



**mercado  
libre**

FACULTAD DE  
INGENIERÍA Y CIENCIAS

**UAI**  
UNIVERSIDAD ADOLFO IBÁÑEZ

## **Informe Final: Rediseño del Proceso de Registro de Clientes de Inteligencia Competitiva en Mercadolibre**

Alumno: Cristóbal Salas Rodríguez  
Ingeniería Civil Industrial e Ingeniería Civil Informática  
Fecha: 28-11-2023

# Índice

<b>Resumen ejecutivo</b>	<b>3</b>
<b>Abstract</b>	<b>4</b>
<b>Introducción</b>	<b>5</b>
Contexto	5
Identificación de la oportunidad	6
Dolores y brechas existentes	8
<b>Objetivos</b>	<b>9</b>
Objetivo general SMART	9
Objetivos específicos	9
<b>Estado del arte</b>	<b>10</b>
<b>Propuestas de solución</b>	<b>12</b>
Elección de la solución	13
Riesgos y mitigaciones	15
<b>Evaluación Económica</b>	<b>16</b>
<b>Metodologías</b>	<b>18</b>
Rediseño iterativo del proceso y especialización de roles	18
Desarrollo del modelo de identificación de clientes	18
Análisis estadístico del modelo	18
Planificación	20
Plan de implementación	20
<b>Medidas de desempeño</b>	<b>22</b>
<b>Desarrollo e implementación del proyecto</b>	<b>23</b>
Definición de la situación As-Is	23
Primera iteración de rediseño del proceso	25
Segunda iteración de rediseño del proceso	27
Tercera iteración de rediseño del proceso	28
● Primer sprint de desarrollo del modelo	28
● Segundo sprint de desarrollo del modelo	29
● Tercer sprint de desarrollo del modelo	31
● Cuarto sprint de desarrollo del modelo	32
Cuarta iteración de rediseño del proceso	33
Implementación final	36
<b>Resultados</b>	<b>37</b>
<b>Conclusiones y discusión</b>	<b>40</b>
<b>Bibliografía</b>	<b>41</b>
<b>Anexos</b>	<b>44</b>

## **Resumen ejecutivo**

Este proyecto de pasantía realizado en Mercadolibre en el área de BI & Analytics de VIS (vehículos, inmuebles y servicios) tuvo su foco en el proyecto de Inteligencia Competitiva, donde a través de acciones que obtengan el interés de usuarios de la competencia en publicar en Mercadolibre, o que los usuarios actuales aumenten su pago por exposición, la empresa busca liderar el mercado de vehículos e inmuebles en todos los países donde opera VIS. Dentro de los ciclos trimestrales que desarrolla el proyecto de Inteligencia Competitiva se encontró una brecha entre el tiempo de vida de las publicaciones de los usuarios, menor a la duración del proceso de registro de clientes de 2 meses aproximadamente, generando a la hora del contacto que los usuarios potenciales ya no tengan interés en publicar en Mercadolibre. Además, se encontraron trabajos manuales de búsqueda de clientes que publican con diferentes nombres en la competencia por parte del área comercial dentro del subproceso de “Gestión Comercial” lo que significaba un costo monetario y de tiempo mayor al 80% en un análisis de Pareto de todos los subprocesos del registro de clientes. También, otra brecha se encontraba en el registro en el CRM de Mercadolibre, donde se presentaban errores de duplicación de cuentas. Todo lo anterior, generaba la oportunidad de agilizar el proceso de registro de clientes enfrentando a la problemática de pérdida de clientes, inversión y tiempo en el proceso de registro de clientes en Inteligencia Competitiva.

Por lo anterior es que se definió el objetivo SMART de “Disminuir el tiempo medio de desarrollo del proceso de registro de clientes a dos semanas en un plazo de 3 meses”. Según lo anterior, a través de conceptos de especialización de roles y BPM (Business Process Management), se planteó un desarrollo iterativo de rediseño del proceso en base a las metodologías Design Thinking y Lean Thinking, complementando esto con el desarrollo de un modelo de identificación de clientes basado en técnicas de NLP (Procesamiento del Lenguaje Natural) como vectorización de textos conocidos como embeddings desarrollada por modelos del transformer BERT, donde en conjunto con métricas de distancias se pueden realizar predicciones de probabilidad de coincidencias entre nombres no identificados de la competencia y Mercadolibre. El modelo se desarrolló siguiendo la metodología CRISP-DM de forma iterativa. Implementando el desarrollo aplicado anteriormente se consiguieron rendimientos del modelo de métricas como la exactitud y F1-Score del modelo mayores al 80%, donde posteriormente en la implementación se identificaron un 66% de los clientes no identificados en el armado de la base de datos. Además, con el rediseño de proceso se estima una duración de 12 días y un ahorro en costos de horas de trabajo de \$5.600 unidades monetarias, cumpliendo con el foco del objetivo SMART planteado y sus objetivos específicos, ayudando a Mercadolibre a agilizar un proceso crítico para el cumplimiento de su objetivo de liderar el mercado de vehículos e inmuebles en base a la captación de clientes.

## **Abstract**

This internship project carried out in Mercadolibre in the BI & Analytics area of VIS (vehicles, real estate and services) was focused on the Competitive Intelligence project, where through actions that obtain the interest of competing users in publishing on Mercadolibre, or that current users increase their payment for exposure, the company seeks to lead the vehicle and real estate market in all countries where VIS operates. Within the quarterly cycles developed by the Competitive Intelligence project, a gap was found between the lifetime of user publications, less than the duration of the customer registration process of approximately 2 months, generating at the time of contact that potential users are no longer interested in publishing on Mercadolibre. In addition, there were manual searches for customers who publish under different names in the competition by the commercial area within the "Commercial Management" sub-process, which meant a monetary and time cost of more than 80% in a Pareto analysis of all the sub-processes of customer registration. Also, another gap was found in the Mercadolibre CRM register, where there were errors of duplication of accounts. All of the above generated the opportunity to streamline the customer registration process, facing the problem of loss of customers, investment and time in the customer registration process in Competitive Intelligence.

Therefore, the SMART objective "To reduce the average development time of the customer registration process to two weeks within 3 months" was defined. According to the above, through the concepts of role specialization and BPM (Business Process Management), an iterative development of process redesign based on Design Thinking and Lean Thinking methodologies was proposed, complementing this with the development of a customer identification model based on NLP (Natural Language Processing) techniques such as text vectorisation known as embeddings developed by BERT transformer models, where, together with distance metrics, predictions can be made of the probability of matches between unidentified names of competitors and Mercadolibre. The model was developed following the CRISP-DM methodology in an iterative way. By implementing the previously applied development, model performance metrics such as accuracy and F1-Score of the model were achieved in excess of 80%, where later in the implementation 66% of unidentified customers were identified. In addition, the process redesign estimates a process duration of 12 days and a cost saving in working hours of \$5.600 monetary units, fulfilling the focus of the SMART objective and its specific objectives, helping Mercadolibre to streamline a critical process for the fulfillment of its objective of leading the vehicle and real estate market based on customer acquisition.

# Introducción

## Contexto

Este proyecto de pasantía se desarrolla en Mercadolibre (MELI), empresa multinacional de origen argentino dedicada al comercio electrónico con operaciones en gran parte de Latinoamérica. El área donde se desarrolla es BI & Analytics de VIS (Vehículos, Inmuebles y Servicios), encargada de realizar proyectos estratégicos y manejo de datos para habilitar proyectos de las otras áreas de VIS. El objetivo de VIS es liderar el mercado de vehículos y de inmuebles en toda latinoamérica, donde uno de los focos principales son los sellers o vendedores, ya que Mercadolibre monetiza la exposición en packs de publicaciones dentro de su plataforma, representando así el revenue que se consigue a través de las ventas.

Uno de los factores importantes en Mercadolibre para apuntar al objetivo de liderar los diferentes mercados de VIS es la captación de más clientes para que publiquen en su plataforma, y el proyecto fundamental que se encarga de atraer dichos clientes directamente desde los diferentes competidores del mercado es el proyecto de Inteligencia Competitiva. Este proyecto busca aplicar acciones de farming, que se define como acciones que consiguen upsellings o mayores pagos de los clientes actuales, y principalmente hunting, que se define como acciones para convertir a usuarios de la competencia en usuarios de MELI. Inteligencia Competitiva se desarrolla en ciclos por cada trimestre del año, implicando acciones de las áreas de BI, Operaciones, Planning y Comercial de VIS para ejecutar los procesos asociados al proyecto. Los datos de la competencia son recolectados por una empresa externa y son recibidos periódicamente.

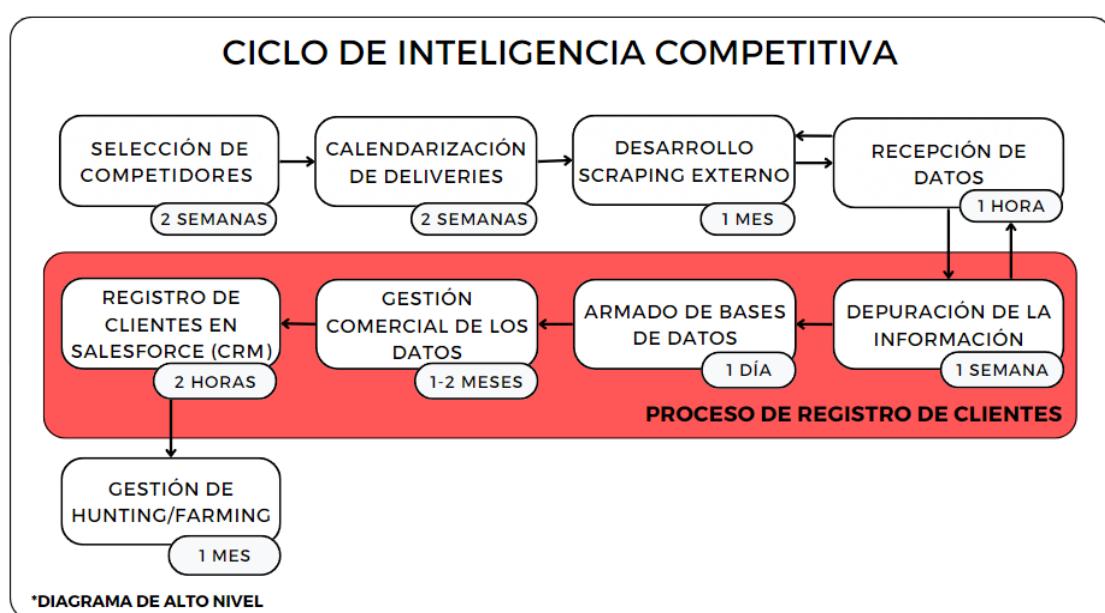


Figura 1: Diagrama de alto nivel de procesos de un ciclo de Inteligencia Competitiva

Los procesos de un ciclo de Inteligencia Competitiva se pueden ver en la figura 1, donde existen etapas estratégicas que van desde la selección de competidores por cada país hasta la calendarización de las entregas del proveedor. Luego, desde el desarrollo de un web scraping externo (extracción de información de sitios web mediante software) y la recepción de estos datos, se realizan acciones para cada país y competidor, donde se destaca el “Proceso de registro de clientes” y sus subprocessos.

Dentro del proceso anterior, en primer lugar, se tiene la “Depuración de la información”, donde el área de BI realiza una validación de los datos. Luego, se realiza el “Armado de bases de datos” a partir de la información recepcionada, definiendo usuarios actuales o potenciales, a través del cruce de datos personales de los usuarios de la competencia con la información registrada en Mercadolibre.

Posteriormente, se disponen las bases de datos al resto de áreas, donde el área comercial realiza la “Gestión comercial de los datos”, realizando una categorización de los datos según los objetivos de cada área ya sean de hunting o farming. Terminada la gestión comercial, el área de operaciones realiza el “Registro de clientes en Salesforce”, el CRM de Mercadolibre, que permite mantener un seguimiento de las oportunidades que se generan en base a los clientes, dando paso a las acciones de gestión de farming o hunting, para conseguir el interés de los usuarios en publicar en Mercadolibre o aumentar la exposición actual de sus publicaciones, convirtiéndose en altas.

