el problema o la oportunidad identificada y estar alineado con la estrategia del negocio o las características del grupo para el cual se plantea el proyecto. Debe garantizar que el producto se puede comunicar de forma apropiada al contexto del negocio o de la comunidad. Adicional se encarga de velar por el cumplimiento de estándares de análisis en todos los entregables.

1. Entender el problema:

Idea: Crear una herramienta de predicción que permita a Chiper, identificar a los clientes en riesgo de dejar la empresa.

Problema: Chiper ha visto una disminución de sus clientes, los cuales no retiene bien, esto genera una pérdida de sus ingresos y desea conocer por qué está pasando esto y cómo puede fidelizar y retener a sus clientes.

Objetivos del proyecto:

- Crear una herramienta que a partir de la ciencia de datos apoye los procesos de fidelización de clientes.
- Identificar y segmentar los clientes con características similares.
- Predecir dadas ciertas características, los usuarios que abandonarán la empresa.
- Identificar las causas por las que un cliente decide disminuir y abandonar la alianza con Chiper.
- Prevenir el abandono por parte de los clientes.




Métricas:


- Número de clientes que abandonaron la empresa en el mes anterior
- % de participación de cada cliente en los ingresos del mes de Chiper.
- Frecuencia de compra de los clientes.
- Ubicaciones geográficas de los clientes que abandonan la empresa.
- Productos de mayor y menor consumo.
- % de los productos en los ingresos de la empresa.
- Número de campañas de fidelización generadas.
- Cantidad de campañas de fidelización por cliente.
- Número de reclamos por parte de los clientes agrupados por categorías.

2. Ideación:

En esta etapa incluimos personas o arquetipos de usuarios, sus dolores e ideas de solución, y seleccionamos una idea de producto y un prototipo de la solución esperada mediante un story board.

Arquetipos de usuarios:

PERSONA	COMPORTAMIENTOS Y ACCIONES	DETALLES DEMOGRÁFICOS	PUNTOS DE DOLOR
Juan José: Director Operaciones Chipper 	Juan José está estudiando una maestría en gerencia de Operaciones. Trabaja como Director de operaciones de Chipper, lleva aproximadamente 8 años. Sale los viernes a compartir con sus amigos y Camila su novia.	Tiene un salario de \$8 millones de pesos mensuales. No tiene hijos, es soltero, vive en Bogotá y tiene 35 años. Le gusta montar bicicleta.	No ha podido identificar por qué sus ventas se han venido reduciendo los últimos meses. No ha logrado fidelizar sus clientes, no los conoce. No ha podido lanzar campañas con resultados exitosos. Está preocupado por la cantidad de clientes que se desvinculan de la empresa. No sabe cuántos clientes se han ido realmente. No sabe cómo evitar el churn en la empresa. No sabe qué hacer para evitar que se vayan cuando informen que se van. No sabe qué les está ofreciendo la competencia.
Don Chepe: Tendero 	Don Chepe es un tendero muy reconocido del barrio Rionegro. Don Chepe abre su tienda todos los días a las seis de la mañana para atender a todos sus vecinos que madrugan y necesiten de alguno de los productos que él ofrece. Don Chepe mantiene su tienda abierta hasta las 8 pm, hora en la que cierra, cena y procede a descansar.	Tiene 56 años Vive con su esposa y un hijo de 26 años. Sus ingresos mensuales son de \$1.200.000. Vive en Bogotá Sufre de Diabetes Le gusta jugar Dominó, sale a caminar los sábados a las 5:30 am. Está a la espera de su pensión.	Los productos han incrementado su valor por lo que le ha tocado pedir menos productos. Hay muchas tiendas a su alrededor y eso ha dificultado mejorar sus ventas. Hay muchos almacenes grandes donde los precios son más bajos.
Doña Sandra Comprador 	Doña Sandra es una ama de casa A Doña Sandra le gusta cocinar, y mantener la casa ordenada mientras sus hijos estudian y su esposo trabaja como conductor de un bus de SITP. Le gusta trabajar con porcelamicrón en las tardes cuando tiene tiempo libre y les vende a sus amistades.	Tiene 50 años Vive con su esposo y dos hijos de 20 y 15 años Vive en Bogotá Su vivienda es estrato 3 Su familia recibe unos ingresos de \$1.500.000	Los productos han subido los últimos meses y ha visto cómo con el mismo dinero compra menos productos. No le alcanza para hacer el mercado para todo el mes por lo que acude a la tienda de barrio a comprar lo suficiente cada semana.

