

**Índice**

[**Introducción**](#_heading=h.klegz8queykc) **2**

**Metodología de trabajo 2**

**Marco de trabajo 2**

**Stakeholders 2**

**Equipo de trabajo y roles 3**

**Ingeniería de datos 4**

**Workflow 4**

**Escalabilidad 5**

[**Machine Learning**](#_heading=h.cfiv8eqwfoh3) **7**

**Modelo** [**Forecasting**](#_heading=h.nh2d9fjokpex) **7**

**Análisis de datos 9**

**Contexto 9**

**Producto Digital Final 12**

**Conclusión 12**

# Introducción

El presente informe es el último del proyecto, fruto del arduo trabajo que el equipo de Racont estuvo realizando durante casi 4 semanas. Logramos transformar los datos en información valiosa, que permitirá a Olist tener nuevas oportunidades de crecimiento para su empresa.

Este reporte tiene como objetivo presentar un resumen de todo el proceso de desarrollo, junto con un exhaustivo análisis de los datos y para finalizar una breve descripción del producto digital que el cliente se llevará, con el objetivo de que pueda medir el rendimiento y predecir las ventas de su negocio a lo largo del tiempo.

# 

# Metodología de trabajo

## Marco de trabajo:

Se utilizó Scrum como metodología ágil, organizado con cuatro Sprints de una semana de duración cada uno.

## Stakeholders

Los interesados identificados del proyecto son:

* Client: Olist
* Product Owner (PO): Juliana Aragón
* Scrum Master | Henry Mentor (HM): Leandro Franchina

## Equipo de trabajo y roles

|  | **Melina Griffo | PM y Data Engineer**  Experta en los procesos de ETL y desarrollo del backend en conjunción con las herramientas cloud. Project Manager, encargada de la dirección del proyecto. |
| --- | --- |
|  | **Isaac Peña | Data Engineer**  Experto en análisis de datos, procesos ETL y desarrollo del backend. |
|  | **Julio Postigo | Data Analyst**  Experiencia en análisis e interpretación de datos. Diseño y presentación de reportes. |
|  | **Emmanuel Corral | Analista Funcional**  Experiencia en procesos EDA y ETL. Especialista en interpretación y exposición de datos. |
|  | **Nicolás Lazarte | Data Scientist**  Científico de datos con experiencia en el desarrollo de modelos supervisados de Machine Learning. |