Luego de explicar sus procesos, en cuanto a su impacto, Inteligencia Competitiva es uno de los proyectos estratégicos más importantes en VIS por varios factores además de poder liderar el mercado de vehículos e inmuebles. En términos económicos, se tiene una inversión de 20.000 UM (unidades monetarias) por la obtención de la información del ciclo actual, además de haber obtenido un revenue de 27.500 UM aprox. hasta el tercer trimestre de este año. En cuanto a demanda, entre los sellers actuales de VIS aproximadamente un 25% han sido obtenidos por los ciclos desarrollados en el proyecto, y el winrate de conversión de contactos, es decir la razón de clientes que se convierten en altas sobre todos los clientes contactados actualmente es de un 19,9%, donde también se destaca un ARPU (revenue promedio por usuario) mensual actual de 74 unidades monetarias, alcanzando así el revenue mencionado.

## **Identificación de la oportunidad**

La oportunidad recae en la agilización de un ciclo de Inteligencia Competitiva para enfrentar la pérdida de recursos y tiempos de desarrollo que se están siendo destinados actualmente al proyecto, puesto que hay ocasiones donde se contactan clientes que ya han vendido sus publicaciones a la hora de hacer el contacto.

La brecha anterior se puede medir mediante la estimación del ciclo de vida de una publicación, donde en vehículos se tiene una media de 41 días (con un coeficiente de variación del 12,4%) y en arriendo de inmuebles de 49 días (con un coeficiente de variación del 15,5%), contrastando dicho valor con la duración del proceso de registro de clientes, siendo este último mayor al tiempo medio de vida de una publicación.

Para profundizar en el análisis del tiempo de desarrollo se realizó el siguiente análisis de Pareto, que nos permite realizar un análisis de prioridad de acciones frente al registro de clientes y los costos de sus subprocesos sujetos a este.

Costo por subprocesso del registro de clientes					
Proceso	Subprocesos	Tiempo (días)	Porcentaje (días)	Costo medio (UM)	Porcentaje (UM)
Registro de clientes	Depuración de la información	7	13,2%	\$616	9,2%
	Armado de bases de datos	1	1,8%	\$88	1,3%
	Gestión comercial de los datos	45	84,8%	\$6.000	89,1%
	Registro de clientes en Salesforce	0,08	0,2%	\$30	0,4%

Figura 2: Diagrama de Pareto de costo por subprocesso del registro de clientes

### Tiempo medio por subprocesso del registro de clientes

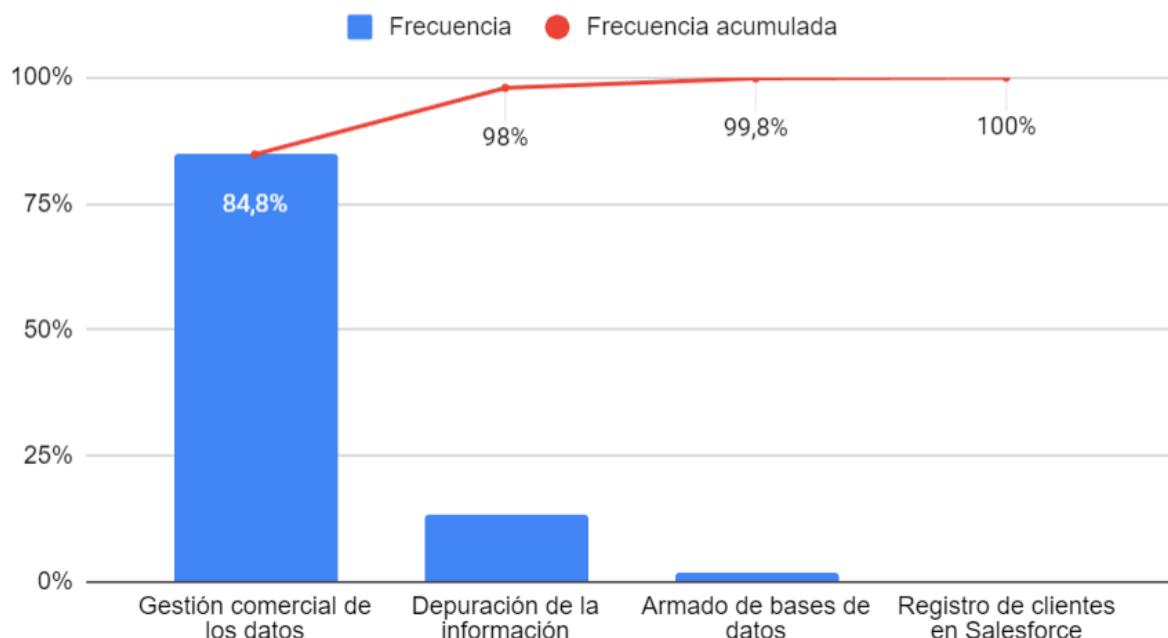


Figura 3: Gráfico de frecuencia de tiempo medio por subprocesso del registro de clientes

### Costo medio (UM) por subprocesso del registro de clientes

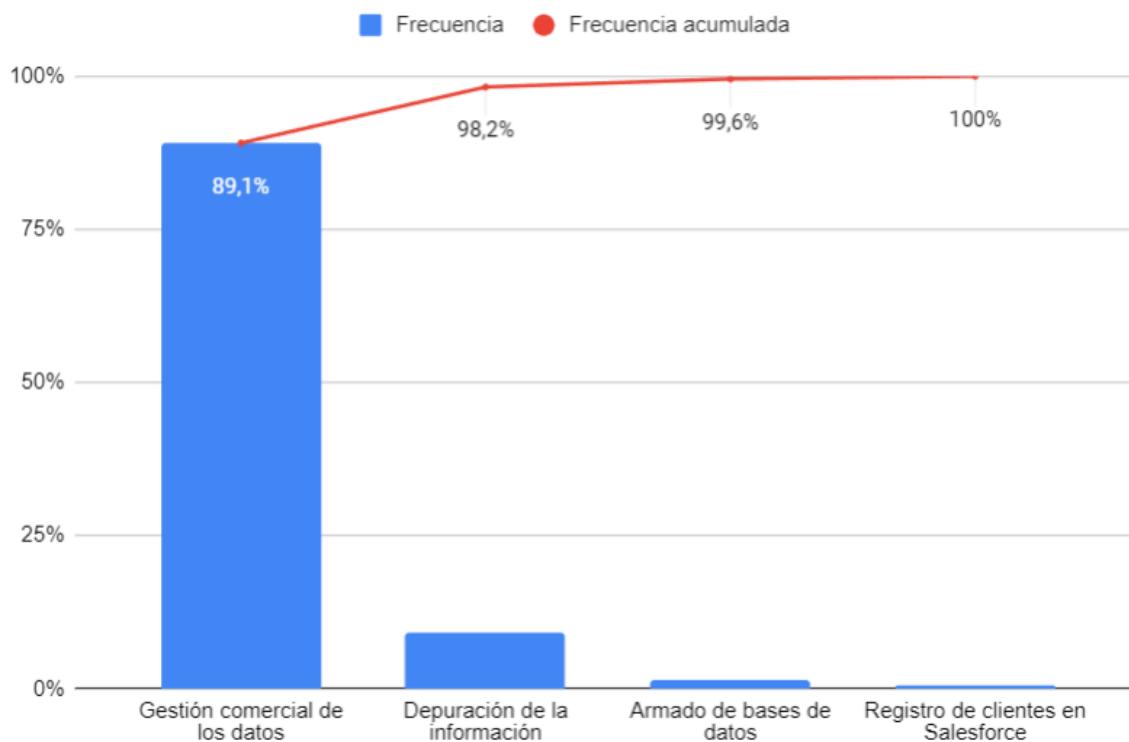


Figura 4: Gráfico de frecuencia de costo medio (UM) por subprocesso del registro de clientes

En la figura 2, 3 y 4 podemos ver a profundidad que la mayoría del tiempo de desarrollo y el mayor costo medio está en la “Gestión comercial de los datos”, donde el área comercial realiza una manipulación de datos según características y filtros de las publicaciones o clientes potenciales, con un 84,77% y 89,1% respectivamente.

Por lo anterior, se profundizó el análisis en dicho subprocesso, donde su tiempo de desarrollo se puede explicar por problemáticas encontradas en su operación, donde en primer lugar, al no identificar a clientes que publican con distintos nombres en la competencia, se llegan a realizar web scrapings manuales de los clientes, es decir, el área comercial realiza una verificación manual de coincidencia de clientes que publican con otro nombre en la competencia, sumando esto al tiempo de filtrado en el que se profundizará más adelante.

También, otro eslabón del registro de clientes que impacta en la calidad de los datos, es que en el registro en Salesforce cuando existen registros duplicados en la información a cargar o cuentas ya existentes en la plataforma, se presentan errores en cerca del 30% de los clientes en el registro, no tomando en cuenta estos clientes en el ciclo de Inteligencia Competitiva, reduciendo la cantidad de clientes a contactar y así, las posibles conversiones o altas.

## Dolores y brechas existentes

Los principales dolores y brechas se conforman a partir de la oportunidad de agilización del proceso de registro de clientes, con el objetivo de captar clientes para que publiquen en la competencia, donde es importante la calidad de ésta en términos de una contactación a tiempo y reducir el costo monetario y el tiempo de desarrollo de este. Para lo anterior, se tiene el análisis de ciclo de vida de las publicaciones y el análisis de Pareto desarrollado anteriormente, sin embargo, se puede profundizar aún más para comprender las causas y problemáticas asociadas a los costos del proceso, las cuales se describen a continuación.

Dentro de la gestión comercial de los datos, tenemos que los equipos comerciales realizan la categorización de datos entre dos analistas comerciales, destinando 1,5 horas por día cada uno durante un mes. Este esfuerzo corresponde a 60 horas de trabajo en total, con un costo aproximado de 780 UM (unidades monetarias), donde según estimaciones de Mercadolibre el costo empresa mensual de un analista es aproximadamente 2.300 UM. Por otro lado, el web scraping manual es dividido entre 10 analistas comerciales para 10.000 clientes, suponiendo 2 horas diarias de cada analista durante un mes, representando 400 horas de trabajo en total, que equivalen a 5.200 UM de costo aproximado por dicho esfuerzo, el cual es más del doble del gasto mensual de la empresa en un analista comercial.

Otra brecha a destacar es respecto al registro de clientes en Salesforce, donde se tiene que en el último registro realizado se encontraron un 33% de usuarios que presentaron error en el registro, afectando el tiempo de procesamiento en dicho porcentaje y haciendo que dichos usuarios no sean tomados en cuenta en el proceso de registro.

Para resumir lo comentado, se presenta el siguiente diagrama de Ishikawa representando gráficamente las causas y fallas explicadas anteriormente.