<p>Luis Fernando: Proveedor de Chipper</p> 	<p>Luis Fernando es el Gerente de Servicios de Alquilería Luis Fernando es el encargado de las relaciones con los clientes a los cuales distribuyen sus productos. Lleva cerca de 4 años en la compañía.</p>	<p>Tiene 40 años Vive con su esposa y un hijo de 2 años Vive en Chía Vive en casa propia estrato 3.</p>	<p>Cada vez son más los proveedores que ingresan al mercado. Están llegando marcas importadas a muy bajos precios. Se torna difícil retener a los clientes por los bajos precios que ofrece la competencia.</p>
---	--	---	---

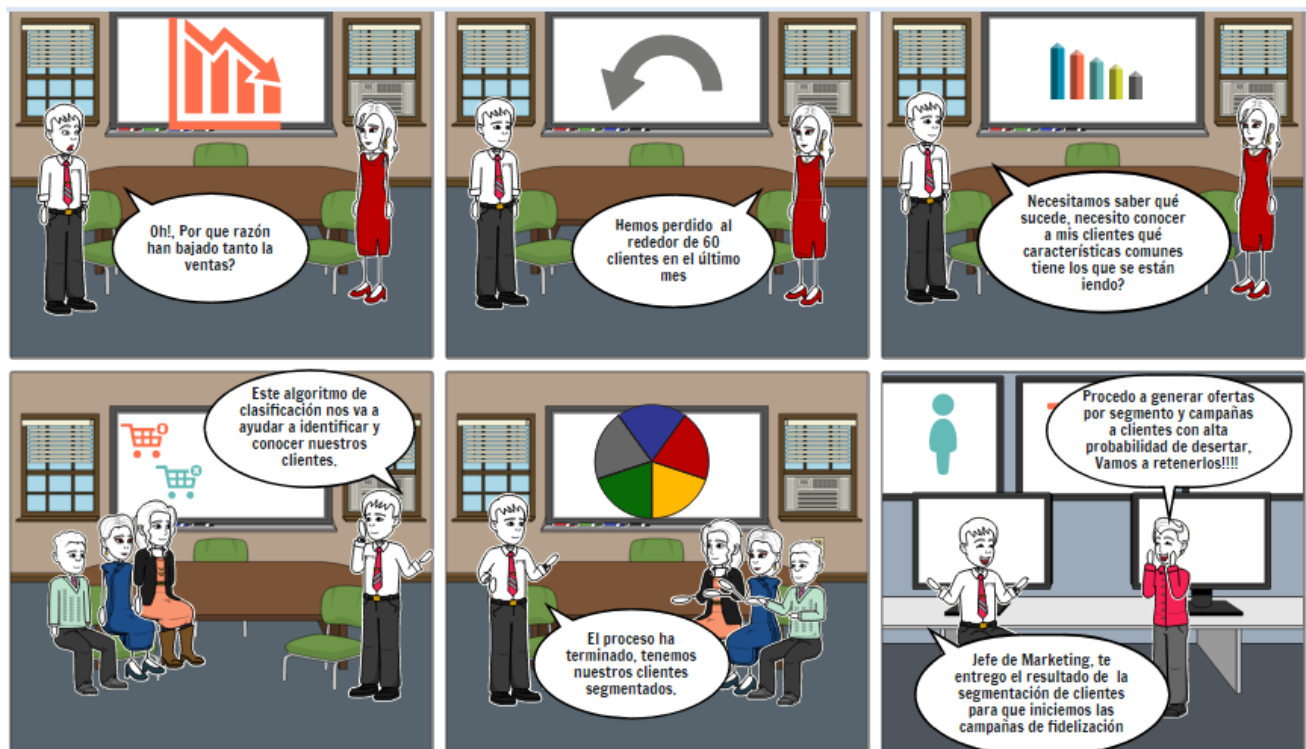
Ideas de Solución:

De los anteriores arquetipos escogimos el arquetipo del **director de Operaciones Chipper**.