# Ingeniería de datos

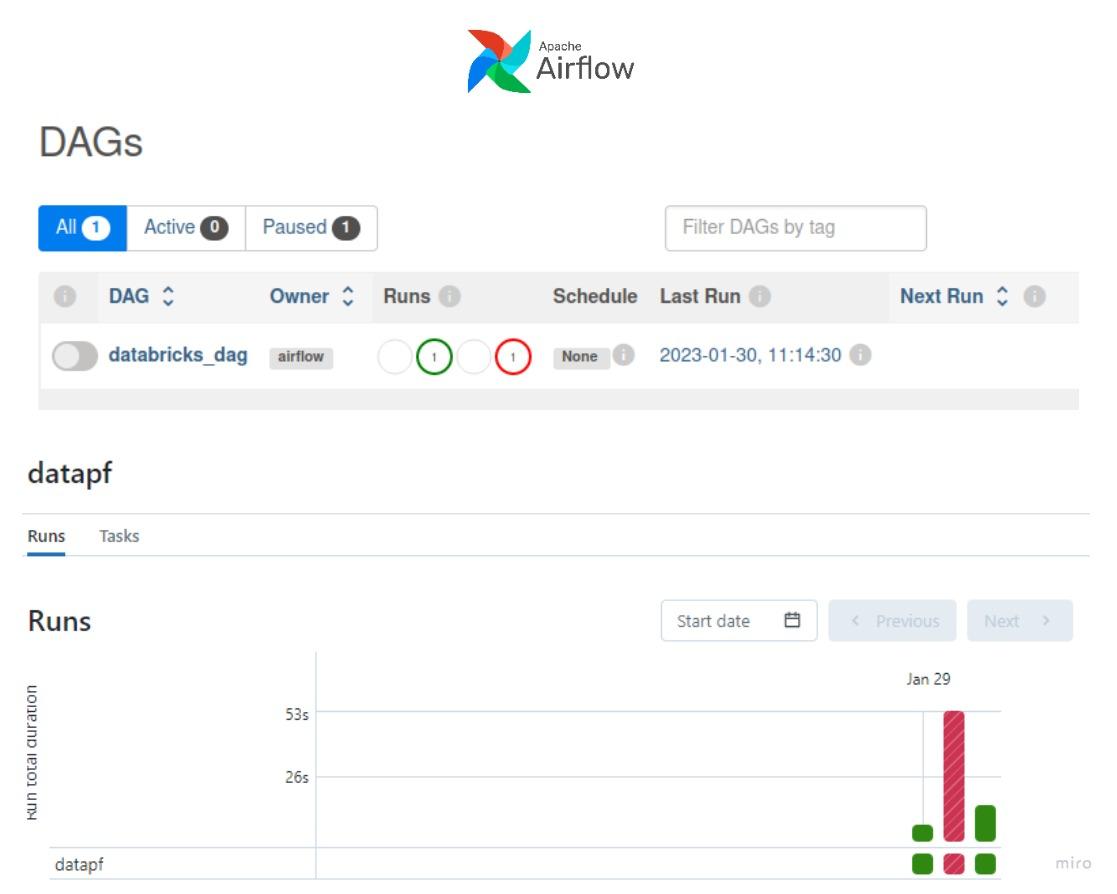
## Workflow - Ciclo de vida del dato

* Recolección de datos: Esta etapa se encarga de obtener los datos necesarios para el proyecto
* Limpieza de datos: Esta se encarga de identificar y corregir problemas en los datos, como valores faltantes, duplicados, outliers o valores atípicos, para garantizar que los datos sean precisos y confiables (***EDA***).
* Integración de datos: Esta etapa se encarga de combinar los datos de diferentes fuentes en una sola estructura de datos, utilizando técnicas la normalización.
* Transformación de datos: Esta etapa se encarga de modificar los datos para que cumplan con las necesidades específicas del proyecto (***Pipeline*** ***ETL***).
* Carga inicial de datos: Los datos limpios se almacenan en la base de datos elegida (***Data Warehouse***).
* Análisis de datos: utilizando técnicas estadísticas, de aprendizaje automático o de visualización de datos, para extraer insights y patrones interesantes.
* Presentación de resultados: Esta etapa se encarga de presentar los resultados del análisis de manera clara, concisa y accesible, utilizando herramientas de visualización de datos o informes.

# 

## Escalabilidad on Cloud

Al considerar la escalabilidad del proyecto, se decidió migrar la carga de datos al cloud (Azure) para aprovechar las ventajas de almacenamiento, procesamiento y disponibilidad que ofrece esta opción, para manejar una mayor cantidad de datos. El pipeline permite que se ejecute la carga incremental, con un gestión del flujo orquestado a través de Apache Airflow.

# Machine Learning

## Modelo Forecasting

Se realizaron tres modelos de Machine Learning utilizando la librería Prophet para el forecasting de ventas totales diarias:

* Modelo simple: Es el que por defecto la librería entrena y obtiene los hiper parámetros según la data histórica ingestada (01/01/2017 al 31/08/2018).
* Modelo simple con optimización de hiperparámetros: Se realiza una función con todas las combinaciones de valores de hiperparámetros que se creen relevantes obteniendose luego la combinación de los mismos que mejor resultado de sMAPE (Error cuadrático Medio absoluto simétrico) obtenga en una ventana de tiempo de 5 días para una predicción de 60 días.
* Modelo con días festivos (holidays) con optimización de hiperparámetros: De la misma manera se aplicó el entrenamiento y evaluación para las combinaciones de hiperparámetros dados, aunque en este caso sí se consideró `holidays\_prior\_scale`. En este modelo se pasó como hiperparámetros `holidays` que corresponde a las fechas obtenidas para determinado rango de tiempo en el que fueron fechas festivas, en este caso se utilizó la librería **holidays** y el método *holidays.country\_name(‘BR’)* para obtener las fechas de festivas en un rango de tiempo. Además se le agregaron los días de BlackFriday al análisis.