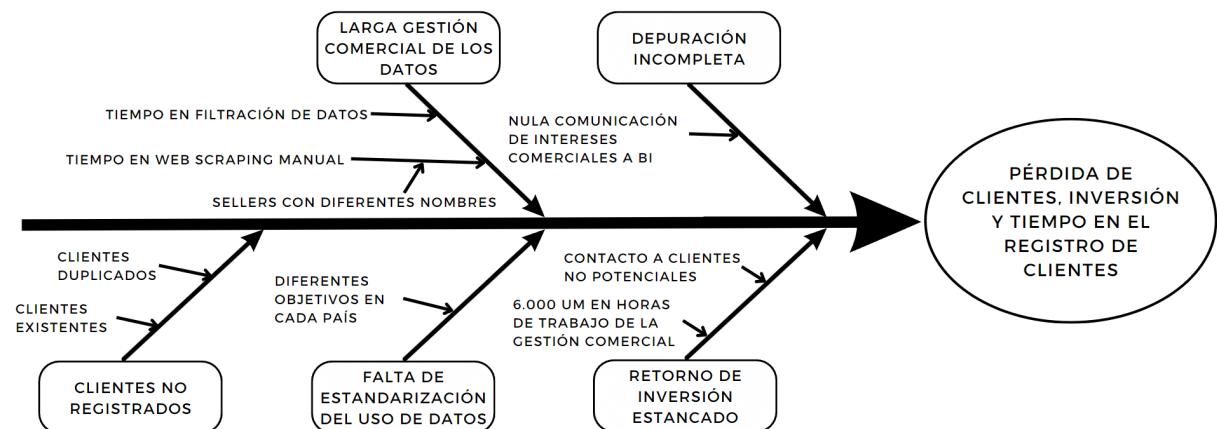


Figura 5: Diagrama de Ishikawa de la problemática encontrada

Como se puede ver en la figura 5, se tienen los focos de la problemática en el proceso de registro de clientes, destacando las brechas del problema en relación a la larga gestión comercial de los datos y clientes no registrados, que desencadenan un retorno de la inversión estancado. Todo lo anterior, apunta a la problemática de pérdida de clientes, inversión y tiempo en el proceso de registro de clientes en Inteligencia Competitiva de VIS en Mercadolibre, apareciendo la oportunidad de agilizar el proceso de registro de clientes para contar con una mayor calidad de los datos, mayor coordinación en el proceso y así un ahorro de recursos junto con una contactación efectiva de los usuarios potenciales.

## **Objetivos**

### **Objetivo general SMART**

Disminuir el tiempo medio de desarrollo del proceso de registro de clientes a dos semanas en un plazo de 3 meses.

### **Objetivos específicos**

- Agilizar el proceso de registro de clientes con foco en el subprocesso de gestión comercial.
- Reducir a 0 la cantidad de clientes duplicados y existentes en el registro de Salesforce.
- Identificar clientes que publican con diferentes nombres en la competencia.

## **Estado del arte**

Como se pudo observar hasta este punto en el desarrollo del proyecto, hay dos focos presentes para la exploración del estado del arte, la agilización del proceso de registro de clientes y la identificación de clientes que publican con diferentes nombres en la competencia.

Dentro de la agilización del proceso y la gestión de clientes potenciales, Salesforce, uno de los CRM más reconocidos de la industria, menciona que el crecimiento de una organización está directamente relacionado con la definición estratégica de un segmento de clientes y su proceso de captación (Salesforce, 2022). La segmentación de clientes es una clasificación de usuarios según diferentes características y se puede desarrollar mediante diversas herramientas. Actualmente en Mercadolibre, se han desarrollado diversos proyectos de segmentación, sin embargo, no han sido comunicados por el área comercial al resto de áreas involucradas en el proceso de registro.

Por otra parte, una de las empresas de ecommerce más grandes de Chile, Yapo.cl, realiza estudios de los mercados automotriz e inmobiliario, además de estudiar el comportamiento de sus usuarios de venta inmobiliaria. Por tanto, en la competencia podemos ver prácticas sobre gestión de clientes y estudio de mercado, lo cual puede ser efectivo a la hora de definir objetivos comerciales, y así agilizar la captación.

Por otro lado, el proceso de registro de clientes en Mercadolibre tiene un funcionamiento centrado en los subprocesos y división de tareas que se componen a través de la colaboración de áreas. En el mismo contexto, tenemos el caso de Frenetic, startup tecnológica dedicada al negocio B2B en Europa y Estados Unidos. Frenetic ha sido destacada dentro de las estrategias para tener una llegada efectiva a los clientes, por su definición de roles y tareas, en la identificación y gestión de clientes potenciales, transfiriendo las responsabilidades de los procesos de contacto, identificación de oportunidades y escalamiento en la relación con los clientes entre áreas, generando un desarrollo de negocios efectivo y eficiente (Escudero, 2021).

Lo desarrollado anteriormente, se puede complementar con el levantamiento de procesos y la gestión de procesos de negocios (BPM), ya que sus elementos son fundamentales a la hora de generar cambios que tengan un impacto significativo en la composición de un proceso. Para lo anterior, existe el aporte de técnicas y herramientas como la matriz RECI, que ayuda a definir roles, responsabilidades y actividades claves. También, hay otras herramientas que pueden consolidar el foco de la reingeniería de un proceso, como lo es el Modelo de Madurez (MMPE) de Hammer, donde se desarrollan facilitadores de procesos y capacidades de la empresa para asegurar un alto desempeño (Hammer, 2007).

Por otra parte, en relación a la brecha de identificación de los clientes de la competencia con diferentes nombres, en el campo de la inteligencia artificial se encuentra presente el estudio del NLP (Procesamiento del Lenguaje Natural), donde se desarrollan modelos y algoritmos para interpretar, manipular y procesar el lenguaje humano. En la actualidad existen diversas aplicaciones del NLP en empresas, donde los principales casos de usos son la eliminación de información confidencial, interacción con clientes y análisis de datos empresariales (AWS, 2023).

En el rubro de NLP, se tiene el caso de DNB (Banco Central de los Países Bajos), donde su equipo publicó un modelo llamado “Matching Company Names”, como solución a la problemática frecuente de recibir diferentes bases de datos de gran tamaño con datos no estructurados. En este desarrollo se implementó un preprocesamiento de texto estandarizando los nombres y se aplicó la similitud del coseno, una métrica de distancia que funciona realizando una vectorización de las palabras y calculando el ángulo entre dos vectores distintos a 0, estandarizando una similitud entre 0 y 1 (GraphEverywhere, 2019). Este algoritmo reduce la cantidad de coincidencias posibles trabajando con las 50 mejores coincidencias. Luego, se aplicaron métricas de coincidencia difusa, que hace referencia a encontrar patrones dentro de la combinación de cadenas de texto, definiendo una distancia de 0 a 1. Por último, se combinan los puntajes dejando abierta la opción de un postprocesamiento que se ajuste al caso de aplicación (Nijhuis, 2022).

También, bajo la misma problemática se tiene el algoritmo de “Fuzzy Name Matching with Machine Learning” de Chris Thornton, donde utiliza machine learning, métricas de distancia y embedding, los cuales apuntan a un modelo entrenado que procesa el lenguaje natural según diversos criterios, utilizando vectorización y corpus (colección de textos), pudiendo realizar una predicción de coincidencia de texto entre grandes cantidades de datos. En este caso para desarrollar los embeddings se utilizaron redes neuronales LSTM siamesas, un tipo de aprendizaje profundo enfocado en capas de nodos que trabajan en paralelo de forma recurrente, lo cual es útil para el manejo de datos secuenciales (GeeksforGeeks, 2023). Luego, utilizando F1-Score como métrica de puntuación, se entrenó al modelo en la biblioteca TensorFlow de Google (Thornton, 2021).

En base al modelo desarrollado por Chris Thornton, en la literatura de NLP el uso de redes neuronales LSTM siamesas ha sido contrastado con la aparición de los “Transformers”, los cuales surgieron a partir del concepto de mecanismos de atención, que permiten que el trabajo secuencial del análisis de texto mantenga la información relevante a lo largo del entrenamiento, siendo capaces de realizar un análisis del contexto de los datos. Este aporte fue tan importante que Google en 2017 publicó el documento “Attention Is All You Need”, el cual apunta a que los mecanismos de atención eran suficientes para resolver por sí solos las tareas de análisis secuencial de texto, definiendo el

concepto de “Transformers”, los cuales son redes neuronales basadas en mecanismos de atención que logran obtener mayores rendimientos que las redes LSTM siamesas (Dot CSV, 2021). Entre las ventajas de los transformers se encuentra el análisis simultáneo de texto, volviendo más rápida la resolución de tareas, además de resolver la pérdida de información propia de los análisis secuenciales. Dentro de los transformers, tenemos el modelo llamado BERT el cual ha sido destacado por su estructura (véase el anexo 1), la cual permite realizar tareas de codificación y decodificación de datos. En el caso de la generación de embeddings, se utiliza la codificación o vectorización en simultáneo de los datos, pudiendo ser aplicado en base a modelos pre-entrenados con anterioridad e incorporar insights a una cantidad de datos limitados a partir del entrenamiento de amplias colecciones de texto de forma más simple y rápida (CodeEmporium, 2020).

## **Propuestas de solución**

Dentro del foco de la agilización del proceso de registro de clientes, se propone realizar un estudio de mercado de los competidores de Inteligencia Competitiva, donde se estudie el contexto de Mercadolibre y se definan los objetivos comerciales de la captura de datos para luego realizar la filtración de datos que esté programada por el área de BI. Otra solución es desarrollar una especialización de roles, profundizando en el levantamiento del proceso de registro de clientes, desarrollando subprocesos especializados con las habilidades de las áreas involucradas, realizando un rediseño con un enfoque iterativo utilizando conceptos de BPM.

Con respecto a la identificación de clientes, una alternativa es desarrollar un algoritmo de identificación de nombres, incorporando un preprocesamiento y realizando una ponderación de métricas de coincidencia difusa en base a las mejores coincidencias según la similitud del coseno. También, para la identificación de nombres se propone aplicar un modelo de embedding basado en BERT, iterando sobre diferentes modelos y métricas para determinar la similitud de los embeddings o vectores generados.

En ambos casos anteriores, se incorporaría una eliminación de duplicados y existentes utilizando una base de clientes de Mercadolibre, y se evaluaría estadísticamente al modelo para unificar los diversos criterios descritos en un mismo algoritmo, probando parámetros de forma iterativa en sprints semanales de desarrollo. Estos modelos se desarrollarían en Fury (entorno colaborativo para programación de máquinas virtuales) utilizando control de versiones con Github (repositorio para almacenamiento de archivos y configuraciones) y el entorno de programación de python JupyterNotebook (ambas integradas en Fury) siendo automatizado para su uso futuro.

## Elección de la solución

Para escoger la solución se analizaron las diferentes alternativas junto con los managers de las áreas de Operaciones, Planning y BI & Analytics de VIS, donde en conjunto se definieron los siguientes criterios para evaluar a cada solución.

Criterio	Definición
Capacidad de datos	Disponibilidad y calidad de los datos para poder desarrollar la solución.
Capacidad de desarrollo	Evaluación económica en términos generales y disponibilidad de las áreas relacionadas a la solución.
Capacidad tecnológica	Disponibilidad de tecnologías para desarrollar la solución.
Tiempo de desarrollo	Viabilidad del desarrollo de la solución en los plazos del proyecto.
Alineación cultural	Si coincide con la cultura organizacional de MELI, es decir, si coincide con la propuesta de valor, uso de los datos, etc.
Alineación con los objetivos	Si se esperan resultados cercanos a los objetivos específicos.
Alineación con las metodologías	Si se puede desarrollar mediante las metodologías elegidas.

Figura 6: Criterios para evaluación de alternativas de solución

En base a los criterios mostrados en la figura 6 se realizaron se aplicó una escala likert para generar matrices de decisión, donde se ponderaron las soluciones con puntajes desde 1 a 5, donde se pueden representar dichos valores en una escala de cumplimiento de bajo, medio y alto (véase figura 7) y su promedio para cada solución. Se buscó despriorizar las alternativas que presentarán niveles bajos de cumplimiento, tanto en promedio como en algún criterio específico, por lo que para que el desarrollo se considere viable, debe contener sólo niveles de cumplimiento medio y alto.

Escala de puntajes	Cumplimiento	Descripción
[1-2[	Bajo	Posibles fallas, problemas de alcance o desajustes con el contexto del problema y de Mercadolibre.
[2-4[	Medio	Possible aparición de riesgos o algún grado de incertidumbre.
[4-5]	Alto	Cumple con el criterio demostrando viabilidad.