La metodología que usamos fue una lluvia de ideas que plasmamos en el padlet (Ver **Anexo 1.**) con el fin de darle solución a los dolores que evidenciamos en la etapa anterior y ya con el arquetipo seleccionado se pidió al equipo de proyecto entregar un puntaje para cada idea y finalmente seleccionar la idea con mayor puntaje. Para esta calificación se tuvo en cuenta la viabilidad, que fuera de valor y muy útil al usuario. Luego del ejercicio, la idea seleccionada fue la número 14:

“Herramienta que permita segmentar los clientes de acuerdo con sus características similares, para poder realizar una predicción con base en éstas y así identificar qué tipo de clientes podrían desertar”

Story Board Para la Solución Planteada: Es importante aclarar que nuestro proyecto llegará hasta la segmentación, los siguientes actos, son actividades que debería seguir la empresa.



3. Implicaciones Éticas:

La compañía puede incurrir en procesos legales debido a la recolección de datos personales para uso predictivos y de big data. Las aplicaciones y/o herramientas que se implementarían para la solución del problema de negocio quedarían expuestas al uso no autorizado de los datos, porque no hay medio de poder asegurar la aprobación de uso. Adicionalmente puede haber un sesgo debido a la cultura organizacional de la compañía, dependiendo si es conservadora o no.

4. Enfoque analítico:

Hipótesis inicial H0: Los usuarios que realizan 1 o menos compras a la semana, son usuarios que desertarán en el siguiente periodo.

Hipótesis nula H1: Los usuarios que realizan 1 o menos compras a la semana, no desertarán.

Tipos de Análisis: Teniendo en cuenta la solución dada, el tipo de análisis predictivo que vamos a trabajar es de clasificación donde se segmentan los clientes en grupos por medio de sus características, de igual manera trabajaremos análisis predictivo de regresión con el fin de identificar los clientes en churn. Lo anterior nos permitirá organizar planes de acción con el fin de retener aquellos clientes con alta probabilidad de fuga.

5. Entender los datos:

Los datos que usaremos para el proyecto, corresponden a la Data Base recopilada por Chiper, allí se encuentran los datos referentes a toda la operación de Chiper, desde la información de proveedores hasta la de clientes.

Así pues, todos los datos proporcionados, corresponden a la búsqueda y descarga de la información en esta herramienta, la cual se alimenta a su vez de MySQL, donde reposa toda la información y datos del negocio. A su vez, los datos contenidos han sido procesados en más de una vez, buscando la uniformidad de los mismos en la empresa, por lo cual, podemos asegurar que contamos con información real, verídica y completa de los diferentes aspectos que queramos conocer. Sin embargo, es importante aclarar que todos los datos recolectados, se toman de la interacción del usuario con la aplicación de Chiper, desde allí se genera el perfil de usuarios, las órdenes de compra y las respectivas rutas.

6. Entender los datos/producto:

Análisis Descriptivo

Fuente de datos:

Como se mencionaba anteriormente, los datos son recolectados directamente por Chiper a través de su aplicación, cada interacción del cliente con la aplicación genera información y datos que son recolectados y guardados en la base de datos LB API MySQL. Dentro de los datos recolectados están datos personales, ubicación, tipo de negocio, productos a vender, etc. A partir de esta base de datos, se crea una manera amigable y colaborativa de participar en la edición y consulta de la información, por medio de Big Query y Looker, que son herramientas complementarias, para el análisis de datos.