### Hiperparámetros

Se consideraron los más comunes a considerar para mejorar el modelo según la documentación de la librería. Entre ellos se encuentran:

* changepoint\_prior\_scale: modifica la cuan flexible es el modelo a la hora de considerar los cambios de tendencia en el corto plazo
* seasonality\_prior\_scale: explica la flexibilidad del modelo para ajustarse a los componentes estacionales, cuanto más alto es el valor más se ajustará a la estacionalidad
* holidays\_prior\_scale: de la misma manera que el anterior, hace referencia al ajuste a los periodos festivos
* weekly\_seasonality: Se tiene en cuenta la estacionalidad semanal a la hora de entrenar el modelo. Por defecto el modelo evalua por si solo si se tiene en cuenta o no.
* yearly\_seasonality: Igual al anterior hiperparámetro solo que para estacionalidad anual.

Para la optimización se consideraron el siguiente diccionario de hiperparámetros para la optimizar el modelo simple y con días festivos.

Consideraciones:

* El periódo de tiempo de predicción y evaluación se tomó en 60 días
* Se tomó una ventana temporal de 5 días, es decir que la métrica diaria es un promedio de los 5 días de ventana posteriores. Esto para evaluar también como se comporta en cuanto a la consistencia en la predicción
* Como métrica a evaluar se considera sMAPE que nos otorga un valor porcentual de error simétrico (mejora al MAPE simple) y se puede considerar como intervalo de confianza de que la predicción puede encontrarse en un intervalo por encima o por debajo de dicho porcentaje.

En la siguiente imagen se ve la lista de hiperparámetros considerados a evaluar.

param\_grid\_simple = {

'changepoint\_prior\_scale': [0.001, 0.01, 0.1, 0.5],

'seasonality\_prior\_scale': [0.01, 0.1, 1.0, 10.0],

#'holidays\_prior\_scale': [0.01, 0.1, 1.0, 10.0],

'weekly\_seasonality': [True,False],

'yearly\_seasonality': [True,False]

}

### Mejores hiperparámetros

Luego de realizar las hiteraciones se obtienen los mejores hiperparámetros para los modelos. Se observa que los valores de hiperparámetros de acuerdo al rango dado por la documentación son relativamente bajo, lo que nos dice que si bien considera las estacionalidades semanal, anual y con días festivos, los mejores modelos son mas flexibles y no se ajustan tanto a los datos.

Mejores parámetros para el modelo simple:

{

'changepoint\_prior\_scale': 0.1,

'seasonality\_prior\_scale': 0.01,

'weekly\_seasonality': True,

'yearly\_seasonality': True

}

Mejores parámetros para el modelo incluyendo holidays:

{

'changepoint\_prior\_scale': 0.1,

'seasonality\_prior\_scale': 0.01,

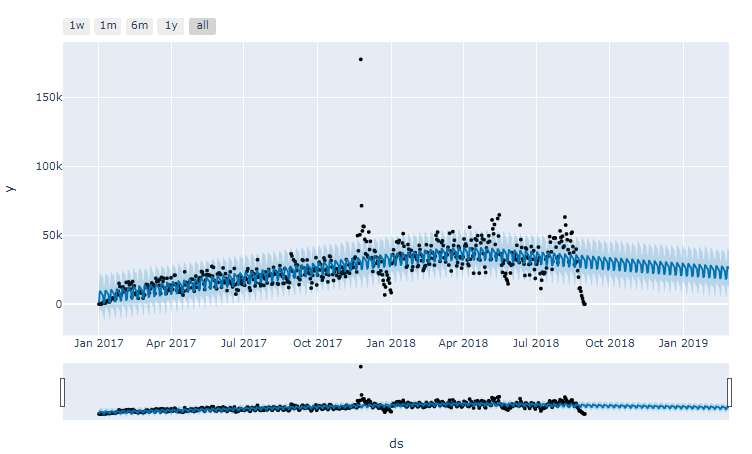
'holidays\_prior\_scale': 0.01,

'weekly\_seasonality': True,

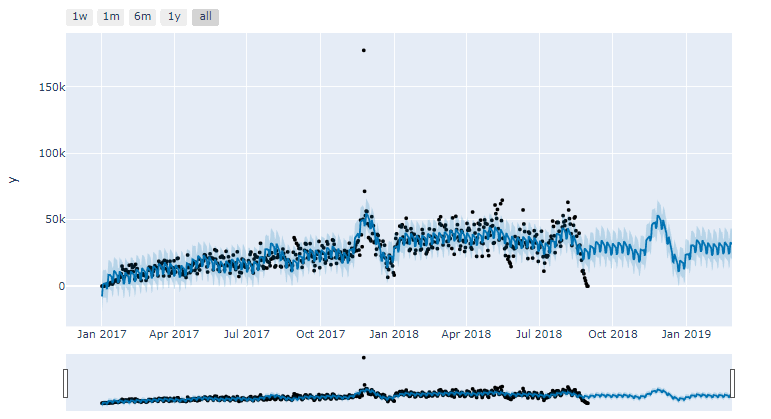
'yearly\_seasonality': True

}

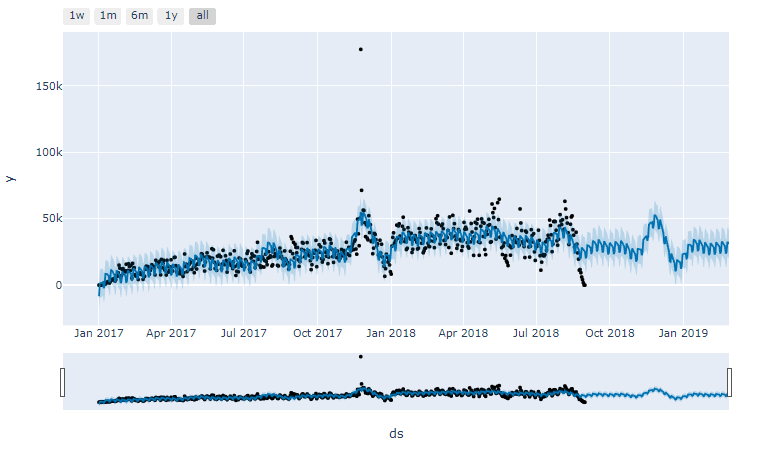