Figura 7: Escala de cumplimiento

Matriz de decisión de soluciones				
Criterio / Solución	Estudio de mercado + Modelo de similitud del coseno y métricas difusas	Estudio de mercado + Modelo de embedding con BERT	Especialización de roles + Modelo de similitud del coseno y métricas difusas	Especialización de roles + Modelo de embedding con BERT
Capacidad de datos	1,3	1,3	3,3	3,3
Capacidad de desarrollo	4	4	4	4
Capacidad tecnológica	5	5	5	5
Tiempo de desarrollo	1,7	1,3	3,3	3
Alineación cultural	2,7	2,7	3,3	3,3
Alineación con los objetivos	3,7	4,3	3,7	4,3
Alineación con las metodologías	4,7	5	4,7	5
<b>Promedio</b>	3,3	3,4	3,9	4
<b>Cumplimiento</b>	Bajo: [1-2]	Medio: [2-4]	Alto: [4-5]	

Figura 8: Matriz de decisión de soluciones

Como se puede observar en la figura 8, se muestran los puntajes promedios otorgados (mediante un formulario luego de una reunión de alineación sobre las soluciones y el desarrollo explicado hasta el momento) y validados por los managers de las áreas de Operaciones, Planning y BI de VIS, donde las soluciones que cumplen con los criterios de selección es la “Especialización de roles” aplicando una reingeniería de procesos para diseñar los subprocesos que permitan agilizar el proceso de registro de clientes, junto con los modelos de “Similitud del coseno y métricas de distancia” y “Modelo de embedding con transformers”, por lo que para elegir cual desarrollar se incluirán ambos en el primer sprint de desarrollo, iterando posteriormente sobre el que presente un mejor rendimiento.

## Riesgos y mitigaciones

Como muestra la figura 9, se desarrolló la siguiente matriz de probabilidad e impacto de riesgos, con una puntuación de 1 a 5 desde “Muy bajo” a “Muy alto”, multiplicando los valores coincidentes en cada caso. Se especifican los niveles de riesgo asignando una calificación.

Matriz de probabilidad e impacto de riesgos						Nivel de riesgo	
Probabilidad / Impacto	Muy bajo	Bajo	Medio	Alto	Muy alto	Puntaje	Calificación
Muy baja	1	2	3	4	5	1-4	Bajo
Baja	2	4	6	8	10	5-10	Medio
Media	3	6	9	12	15	11-16	Alto
Alta	4	8	12	16	20	17-25	Muy alto
Muy alta	5	10	15	20	25		

Figura 9: Matriz de probabilidad e impacto de riesgo junto con calificaciones de nivel de riesgo

Luego, se definió la siguiente matriz de riesgos con sus probabilidades, impactos, niveles de riesgo y mitigaciones, siendo controlados semanalmente (véase en la figura 10).

Evento	Probabilidad	Impacto	Nivel de riesgo	Mitigación
Desconocimiento del proceso rediseñado	Alta	Muy alto	Muy alto	Realización de kick-off donde se expliquen los cambios del proceso, junto con documentación y capacitaciones de ser necesario.
Incumplimiento del plazo de la implementación del proceso rediseñado	Media	Muy alto	Alto	Coordinación de reuniones de alineación programadas y horas de trabajo definidas según planificación.
Incumplimiento del plazo de la implementación del modelo de identificación	Media	Muy alto	Alto	Fijar cuatro sprints de desarrollo de una semana con horas de trabajo definidas con la empresa.
Desconocimiento del uso del modelo de identificación	Media	Alto	Alto	Generar documentación y realizar capacitación sobre desarrollo y casos de uso del modelo.
Resistencia al cambio de áreas involucradas	Baja	Medio	Medio	Explicación de principios lean y mostrar casos exitosos en encuentros personales en reuniones de coordinación.
Caída plataforma Fury (modelo de identificación)	Muy baja	Muy Alto	Medio	Contacto directo con soporte de Fury y respaldo local y en Github del modelo.
Pérdida del desarrollo del modelo	Baja	Muy Alto	Medio	Control de versiones del modelo en repositorio interno de BI de Github.
Parametrización errónea del modelo	Baja	Alto	Medio	Ánálisis estadístico e iteraciones ágiles en definición de parámetros de los modelos.
Baja representatividad de los datos en el desarrollo del modelo	Baja	Muy alto	Medio	Uso de datos reales de nombres de clientes para MVP y posterior uso de base de datos de todos los clientes de VIS.

Figura 10: Matriz de riesgos del proyecto

## Evaluación Económica

Para la evaluación económica del proyecto se tomaron en cuenta los costos de tecnologías, mantenimiento, implementación y horas de trabajo de los involucrados en el proyecto. También, se incluyeron los ahorros esperados en horas de trabajo con la solución implementada que en conjunto con los costos determinan el flujo del proyecto, para posteriormente determinar el Valor Actual Neto (VAN) y la Tasa Interna de Retorno (TIR), que permiten analizar la viabilidad económica y la tasa interna de rentabilidad del proyecto respectivamente. Además, se realizó un análisis de sensibilidad presentando tres escenarios, pesimista, neutro y optimista.

Siguiendo con lo anterior, el escenario pesimista, corresponde a la reducción a un mes en el tiempo de desarrollo del proceso de registro de clientes, el escenario neutro, una reducción a dos semanas y el optimista a una reducción a una semana. Para lo anterior, se modificaron los costos y ahorros según cada escenario. En base a todo lo mencionado se presenta el flujo del escenario neutro (véanse los escenarios pesimista y optimista en el anexo 2).

Flujo de ingresos y costos para evaluación económica del proyecto (unidades monetarias)					
Ingresos y costos / Periodos	Trimestre 0	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
<b>Ingresos</b>					
Ahorro de costos de empresa de horas de trabajo	-	\$4.810	\$4.810	\$4.810	\$4.810
<b>Costos de mantención</b>					
Costo de plan de entorno de programación Fury	-	-\$510	-\$510	-\$510	-\$510
Costo empresa de horas de trabajo para revisión de alertas de Fury	-	-\$113	-\$113	-\$113	-\$113
Costo empresa de revisión retrospectiva del proceso de registro de clientes	-	-\$100	-\$100	-\$100	-\$100
<b>Inversión fija</b>					
Costo de datos para MVP del modelo	-\$1.540				
Costo empresa por horas de trabajo del desarrollador del proyecto	-\$162				
Costo empresa por horas de trabajo por personal involucrado en el proyecto	-\$117				
Costo empresa por horas de trabajo para capacitación del proceso	-\$821				
Costo empresa por horas de trabajo para capacitación del modelo	-\$332				
<b>Flujo total</b>	<b>-\$2.972</b>	<b>\$4.087</b>	<b>\$4.087</b>	<b>\$4.087</b>	<b>\$4.087</b>

Figura 11: Flujo de ingresos y costos del escenario neutro

La figura 11 muestra el flujo de ingresos y costos de la situación neutra, dividiendo dicho análisis en trimestres debido a la periodicidad trimestral de los ciclos de Inteligencia Competitiva, siendo el “Trimestre 0” el periodo donde se está realizando el proyecto e incluyendo en dicho punto la

inversión fija y costos de implementación, luego a lo largo de los semestres podemos ver los ahorros en horas de trabajos traducidos al costo empresa, además de la mantención de las tecnologías y revisiones al proceso rediseñado.

Luego, para visualizar la sensibilidad de los escenarios se calcularon los VAN y TIR correspondientes a cada uno, con una tasa de descuento calculada sumando la tasa del 20% que utiliza Mercadolibre, a una tasa calculada con el modelo CAPM (cálculo de la tasa de retorno de un activo financiero en relación al riesgo) para castigar la tasa utilizada en Mercadolibre con datos de la actualidad económica de la empresa, con una tasa del 6,7% (Guidi, 2023), resultando en una tasa de descuento del 26,7%. Por tanto, en el escenario neutro, el VAN resulta en \$6.394 unidades monetarias y un TIR del 133%. Para el análisis de los indicadores económicos, se aplicó el mismo desarrollo para todos los escenarios.

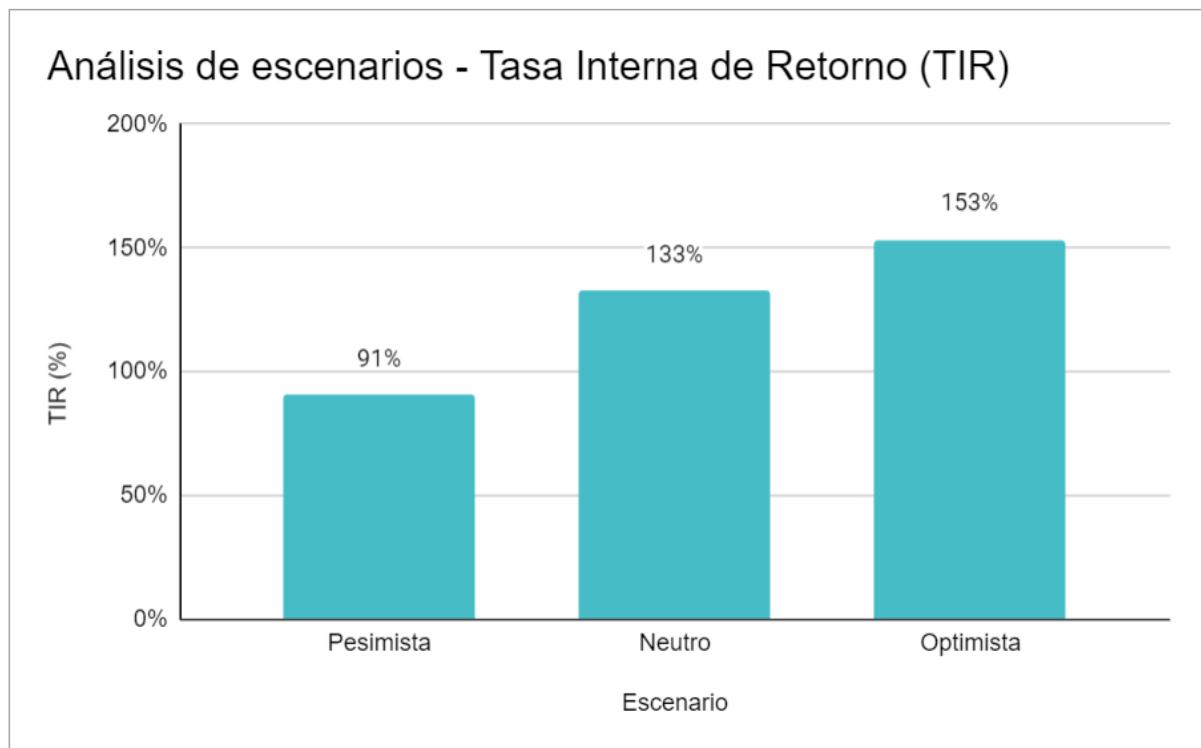


Figura 12: Análisis de escenarios pesimista, neutro y optimista para el Valor Actual Neto (TIR)

En la figura 12 se puede ver el desarrollo del TIR para los diferentes escenarios, donde el escenario pesimista presenta una tasa interna de retorno del 91%, elevándose al 133% y al 153% en los escenarios neutro y optimista respectivamente. En todos los casos, el TIR es mayor a 0 y a la tasa de descuento definida anteriormente del 26,7%, por lo que el proyecto tendría una tasa interna de retorno que garantiza la rentabilidad del proyecto en todos los casos desarrollados, siendo aceptado bajo esta métrica para destinar recursos por parte de la empresa.