Base de Datos:

Por una parte, Big Query presenta la información a modo de consultas de SQL dónde se tienen dos vistas o bases de datos para distintos procesos dentro de la empresa.

Por su parte Looker, permite realizar la analítica de los datos, a manera de gráficas, dimensiones y medidas, lo que posibilita el rápido entendimiento de los datos, tendencias y problemas. A su vez Looker se alimenta de MySQL, por lo cual la información es verídica y está en tiempo real.

Información:

Sobre la información, ésta la tomaremos directamente de Big Query como se mencionó anteriormente, se presenta mucha y variado tipo de información de todas las áreas operacionales de la empresa. Para nuestro caso, tomaremos la información del conjunto Silver y las tablas allí contenidas, en primera instancia StoreDetails y OrderProperties, que contienen la información necesaria para nuestro proyecto.

Una vez realizado un análisis exploratorio de los datos, en el perfilamiento realizado con pandas profiling, encontramos los siguientes datos relevantes de mención:

Verificando la información se notó que el campo de NIT, debe normalizarse porque hay registros en dónde éste está vacío.



Imagen 1: Pandas Profiling NIT

Adicionalmente, notamos que la información de pago nos va a ser muy útil a la hora de detectar anomalías o comportamientos de deserción, donde el gran porcentaje se sesga sobre el pago efectivo contra entrega, este dato fue necesario codificarlo con el fin de poder procesar la información.



Imagen 2: Pandas Profiling PaymentMethod

Además, la ubicación de los puntos se centra en FPI Vallejo en México, lo cual resulta curioso entendiendo que Chiper es una compañía con cubrimiento nacional e internacional, pero con mayor ubicación en Colombia. Esto nos guiará a realizarnos preguntas como participación del mercado de Chiper.

warehouse

Categorical

HIGH CORRELATION

HIGH CORRELATION

Distinct	8
Distinct (%)	< 0.1%
Missing	52
Missing (%)	< 0.1%
Memory size	1.1 MiB

FPI Vallejo	73047
CEDI Cali	42486
CEDI Medellín	31436
CEDI Álamos	2249
CEDI Bogotá	645
Other values (3)	85

Imagen 3: Pandas Profiling Warehouse

Además de las anteriores observaciones, realizamos los siguientes hallazgos, de acuerdo a las gráficas generadas:

```
## Diagramas un box plot
#Revisar los outliers(datos atípicos de la distribución de longitud del producto)
fig = px.box(data, y="city")
fig.show()
```

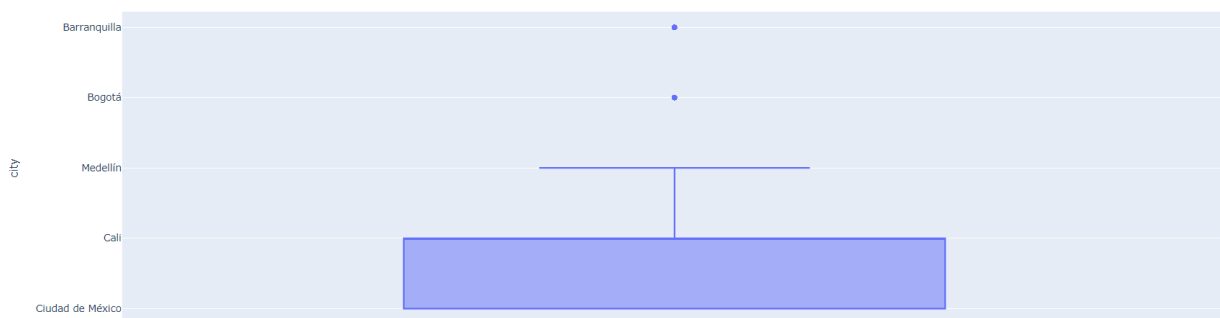


Imagen 4: Diagrama de caja

El diagrama nos indica que la mayor cantidad de bodegas se ubican en Ciudad de México, mientras que, en Medellín, Cali, Bogotá y Barranquilla, se encuentra una distribución normal de bodegas.

```
[ ] # Podemos utilizar la tabla de frecuencias para realizar el gráfico
data_raza = data['paymentMethod'].value_counts().reset_index()
# cambiamos el nombre de las columnas
data_raza.columns=['paymentMethod','conteo']
fig = px.bar(data_raza, x='paymentMethod', y='conteo')
fig.show()
```