### Modelo simple por defecto



### Modelo simple optimizado



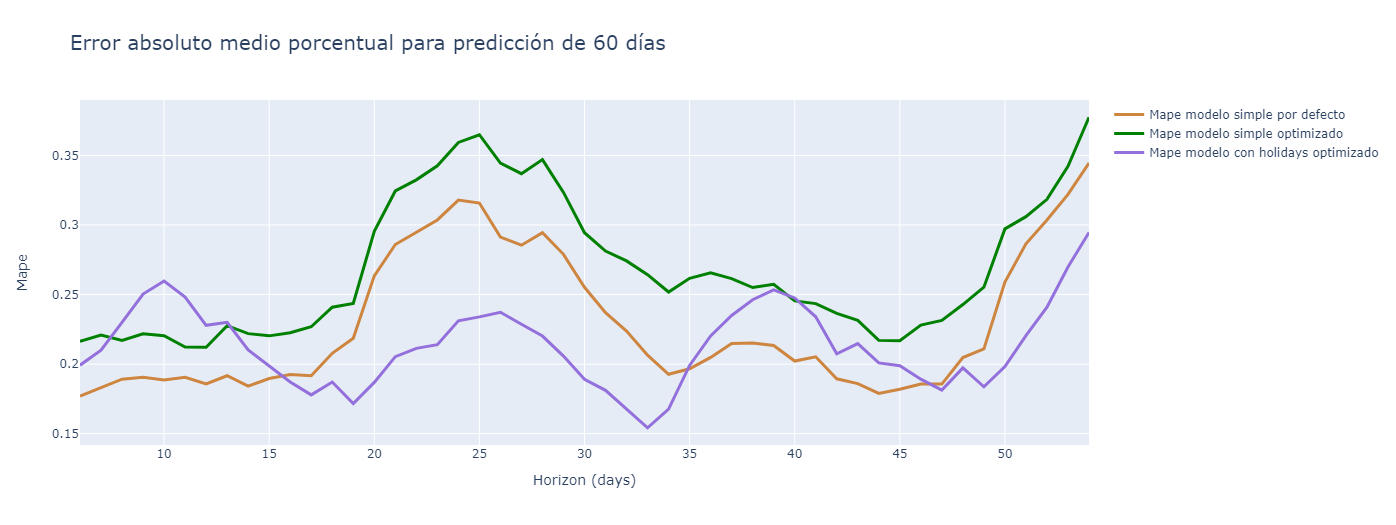
### Modelo con días festivos optimizado

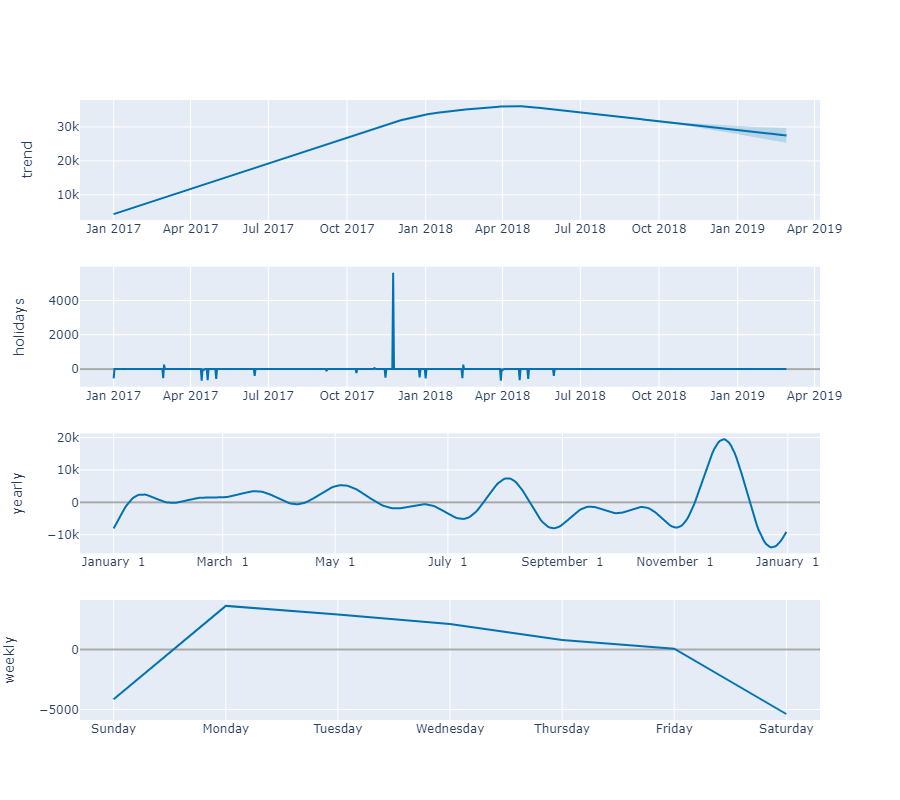


Vemos que las predicciones a simple vista son muy parecidas en los modelos optimizados, considerando un nuevo pico de ventas para noviembre de 2018.