## Análisis de escenarios - Valor Actual Neto (VAN)

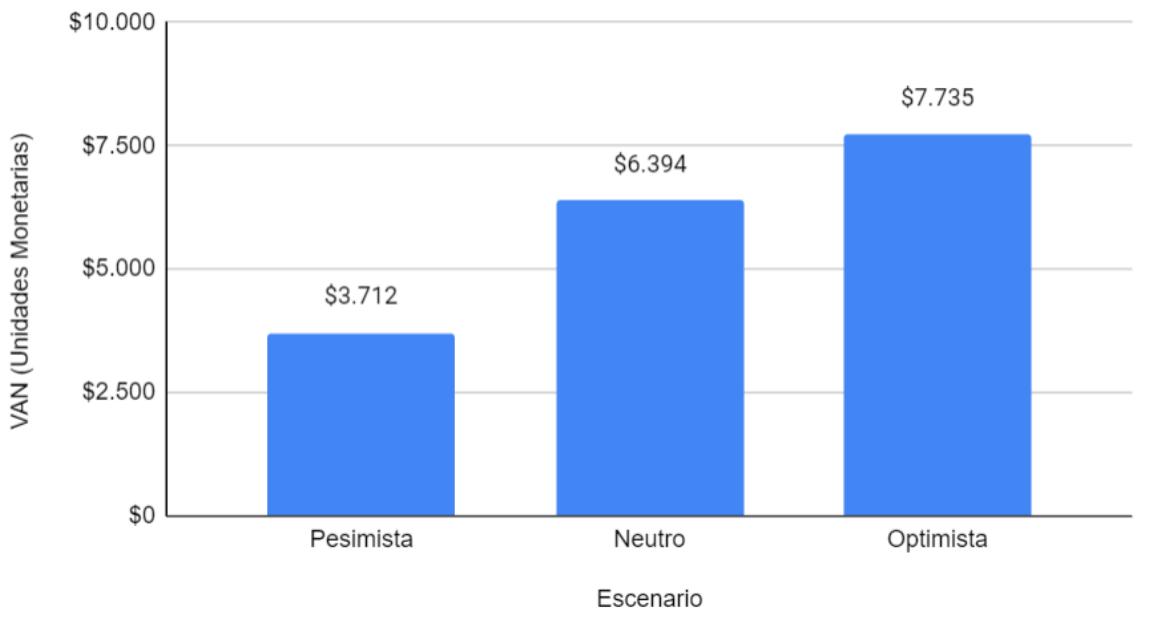


Figura 13: Análisis de escenarios pesimista, neutro y optimista para el Valor Actual Neto (VAN)

Por otra parte, complementando el análisis del TIR, en la figura 13 se puede ver la variación del VAN entre los diferentes escenarios, donde el escenario pesimista presenta un VAN de \$3.712 unidades monetarias, mientras que el escenario neutro muestra un VAN de \$6.394 unidades monetarias y el escenario optimista eleva el VAN a \$7.735 unidades monetarias. En todos los casos, el VAN es mayor a 0, es decir, la suma de la inversión inicial fija del proyecto y el valor actual (valor de los flujos a recibir en el futuro al momento de la evaluación) de los flujos definidos en el análisis son mayores a dicha inversión inicial cercana a las 3.000 U.M. necesarias para el desarrollo del proyecto, demostrando que su desarrollo es viable.

Todo lo anterior, los análisis del VAN y TIR aseguran la viabilidad económica en todos los escenarios definidos en base al objetivo del proyecto, contando con la validación de los managers de las áreas de Operaciones, Planning y BI & Analytics para confirmar el análisis mostrado, destinando los recursos descritos por parte de Mercadolibre a la realización del proyecto.

## **Metodologías**

### **Rediseño iterativo del proceso y especialización de roles**

En cuanto al rediseño del proceso, se utilizaron las metodologías Design Thinking y Lean Thinking, donde los principios de este último apuntan a un trabajo cílico en base al cliente, definiendo el valor a entregarle, el flujo que seguirá dicho valor, asegurarse del flujo de las actividades, que el cliente pueda testear lo desarrollado y optimizar constantemente el sistema propuesto. Por otro lado, Design Thinking pone foco en el contexto y forma, donde busca responder a las preguntas de “¿Cuál es el problema?”, “¿Por qué es importante?”, “¿Cómo lo resolvemos?”, “¿Cómo lo creamos?” y “¿Funciona?”, pudiendo desarrollar soluciones para refinar el proceso a medida que sea necesario (HubSpot, 2023).

Para evaluar el proceso en primera instancia, como se comentó en la propuesta de solución, se define la situación As-Is para establecer los principales elementos a tomar en cuenta para el rediseño. En base a lo anterior, se desarrolló la matriz RECI del proceso, con el objetivo de levantar las actividades y roles dentro del proceso. También se incorpora un Modelo de Madurez de Hammer para medir las capacidades actuales del proceso en busca de un alto desempeño, determinando a qué puntos darle foco. Además, se agrega un modelado en Bizagi para documentar el proceso rediseñado a bajo nivel, detallando las actividades según el modelo BPMN.

### **Desarrollo del modelo de identificación de clientes**

En cuanto al desarrollo del modelo para identificar clientes que publican con distintos nombres en la competencia, la metodología a seguir es CRISP-DM con cuatro sprints de desarrollo, donde a través de un entendimiento del negocio y de los datos, se realiza una preparación de estos para el posterior desarrollo del modelo y evaluación de su rendimiento. Luego, se sigue un desarrollo cílico volviendo al entendimiento del negocio, obteniendo feedback de los usuarios para finalmente llegar al despliegue e implementación del modelo automatizado.

### **Análisis estadístico del modelo**

Con el objetivo de ajustar el modelo de identificación para maximizar los resultados del modelo, se propuso la realización de un análisis estadístico basado en los conceptos asociados a la matriz de confusión, la cual consta de un conjunto de métricas que permiten cuantificar una tarea de clasificación entre los resultados de una predicción y los datos reales. En este caso se clasifica la coincidencia entre nombres de la base de datos destinada al desarrollo del modelo. Este análisis tiene relación con las pruebas de hipótesis donde la hipótesis sería que el cliente de la competencia

es el mismo usuario de MELI y bajo un umbral probabilístico se acepta o rechaza dicha hipótesis para cada caso pudiendo analizar cada métrica de exactitud de la predicción y clasificación en la matriz de confusión.

Bajo lo anterior, la matriz de confusión permite medir los diferentes tipos de errores que se generan en la tarea predictiva del modelo, donde se tienen el error de tipo I o falsos positivos, que hace referencia a que el modelo predice que el nombre de mayor coincidencia en Mercadolibre efectivamente coincide con el nombre del nuevo cliente, cuando en realidad no es el mismo cliente. Por otro lado, se tiene el error de tipo II o falsos negativos, que consta de que el modelo predice que el nombre de mayor coincidencia en Mercadolibre no coincide con el nombre del nuevo cliente, cuando en realidad si es el mismo cliente (DataSource, 2023). Lo anterior, se puede complementar con la siguiente figura.

Matriz de confusión		Predicción	
		Positivo	Negativo
Datos reales	Positivo	Verdadero Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadero Negativo (VN)

Figura 13: Matriz de confusión y métricas principales

Como se puede ver en la figura 13, se presenta la clasificación de valores entre los resultados de la predicción y los datos reales que se tienen en la construcción del modelo. Esta matriz nos permite medir el rendimiento del modelo con métricas que se derivan del análisis de clasificación comentado, profundizando en ellas en el apartado de medidas de desempeño.

## Planificación

La planificación del proyecto, orden de etapas, tareas y sus plazos fueron definidos mediante la siguiente Carta Gantt.



Figura 14: Carta Gantt del proyecto (véase el anexo 3)

Como muestra la figura 14, se realizó una división semanal del avance de las actividades en base a tres etapas, el entendimiento del negocio, desarrollo de la solución y despliegue de la solución.

## Plan de implementación

Para llevar a cabo el proyecto, en la figura 15 presentada a continuación, se puede apreciar el plan de implementación desarrollado, dividiendo dicho plan en etapas y actividades, incluyendo una descripción para cada actividad. Las etapas de desarrollo de la especialización de roles y rediseño del proceso de registro de clientes y desarrollo del modelo de identificación de clientes, se realizaron en paralelo según la planificación establecida.

Plan de implementación del proyecto		
Etapa	Actividades	Descripción
Desarrollo de especialización de roles y rediseño de proceso de registro de clientes.	Coordinación de reuniones de alineación	Coordinación de reuniones con analistas comerciales de Inteligencia Competitiva.
	Definición de la situación As-Is del proceso	Definición de matrices de descubrimiento, RECI y SIPOC, aplicación de cuestionario de Modelo de Madurez de Hammer y modelado BPMN.
	Reuniones de alineación	Reuniones semanales con analistas comerciales y encargado del proyecto del área de operaciones con enfoque al diseño y mejora del proceso de registro de clientes.
	Modelado iterativo del proceso	Modelamiento iterativo del proceso rediseñado con los conceptos destacados en las reuniones de alineación.
	Evaluación de cambios al proceso	Validación de cambios en el proceso con resultados del modelamiento.
	Relevamiento del proceso	Consolidación de cambios en documentos compartidos de Google Sheets donde se la hace seguimiento al proyecto.
	Kick-off del nuevo registro de clientes	Envío a las áreas involucradas del relevamiento del proceso y la forma de trabajar en base a este.
Desarrollo del modelo de identificación de clientes	Definición de bases de datos para desarrollo del modelo	Utilización de base de clientes del web scraping manual que cuenta con nombres de Mercadolibre, id y nombres de la competencia que ya fueron identificados.
	Desarrollo del procesamiento	Desarrollo del procesamiento con la base de datos del web scraping manual realizado por analistas comerciales, que cuenta con nombres de MEI y nombres de la competencia.
	Sprints semanales de desarrollo del modelo	Sprints de desarrollo donde se comparan las performances de diferentes modelos y parámetros, analizando estadísticamente los resultados.
	Aplicación del modelo	Utilización de base de datos de clientes de MEI y base de datos de la competencia para detectar clientes con distintos nombres de publicación.
	Automatización del modelo en Fury	Creación de job (trabajo cíclico que corre un código) en Fury diseñado para recibir una base de datos y entregar coincidencias de nombres, asignando un id de MEI.
Implementación final del proyecto	Generación de documentación del proyecto	Documentar el proceso de rediseño y el desarrollo, automatización y casos de uso del modelo de identificación de nombres de clientes.
	Capacitaciones	Coordinar capacitaciones para explicar lo generado en la documentación.
	Medición de resultados	Medir los resultados del modelo y del proceso rediseñado durante finales de noviembre.
	Control del modelo	Configurar alertas sobre funcionamiento en Fury y registro de última versión en GitHub integrado a este.
	Asignación de roles para el uso del modelo automatizado	Asignación de un responsable de BI capacitado para controlar el modelo (recibir alertas y asegurar su funcionamiento).
	Asignación de roles para responsables de BI del proceso de registro de clientes	Asignación de un responsable de BI capacitado para ejecutar los subprocesos de Inteligencia Competitiva.

Figura 15: Plan de implementación del proyecto

## Medidas de desempeño

En primer lugar, para evaluar el objetivo de agilización del proceso y su impacto se tienen las siguientes métricas asociadas al tiempo de flujo del registro de clientes y al valor monetario de los analistas comerciales que son el principal foco de los costos asociados al proceso.

- Tiempo de flujo del proceso de registro de clientes (días):
  - $T_1 - T_0$
- Valor monetario de horas de trabajo de analistas comerciales en Gestión Comercial (UM):
  - $\$ \text{Valor monetario de horas de trabajo de analistas en el ciclo IV}$

Luego, para medir el objetivo de la reducción de clientes fuera del ciclo por errores de duplicación y cuentas ya existentes dentro del registro en el CRM Salesforce, se tienen las métricas de la diferencia porcentual entre el Ciclo III correspondiente al tercer trimestre del año y el Ciclo IV correspondiente al último trimestre del año y la cantidad de errores que se identifiquen en el registro porcentualmente.

- Diferencia porcentual de clientes contactados entre Ciclo III y IV (%):
  - $\frac{(C_{IV} - C_{III})}{C_{III}} \times 100$
- Cantidad de errores en el registro de clientes en Salesforce (%):
  - $\frac{E_{IV}}{T_{IV}} \times 100$

En relación al objetivo de identificar a los usuarios de la competencia que publican con diferentes nombres en la competencia, se tienen diversas métricas para evaluar el rendimiento del modelo de identificación, donde aparecen métricas estadísticas de exactitud, precisión, sensibilidad, F1-score, junto con añadir el tiempo de flujo del modelo.