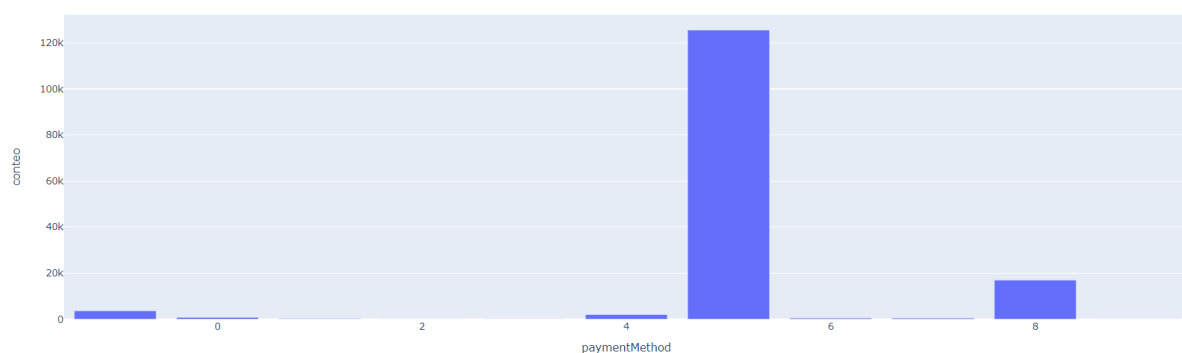


Imagen 5: Gráfico de columnas

Hemos codificado los datos de PaymentMethod con el fin de realizar un buen procesamiento de datos, evidenciamos que el método de pago por excelencia es el número 5, el cual corresponde a Pago Efectivo Contraentrega.

Diccionario de Datos - Payment Method		
#	Descripción	Código
1	Con mi cupo	0
2	Consignación	1
3	Megared	2
4	Pago con Nequi	3
5	Pago con tarjeta contraentrega	4
6	Pago efectivo contraentrega	5
7	Pago punto de recaudo	6
8	PayU	7
9	Tarjeta Chiper	8
10	Transferencia Bancolombia	9
11	No hay registro	-1

Se realizó un análisis descriptivo de la información por medio del notebook que se encuentra en la ruta: <https://github.com/SanchezSara/CienciaDatos>.

7. Conclusiones y primeras acciones:

Se concluye que para este primer sprint se entendió la dimensión de la información a procesar para la solución de negocio. Esta dimensión es la que a manera general canaliza el flujo a seguir sobre la propuesta de la herramienta de predicción a construir.

Seguidamente se identificó que hay un gran porcentaje de sesgo en cuanto a que el proyecto sea implementado productivamente o no. Eso depende mucho de la cultura organizacional de la compañía y de su resistencia al cambio. Cabe aclarar que para proyectos de ciencia de datos y asegurar el éxito empresarial, es imprescindible maximizar el retorno de la inversión existente en la herramienta, a la vez que innovar y adoptar nuevas tecnologías, ya que todo se verá reflejado en los beneficios económicos y/o financieros de la compañía.