## Comparación y elección del modelo

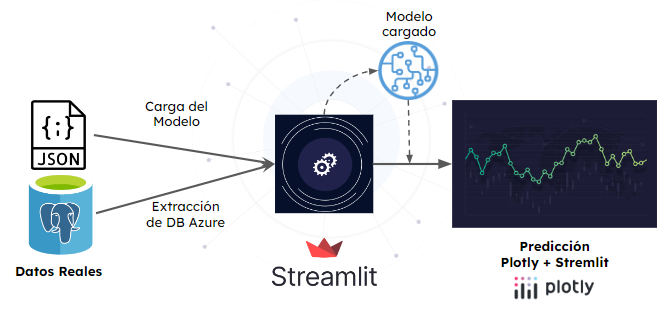
Para cada modelo se realiza un `cross validation` para evaluar la métrica de sMAPE con una predicción a futuro de 60 días desde el 01/09/2018.

Se puede observar que la métrica el mejor y más consistente a lo largo del tiempo para el caso del modelo considerando días festivos. Esto daría una mayor confianza en las predicciones por lo que se escoge dicho modelo.

En el siguiente gráfico se puede entonces evaluar la estacionalidad dada por la predicción del modelo elegido.

## Deploy del modelo optimizado

Una vez elegido el modelo, se lo guarda utilizando un serializador dado por la misma librería profet y se decide realizar el deploy del mismo en Streamlit para evaluar de manera interactiva las predicciones. El proceso es descripto a partir del siguiente esquema.



# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Análisis de datos

## Contexto

Ahora vamos a centrarnos en un punto de mucha importancia en todo lo que es e-commerce, que es la logística. El objetivo de la logística es acercar al vendedor y comprador con un solo ‘click’, salvando sus distancias físicas reales.

Para entender la logística hay que analizar la distribución geográfica de Brasil. Comenzaremos con el mapa que muestra la población por estados. Esta información es necesaria si queremos saber hacia dónde convendría que Olist se expandiera o fijara objetivos específicos por región.

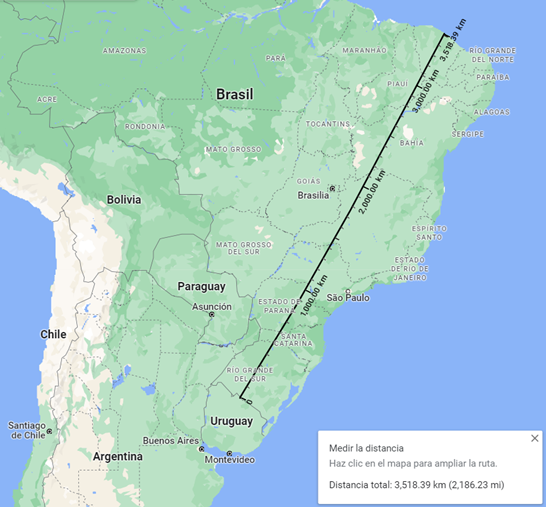


Brasil tiene una población estimada en más de 180 millones de habitantes, de los cuales casi 75 millones viven en los 3 principales estados del Sudeste (Sao Paulo, Minas Gerais y Rio de Janeiro).

Sin embargo, en el 4to lugar en densidad poblacional encontramos a Bahía, un estado del Nordeste. Y si analizamos un poco más, dentro de los 10 primeros estados con más población, encontramos 4 del Nordeste y 1 del Norte.