- Rendimiento del desarrollo del modelo de identificación:
  - “Exactitud” (%): porcentaje de predicciones correctas del modelo.
    - $\frac{VP + VN}{VP + FP + VN + FN}$  donde,  
VP: Verdadero positivo, VN: Verdadero negativo, FP: Falso positivo y FN: Falso positivo.

- “F1-Score” (%): medida de efectividad del modelo que realiza una media armónica entre la precisión y la sensibilidad donde,

- “Precisión” (%): porcentaje de predicciones positivas correctas del modelo.

$$\bullet \quad \frac{VP}{VP + FP}$$

- “Sensibilidad” (%): porcentaje de casos positivos detectados en el modelo.

$$\bullet \quad \frac{VP}{VP + FN}$$

- “F1-Score” (%):  $\frac{2 \times (\text{precisión} \times \text{sensibilidad})}{\text{precisión} + \text{sensibilidad}}$

- Tiempo de flujo del modelo de identificación (horas):

- $Tiempo \ de \ finalización \ del \ modelo \ (X_1) - Tiempo \ inicio \ del \ modelo \ (X_0)$

## Desarrollo e implementación del proyecto

### Definición de la situación As-Is

En primera instancia, luego de la coordinación de las reuniones y definir horas de trabajo, se realizó el análisis y desarrollo de la situación As-Is, donde se define en primer lugar la matriz RECI de la situación actual del proceso, donde según las actividades de cada subprocesso y sus participantes se revisan los roles asignados actualmente.

Matriz RECI de la situación actual del registro de clientes					
Actividades / Roles	Encargado de BI & Analytics	Proveedor externo	Encargado de operaciones	Analistas Comerciales	Jefe de la vertical y país
<b>Depuración de la información</b>					
Revisión de inconsistencias de formato y duplicados de base de datos externa.	R/E	I	I		
Disponibilizar la base de datos externa corregida.	R/E		I		I
<b>Armado de la base de datos</b>					
Subir base de datos a Cloud Storage (servicio de almacenamiento de datos de Google).	R/E				
Crear una tabla en Bigquery de la base externa.	R/E				
Cruzar datos de la tabla creada con registros de MELI para crear una base de datos de clientes potenciales.	R/E				
Disponibilizar la base de datos de clientes potenciales.	R/E		I		I
<b>Gestión comercial de los datos</b>					
Web scraping manual para identificar clientes con distinto nombre en la competencia.				R	E
Categorización y filtrado de los datos en base de clientes potenciales.			I	R	E
<b>Registro de clientes en Salesforce (CRM)</b>					
Registrar base filtrada en Salesforce.	C		R/E	I	
<b>R: Responsable</b>	Rol encargado de realizar la actividad.				
<b>E: Encargado</b>	Rol que aprueba el trabajo realizado por el responsable.				
<b>C: Consultado</b>	Aquellos expertos que son consultados sobre algún aspecto de la tarea.				
<b>I: Informado</b>	Aquellas personas que deben ser informadas sobre la evolución de la tarea.				

Figura 16: Matriz RECI de la situación actual del registro de clientes

Como se puede ver en la figura 16, la matriz RECI de la situación actual muestra sólo aprobadores en el subprocesso de gestión comercial y sólo un rol de consultado, es decir, la petición de un punto de vista experto, en el último subprocesso del proceso. Además, los analistas comerciales no participan en el proceso hasta la gestión comercial. En relación a lo mencionado, podemos ver como en relación al uso de datos de la gestión comercial no se le da participación al área especializada en datos de BI & Analytics, donde esta última tampoco genera una conexión con el negocio, solo informando al área de operaciones y altos mandos cuando se finalizan los subprocessos en los cuales es responsable.

También, se sumó al análisis y definición de la situación actual del proceso, la aplicación del modelo de madurez de Hammer al proceso de registro de clientes (véase el anexo 4), obteniendo los siguientes resultados.

MMDP aplicado al proceso de registro de clientes						
Facilitador		P-1	P-2	P-3	P-4	Nivel
Diseño	Propósito					P-1
	Contexto					
	Documentación					
Ejecutores	Conocimientos					P-1
	Destrezas					
	Conductas					
Responsables	Identidad					P-3
	Actividades					
	Autoridad					
Infraestructura	Sistemas de información					P-3
	Sistemas de recursos humanos					
Indicadores	Definición					P-1
	Usos					

Cumplimiento de cada nivel
Frecuencia de <20%
Frecuencia de 20%-80%
Frecuencia de >80%

Figura 17: MMDP de la situación actual del registro de clientes

En la figura 17 podemos ver la evaluación de los facilitadores del proceso según su cumplimiento en cada nivel, donde se considera que el facilitador alcanza un nivel cuando todos sus puntos alcanzan como mínimo una frecuencia mayor del 80% en dicho nivel, por lo que se considera que el diseño, los ejecutores y los indicadores estarían en un nivel de P-1, siendo los principales focos de madurez a tomar en cuenta en el rediseño. El nivel P-1 de los facilitadores mencionados hacen que el proceso pertenezca al mismo nivel, definiéndose según la teoría de Hammer como un proceso confiable y predecible (Hammer, 2007).

Finalizando la definición de la situación actual As-Is, se tiene el modelamiento del proceso desarrollado en Bizagi siguiendo los conceptos de BPMN (véase el anexo 5), que siguen la descripción del proceso mencionado en el transcurso del desarrollo del informe, donde podemos ver sus subprocesos y actividades para evaluar y realizar cambios conociendo a profundidad todos los pasos por los que pasa el proceso de registro de clientes.

## Primera iteración de rediseño del proceso

El análisis realizado anteriormente en la definición de la situación As-Is fue el punto de partida para la primera iteración del rediseño del proceso, que en conjunto con las brechas comentadas en el contexto del proceso se desarrolló la siguiente matriz de mejora, utilizada para levantar los puntos relevantes a evaluar en el rediseño.

Matriz de mejora del proceso de registro de clientes: 1ra iteración	
¿Cuál es el problema?	Pérdida de tiempo en la manipulación de datos en la Gestión Comercial. Desconocimiento y falta de métricas del registro de clientes.
¿Por qué es importante?	Inversión y costo de horas de trabajo.
¿Cómo lo resolvemos?	Mayor coordinación del área comercial y BI. Reasignación de roles en los subprocesos del registro de clientes. Medición de los subprocesos con métricas del proyecto.
¿Cómo lo creamos?	Archivo compartido para ingreso de intereses comerciales y modelamiento de nuevo proceso de armado de base de datos integrando dichos intereses.
¿Funciona?	Prueba del nuevo formato con analistas.
¿Genera valor para el cliente?	Cliente interno: ahorro de horas de trabajo y comunicación efectiva. Cliente MELI: mayor agilidad del proceso (contactación más rápida).
¿Cuál es el flujo de valor?	

Figura 18: Matriz de mejora del proceso de registro de clientes en su primera iteración

Como muestra la figura 18, se profundizaron en conceptos conocidos dentro del análisis del problema y la solución propuesta, donde el foco de la primera iteración se concentró en la Gestión Comercial y la brecha que existe en los indicadores del proceso. También, se proponen soluciones ágiles para que no tengan un costo asociado y se definen roles, tareas y pruebas, aplicándolas al proyecto en un país específico sin afectar el desarrollo completo del ciclo de Inteligencia Competitiva.

En base a lo anterior, se implementó un nuevo proceso de definición de intereses comerciales paralelo a la depuración de la información, donde se comparte la información comercial para el armado de bases de datos, sin afectar a su tiempo de desarrollo por la facilidad de filtración en Bigquery. Lo anterior, se puede apreciar en la figura 19.

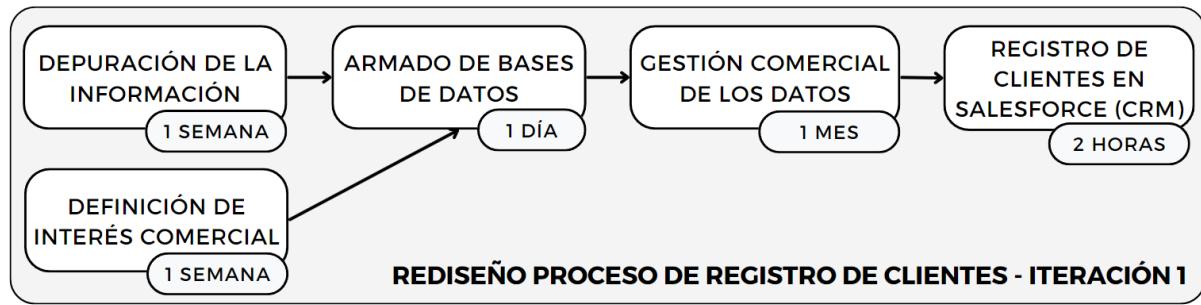


Figura 19: Rediseño del proceso de registro de clientes en la primera iteración

A partir del desarrollo anterior, incluyendo el nuevo subproceso de definición de interés comercial, dejando de lado la categorización de datos que pertenecía a la gestión comercial de los datos, se tienen los siguientes resultados en términos de la medición del proceso.

Medición de resultados del rediseño del proceso		
Medida de desempeño	Última medición	Estimado iteración 1
Tiempo de flujo del proceso de registro de clientes (días)	53 días aprox.	38 días aprox.
Horas de trabajo de analistas comerciales en Gestión Comercial	460 horas aprox. (6.000 UM)	400 horas aprox. (5.200 UM)

Figura 20: Resultados de la primera iteración

La figura 20 muestra la medición de los resultados de la primera iteración, donde se tiene una reducción de 15 días en el tiempo de flujo del proceso de registro de clientes y una reducción de 60 horas de trabajo de los analistas comerciales, significando un ahorro de 800 unidades monetarias.

## Segunda iteración de rediseño del proceso

El desarrollo correspondiente a la segunda iteración del rediseño, apuntó principalmente a la verificación de la correcta aplicación de los datos entregados del subprocesso de “Definición de intereses comerciales” para el armado de las bases de datos, desarrollando lo siguiente en la matriz de mejora del proceso presentada en la figura 21.

Matriz de mejora del proceso de registro de clientes: 2da iteración	
¿Cuál es el problema?	Posibles errores de categorización de datos en el nuevo armado de bases de datos.
¿Por qué es importante?	Inversión y costo de horas de trabajo.
¿Cómo lo resolvemos?	Integración del rol especializado del área comercial en los datos resultantes de las bases de datos y feedback al área de BI & Analytics.
¿Cómo lo creamos?	Trabajar cíclicamente el armado de la base de datos con la gestión comercial de los datos, con tablas a diferentes niveles de detalle de información de los clientes (todos los clientes a los clientes que son objetivos comerciales).
¿Funciona?	Realización de un trabajo de una base de datos en conjunto con una verificación cíclica de los datos.
¿Genera valor para el cliente?	Cliente interno: ahorro de horas de trabajo y comunicación efectiva. Cliente MELI: mayor agilidad del proceso (contactación más rápida).
¿Cuál es el flujo de valor?	

Figura 21: Matriz de mejora del proceso de registro de clientes en su segunda iteración

En esta iteración, su desarrollo resumido en la figura 21 se concentró en la efectividad y especialización que aporta el área comercial en los posibles errores del nuevo armado de bases de datos donde se definen los intereses comerciales, formando un trabajo colaborativo para ir definiendo bases específicas de los clientes que son parte de los objetivos del proceso. Las pruebas de este cambio se realizó con una base específica de un competidor realizando la prueba del desarrollo del armado de la base de datos y las verificaciones que realizará el área comercial de la completitud de los datos.