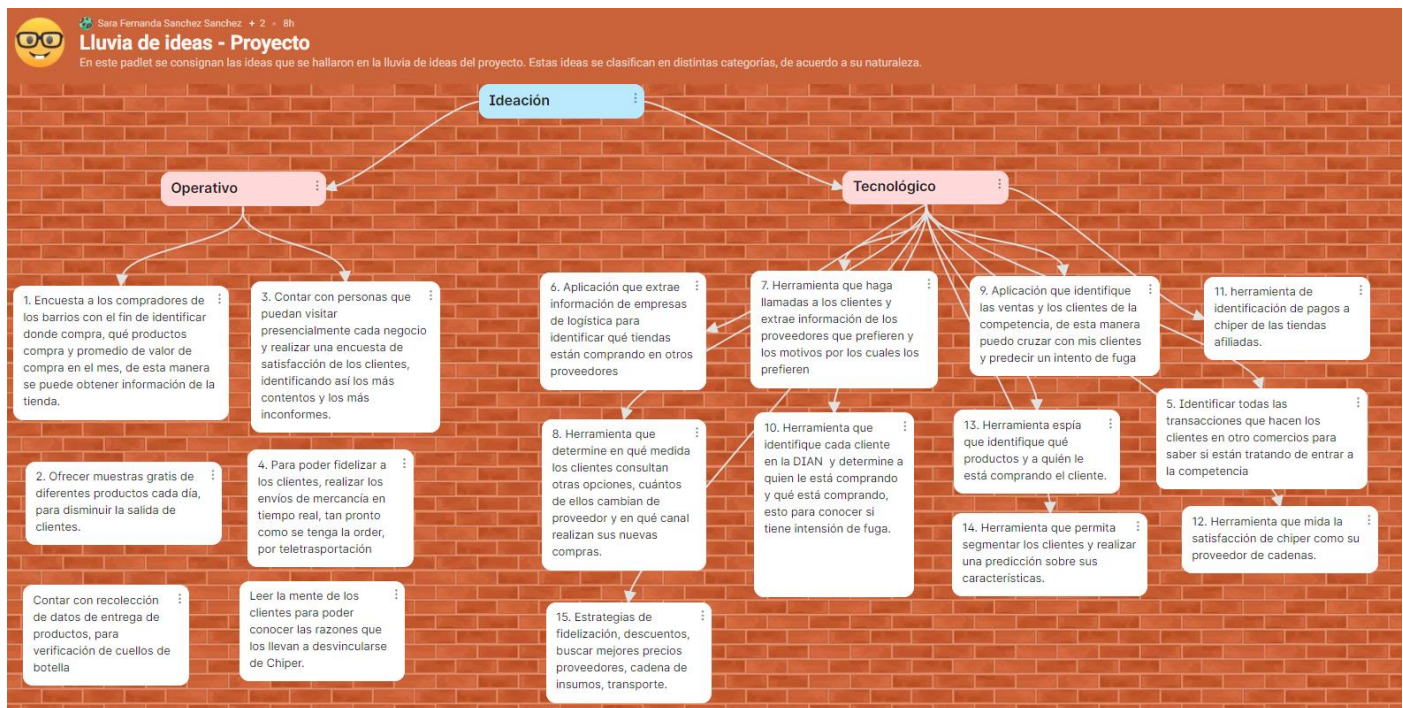
Por otro lado, se analizó que la compañía tiene medianamente organizada la información con la cual maneja el negocio, lo cual es muy efectivo a la hora de realizar proyectos de predicción de datos, pues la calidad del dato entregado es alto y por tanto habrá mejores resultados y coherentes.

Los datos que son relevantes de todo el universo serían NIT, método de pago, ciudad, país, monto, frecuencia de compra por cliente y fecha.

Finalmente, las acciones sugeridas son las siguientes:

- Solicitar a través de formatos de confidencialidad la información de los datasets que se implementarían en la solución.
- Iniciar luego la depuración de la información u/o solicitud de información adicional para complementar el universo de datos que se implementaría en la predicción.
- Escoger modelo adecuado detallado de solución de predicción.
- Comunicar a la junta directiva los resultados obtenidos del análisis inicial con tal de obtener feedback del mismo.

Anexos



Anexo 1: Ideaación

Padlet en: <https://uniandes.padlet.org/sfsanchez101/jh0hwdmmmln7qtdw>