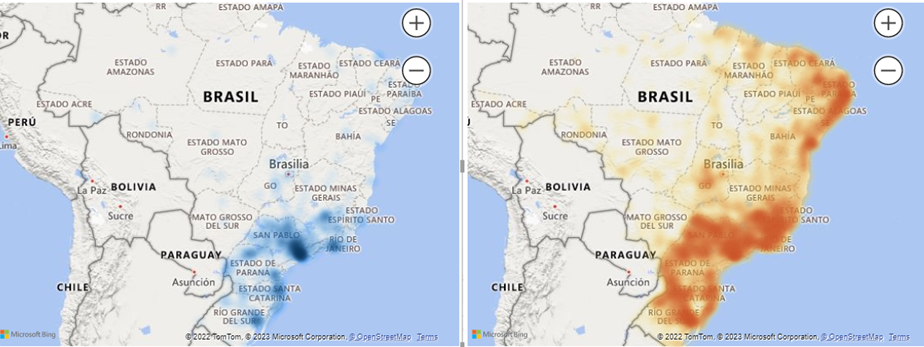


Ahora vamos a ver el siguiente mapa, que simplemente mide en línea recta la distancia en km entre un punto Sur y un punto Norte de Brasil.

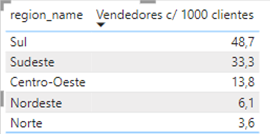


Tenemos la no menor distancia de 3518,39 km y solo en una línea recta, con lo cual podríamos decir que una logística no muy aceitada de la compañía podría resultar insatisfactorio para los clientes que ven demorados sus pedidos, por ejemplo, y esto, significa una desventaja frente a otras empresas que tienen mejor resuelto este punto.

Ahora veamos la distribución de vendedores y clientes que forman parte de Olist:



Podemos ver en azul los vendedores, en rojo anaranjado los clientes. Muy concentrados en ambos en el Sur y el Sudeste, pero cuando nos movemos hacia el Centro, Norte y Nordeste del país, la diferencia de distribución cambia significativamente entre unos y otros.



La tabla anterior, nos brinda mucha información. La diferencia de vendedores cada 1000 clientes es notoria, presentando un apartado en rojo en los ya mencionados estados de Norte y Nordeste.

Una vez establecido el contexto, podemos pasar a observar las medidas clave de rendimiento (KPIs) que fueron desarrolladas y analizadas dentro de un dashboard. Finalmente, también se detalla el link del informe detallado, con un análisis profundo de los datos recopilados y las conclusiones extraídas.

LINKS:

[Dashboard](https://app.powerbi.com/view?r=eyJrIjoiMDllNzkwNjYtYWVlOS00MzZlLTgxOGItYTA3ZTQ2ZTI2ODNiIiwidCI6ImRmODY3OWNkLWE4MGUtNDVkOC05OWFjLWM4M2VkN2ZmOTVhMCJ9&pageName=ReportSection3b8efcfb89bf1de353de)

[Reporte final](https://emm-ac-streamlitpf-productoolist-a2dn26.streamlit.app/)

# Producto Digital Final

Link al MVP

Funcionalidades de la plataforma:

* Home: página de inicio con una breve descripción del proyecto y de nuestra consultora (about us)
* Productos: 4 tarjetas que tienen links que nos llevan a diferentes páginas: Dashboard Modelo de ML Reporte del Análisis de datos Repositorio de Github
* Equipo de trabajo: integrantes y sus respectivos roles

# Conclusión

## La combinación de ingeniería y análisis de datos fue clave para el éxito del proyecto. A través del procesamiento y la manipulación de los datos, se pudo obtener una comprensión profunda de los patrones y tendencias presentes en los datos, lo que permitió tomar decisiones informadas para mejorar los procesos existentes e implementar soluciones eficientes y escalables.

Además de lo mencionado anteriormente, cabe destacar que durante el desarrollo del proyecto se encontraron algunos desafíos y obstáculos que fueron abordados con éxito, lo que nos permitió continuar avanzando hacia el objetivo final, reflejando la importancia de tener un enfoque integrado en la resolución de problemas.

Destacamos por último la combinación de habilidades, perspectivas y experiencias únicas de cada miembro del equipo que permitió una sinergia que no se hubiera logrado de otra manera. La unión y la comunicación constante fueron claves para resolver los desafíos y alcanzar los objetivos propuestos.