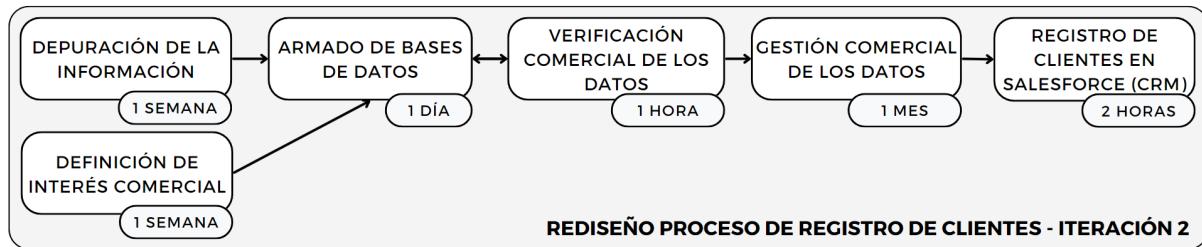


Figura 22: Rediseño del proceso de registro de clientes en la segunda iteración

Continuando, como se muestra en la figura 22, el modelado del proceso incluye un nuevo subprocesso de "Verificación comercial de los datos", apuntando en un trabajo colaborativo donde cada área se especializa desde sus capacidades en mantener el flujo de valor del rediseño del proceso. En cuanto al nuevo subprocesso, al ser un filtrado en Bigquery y tener definidos con anterioridad los intereses comerciales, no se afecta al proceso de forma considerable en términos de tiempo del proceso.

### Tercera iteración de rediseño del proceso

Esta iteración se centra en el desarrollo del modelo de identificación de clientes para reducir el tiempo de web scraping manual realizado en el subprocesso de Gestión Comercial, automatizando dicha tarea manual.

- **Primer sprint de desarrollo del modelo**

En la primera parte del desarrollo, en el entendimiento del negocio se definió la base de datos a utilizar en los sprints en base a la realización de un web scraping manual donde se identificaron nombres de la competencia asociados a nombres de Mercadolibre, contando con más de 1.500 clientes. Luego se desarrolló el preprocesamiento aplicando diferentes funciones de python en los nombres para la eliminación de duplicados y caracteres especiales, uso de minúsculas, etc. En la siguiente figura se puede ver un ejemplo del preprocesamiento de un par de nombres.



Figura 23: Ejemplo del input y output del procesamiento del primer sprint

Siguiendo el ejemplo mostrado en la figura 23, aplicando el preprocesamiento a todos los nombres y realizando una coincidencia entre nombres preprocesados se obtuvo una identificación efectiva de un 9% de los clientes.

También se realizó el desarrollo del algoritmo de similitud del coseno y métricas difusas con tiempo de procesamiento de 65 minutos aproximadamente y alcanzó un 41,6% de coincidencias efectivas. Por otro lado, también se desarrolló el modelo de embeddings con BERT, basado en la aplicación del modelo “all-MiniLM-L6-v2” de BERT presente en la librería “Sentence-Transformers” con el objetivo de cumplir múltiples tareas, ya que está entrenado con más de mil millones de datos de entrenamiento (Sentence-Transformers, 2023). Las similitudes de los embeddings generados por BERT se midieron con la métrica de distancia del coseno, obteniendo un 62% de coincidencias efectivas de nombres con un tiempo aproximado de 2 minutos. Según los resultados descritos, se definió al modelo de embedding basado en BERT como la solución a desarrollar para el foco del modelo identificación de clientes.

- **Segundo sprint de desarrollo del modelo**

Luego del primer sprint, se analizaron los resultados del preprocesamiento, donde se encontró desde la mirada del negocio que muchos nombres no llegaban a tener una coincidencia exacta por palabras que se repetían a lo largo de los datos. Por lo anterior, se analizaron la frecuencia de palabras comunes, donde “servicios”, “propiedades”, “negocios”, “inmobiliarios” e “inmobiliaria” aparecen más de 950 veces entre los 3000 nombres. Por tanto, se incluyó la eliminación de palabras comunes (véase el código en el anexo 6) obteniendo una identificación efectiva de un 20% de los clientes dentro de la base destinada al desarrollo, aumentando considerablemente el resultado del primer sprint en el preprocesamiento.

Según lo comentado, el objetivo del modelo es determinar un valor probabilístico de coincidencia, definiendo una similitud entre los vectores de nombres, la cual posteriormente se evalúe bajo un umbral que determinará si la máxima coincidencia de Mercadolibre corresponde al nuevo cliente. Para lo anterior, con la ayuda de la biblioteca Scikit-Learn, se aplicaron las métricas de similitud más utilizadas en desarrollos de NLP, la similitud del coseno, producto punto y la distancia euclíadiana (Briggs, 2022).

Se utilizaron los mismos conceptos asociados a la matriz de confusión explicada anteriormente, donde aparece las curvas de “Precision-Recall” (Precisión-Sensibilidad), graficando dichos valores para diferentes umbrales, donde la intersección de las curvas determina el equilibrio entre la precisión y la sensibilidad, definiendo un umbral óptimo, maximizando el F1-Score (véase el desarrollo en python en el anexo 7). A modo de ejemplo se muestra la curva de “Precision-Recall” en la siguiente figura.

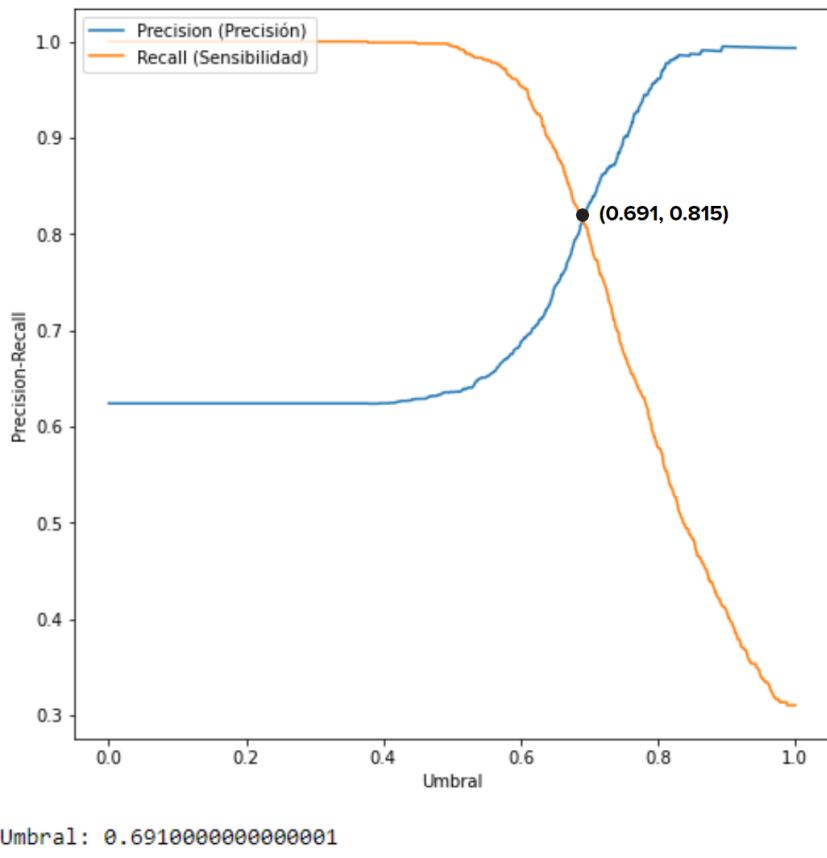


Figura 25: Gráfico “Precision-Recall” para determinar el umbral de la similitud del coseno junto el umbral

En la figura 25, podemos ver el ejemplo de la aplicación de la métrica de similitud del coseno y el análisis del umbral de probabilidad mediante las curvas Precision-Recall, donde para un umbral de 0,691, es decir, un 69,1% de similitud del coseno entre las máximas similitudes, donde las que sean mayor a dicho umbral se consideren como el mismo cliente. Con lo anterior, se obtiene el equilibrio entre los dos tipos de errores y se muestra el F1-Score que sería aproximadamente de un 81,5%. Replicando el desarrollo mencionado y determinando los umbrales de la misma forma que muestra la figura 25, se obtuvieron los siguientes resultados del rendimiento de cada métrica de similitud.

Medidas de desempeño / Métricas de similitud	Similitud del coseno	Producto punto	Distancia euclíadiana
Umbral	69%	98.6%	45.5%
Exactitud	76.95%	76.62%	76.82%
F1-Score	81.53%	81.36%	81.45%

Figura 26: Resultados de las medidas de desempeño para cada métrica de similitud

La figura 26 muestra cada umbral junto con los resultados de las medidas de desempeño según cada métrica de similitud. La métrica que destaca levemente sobre las demás fue la similitud del coseno, alcanzando una precisión del 76,95% y F1-Score del 81,53% mediante un umbral del 69%, siendo esta

la métrica a utilizar en el resto del desarrollo. Además, el tiempo de flujo fue aproximadamente de 3 minutos, siendo levemente superior al tiempo de flujo del primer sprint.

- **Tercer sprint de desarrollo del modelo**

Llevando los resultados del desarrollo del modelo en el segundo sprint al entendimiento del negocio, se destaca la importancia de iterar sobre modelos de BERT, para probar diferentes modelos que se ajusten al caso de aplicación. Utilizando los modelos de BERT documentados en SBERT.net y HuggingFace, plataformas donde se publican diversos modelos pre-entrenados, se realizó una búsqueda de modelos en tareas de similitud textual de textos y colecciones de textos en español. En base a lo anterior, se comparó el rendimiento de 21 modelos, obteniendo los siguientes resultados.

Modelos BERT / Medidas de desempeño	Exactitud	F1-Score	Tiempo de flujo (min)
multi-qa-MiniLM-L6-cos-v1	78%	82.9%	2
paraphrase-MiniLM-L3-v2	77.1%	82.4%	2
multi-qa-distilbert-cos-v1	77%	81.8%	3
all-MiniLM-L6-v2	77%	81.5%	3
multi-qa-mpnet-base-dot-v1	75.6%	81%	4
paraphrase-albert-small-v2	76.6%	80.9%	3
all-MiniLM-L12-v2	76.4%	80.4%	3
mrm8488/RuPERTa-base	77.9%	79%	3
hiiamsid/sentence_similarity_spanish_es	76.9%	78.8%	3
all-distilroberta-v1	75%	78.3%	3
<b>mrm8488/TinyBERT-spanish-uncased-finetuned-ner</b>	<b>87%</b>	77.8%	4
nli-distilroberta-base-v2	72.4%	77.4%	3
paraphrase-multilingual-mpnet-base-v2	76.3%	76%	5
bert-base-uncased	81.5%	75.9%	3
all-mpnet-base-v2	72.6%	75.3%	4
mrm8488/bert-spanish-cased-finetuned-ner	83.4%	75.1%	6
bert-base-nli-mean-tokens	76.4%	75%	3
distiluse-base-multilingual-cased-v1	74.9%	74.3%	5
hackathon-pln-es/paraphrase-spanish-distilroberta	73.5%	73.7%	3
paraphrase-multilingual-MiniLM-L12-v2	76%	72.4%	3
distiluse-base-multilingual-cased-v2	75%	70.6%	5

Figura 27: Resultados de las medidas de desempeño para los diferentes modelos de BERT aplicados

Como se aprecia en la figura 27, se presentan los resultados en las diferentes medidas de desempeño. El modelo “multi-qa-MiniML-L6-cos-v1” presenta los valores máximos en términos de F1-Score con un 82,9% y el menor tiempo de flujo con un aproximado de 2 minutos, mientras que el modelo “TinyBERT-spanish-uncased-finetuned-ner” presenta el valor máximo de exactitud con un 87%. La elección del modelo a utilizar se define por la comparación entre la exactitud y el F1-Score. Si

bien la exactitud representa las predicciones acertadas, en dichas predicciones pueden haber falsos negativos o falsos positivos, que correspondan a una pérdida de información fundamental a la hora de identificar clientes correctamente, y que se puede equilibrar con la métrica de F1-Score, ponderando la precisión y sensibilidad del modelo. Por esto el modelo a utilizar en el resto del desarrollo corresponde es el modelo “multi-qa-MiniML-L6-cos-v1”, el cual fue entrenado con millones de datos en forma de pregunta-respuesta, siendo útil para diversas tareas y contando con un procesamiento rápido.

Comparando los resultados del sprint anterior, el modelo actual presentó un aumento del 1,4% en la medida de F1-Score y 1% en exactitud, con un tiempo de flujo menor con un total de 2 minutos.

- **Cuarto sprint de desarrollo del modelo**

En el último sprint de desarrollo, se realizó un entendimiento del negocio del avance realizado hasta este punto, donde como último foco está la posible incorporación de métricas de similitud difusa, para complementar el desarrollo de BERT con tratamiento directo del texto. Dentro de las métricas de distancia se tiene a la librería “Fuzzy-Wuzzy”, la cual contiene diversas métricas, donde se tienen ratios de similitud entre dos textos con diversas variantes. Por lo anterior, se ponderaron las diferentes métricas con la similitud del coseno, obteniendo los siguientes resultados.

Métricas Fuzzy Wuzzy / Medidas de desempeño	Exactitud	F1-Score	Tiempo de flujo (min)
<b>Partial Ratio</b>	<b>82.7%</b>	<b>88.2%</b>	<b>4</b>
Partial Ratio + Simple Ratio	79.5%	86.2%	6
Partial Ratio + Token Sort Ratio	79.2%	85.9%	6
Partial Ratio + Simple Ratio + Token Sort Ratio	79%	85.7%	9
Partial Ratio + Simple Ratio + Token Set Ratio	78.9%	85.6%	9
Partial Ratio + Token Set Ratio	78.9%	85.1%	6
Partial Ratio + Token Sort Ratio + Token Set Ratio	78.3%	84.9%	9
Todas las métricas	77.8%	84.7%	14
Token Set Ratio	79.4%	84.5%	<b>4</b>
Simple Ratio	76.7%	83.4%	<b>4</b>
Token Sort Ratio	76.2%	82.4%	<b>4</b>

Figura 28: Resultados para cada métrica de Fuzzy-Wuzzy ponderada con la similitud del coseno

La figura 28 muestra los resultados de la aplicación de diferentes métricas de Fuzzy-Wuzzy, donde se tenía “Simple Ratio”, el cual se basa en la distancia de Levenshtein (cantidad de ediciones para conseguir una similitud exacta), “Partial Ratio” que otorga mayor similaridad a nombres que estén contenidos en otros, “Token Sort Ratio” que ignora el desorden de las palabras que componen al nombre y “Token Set Ratio” que ignora palabras repetidas en el cálculo de la similaridad.

La métrica que mayor rendimiento muestra es “Partial Ratio” con una exactitud del 82,7% y un F1-Score del 88,2%, alcanzando un rendimiento superior al del sprint anterior. Vale destacar, que el umbral de esta aplicación se determinó en un 72,2%, siendo este el seleccionado para evaluar las máximas similitudes generadas por el modelo en su posterior implementación.

Según lo anterior, el modelo de identificación de clientes se definió por los desarrollos aplicados a lo largo de los sprints. Luego, se tiene que el output definido para el modelo de una base que centralice toda la información de los clientes y una etiqueta de la predicción de que el cliente de la competencia tiene un nombre coincidente en Mercadolibre para su posterior análisis (véase el notebook de todo el modelo desarrollado en el anexo 8).

#### Cuarta iteración de rediseño del proceso

En esta última iteración del rediseño el foco se instaló en la incorporación del modelo de identificación de clientes desarrollado al proceso de registro de clientes de Inteligencia Competitiva.

Matriz de mejora del proceso de registro de clientes: 4ta iteración	
¿Cuál es el problema?	Incorporar el modelo de identificación de clientes en el proceso en vez del web scraping manual realizado en la Gestión Comercial.
¿Por qué es importante?	Inversión y costo de horas de trabajo en el web scraping manual y en el proyecto.
¿Cómo lo resolvemos?	Sumar al nuevo subprocesso de "Verificación comercial de los datos", la verificación de los clientes detectados, sumando anteriormente un subprocesso de BI & Analytics para la aplicación del modelo.
¿Cómo lo creamos?	Implementando el modelo en Fury, definiendo responsabilidades internas de cada área y definiendo horas de trabajo para la verificación de los datos.
¿Funciona?	Realización de una prueba del uso del modelo y la base de datos que recibirá la Gestión Comercial.
¿Genera valor para el cliente?	Cliente interno: ahorro de horas de trabajo y definición de responsabilidades. Cliente MELI: mayor agilidad del proceso (contactación más rápida).
¿Cuál es el flujo de valor?	 <p>ASEGURAR LA CALIDAD DE LOS DATOS</p> <p>MANTENER UNA COORDINACIÓN INTERNA EFECTIVA</p> <p>REDUCIR TIEMPOS INNECESARIOS</p> <p>MANTENER UNA MEDICIÓN DEL PROCESO</p>

Figura 29: Diagrama del modelo de identificación de nombres

Como se muestra en la figura 29, se realizó la matriz de mejora para la última iteración con el foco comentado anteriormente, donde se fijó la creación de un subprocesso de “Aplicación del modelo de identificación” para posteriormente entregar los resultados al área comercial.

Para definir el tiempo estimado para la aplicación del modelo se realizó una implementación del modelo con una base de clientes no identificados de la competencia según su país y vertical (vehículos o inmuebles), sumando los tiempos de subida de archivos e inicio del laboratorio en Fury que aloja el modelo (véase el anexo 9), resulta un tiempo de flujo del modelo de 1,5 horas.

En la aplicación anterior del modelo de un total de 6.700 clientes no identificados aproximadamente se detectaron cerca de un 66% con alerta de coincidencia (se identificaron un 67% de clientes en el armado de la base de datos por información de contacto, por lo que dicha base de no identificados corresponde al 33% y se identificaron un 22% del total de clientes). Por tanto, según la capacidad que se mencionó en un principio de los analistas comerciales en la “Gestión comercial de los datos” y la implementación del modelo, la revisión de los clientes no identificados, donde se revisan variables descriptivas y son tratados como clientes existentes dentro del registro, reportando esto a los ejecutivos comerciales a la hora de hacer el contacto aplicando acciones preventivas de falsos positivos.

Lo anterior, equivaldría a 27 horas de trabajo aproximadamente, dejando una holgura de hasta 4 días en coordinación con los equipos comerciales. Por lo anterior, se excluye el subprocesso de “Gestión comercial de los datos” debido a la eliminación de los web scrapings manuales por parte del área comercial.

Según lo comentado, podemos ver el modelamiento final del nuevo proceso de registro de clientes.

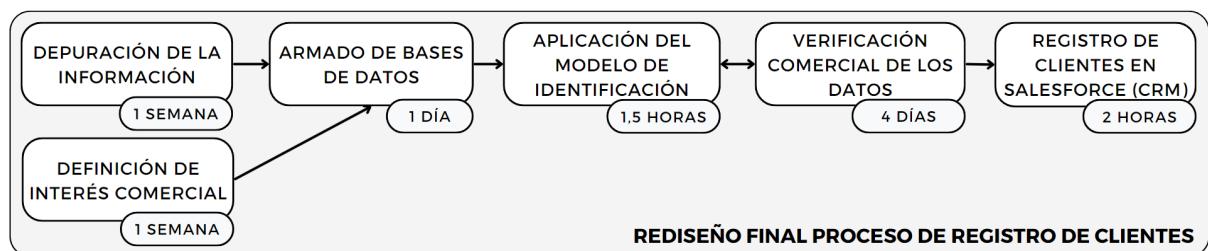


Figura 30: Modelamiento del rediseño del proceso de registro de clientes

En la figura 30 podemos revisar los subprocessos finales del nuevo proceso de registro de clientes en Inteligencia Competitiva, con los tiempos mencionados anteriormente según el rediseño. Para mayor detalle de los subprocessos y sus actividades se puede ver el modelamiento BPMN del proceso rediseñado en el anexo 10.

Para realizar el relevamiento del nuevo proceso de registro de clientes, en el archivo compartido definido en la primera iteración se documentaron los principales cambios, agregando el modelamiento del proceso rediseñado y su nueva matriz RECI presentada a continuación.

Matriz RECI del proceso de registro de clientes rediseñado					
Actividades / Roles	Encargado de BI & Analytics	Proveedor externo	Encargado de operaciones	Analistas Comerciales	Jefe de la vertical y país
<b>Depuración de la información</b>					
Revisión de inconsistencias de formato y duplicados de base de datos externa.	R/E	I	I		
Disponibilizar la base de datos externa corregida.	R/E		I		I
<b>Definición de interés comercial</b>					
Disponibilizar filtros a aplicar a la base externa.	I		I	R/E	E
<b>Armado de la base de datos</b>					
Subir base de datos a Cloud Storage (servicio de almacenamiento de datos de Google).	R/E				
Crear una tabla en Bigquery de la base externa.	R/E				
Filtrar base de datos de Bigquery según interés comercial.	R/E			C	
Cruzar datos de la tabla creada con registros de MELI para crear una base de datos de clientes potenciales.	R/E			C	
<b>Aplicación del modelo de identificación</b>					
Subir base de datos de la competencia a Github.	R/E				
Activar laboratorio en Fury y verificar correcto funcionamiento del modelo.	R/E				
Disponibilizar la base de datos de clientes potenciales con etiquetado de posible coincidencia de nombre en MELI.	R/E		I	I	
<b>Verificación comercial de los datos</b>					
Verificar la correcta filtración de la base de datos de clientes potenciales.	I			R	E
Verificar la información de clientes detectados por el modelo de identificación.	C		I	R	E
<b>Registro de clientes en Salesforce (CRM)</b>					
Registrar base filtrada en Salesforce.	C		R/E	I	
<b>R: Responsable</b>	Rol encargado de realizar la actividad.				
<b>E: Encargado</b>	Rol que aprueba el trabajo realizado por el responsable.				
<b>C: Consultado</b>	Aquellos expertos que son consultados sobre algún aspecto de la tarea.				
<b>I: Informado</b>	Aquellas personas que deben ser informadas sobre la evolución de la tarea.				

Figura 31: Matriz RECI del proceso de registro de clientes rediseñado

Como se ve en la figura 31, se visualizan los roles y actividades según lo definido a lo largo del rediseño, donde cada área se especializa en ser el responsable de las nuevas actividades, donde BI & Analytics realiza las actividades asociadas a los datos y es consultado en actividades que conciernen de la verificación de estos, mientras que el área comercial se encarga de asegurarse de la calidad de la información del proceso, verificando la correcta filtración y detección de clientes con diferente nombre en la competencia.

Finalizando con el rediseño del proceso, se envió vía email un comunicado de los cambios del proceso rediseñado a todos los involucrados del proyecto (véase en el anexo 11 ocultando su contenido por información confidencial entre los cambios del proyecto), utilizando el archivo compartido para centralizar toda la documentación generada por el proyecto.

### **Implementación final**

Para la etapa de implementación final, siguiendo con el plan de implementación, se generó la documentación de cómo trabajar con los cambios realizados al proceso de registro y los detalles técnicos del modelo de identificación, además de un manual para sus casos de uso del área de BI & Analytics. En cuanto a las capacitaciones, se coordinaron dos reuniones (rediseño del proceso y utilización del modelo de identificación) según la planificación inicial.

En cuanto a las últimas actividades de la implementación final, la asignación de responsabilidades, fueron definidas a lo largo del rediseño para el proceso de registro de clientes y para el modelo se fijaron en las reuniones semanales de BI & Analytics. En relación al control del modelo, se configuró dentro de la plataforma de Fury alertas para errores de ejecución y tiempos de desarrollo mayores a 9 horas, donde se envían automáticamente emails a una línea de correo del equipo de BI & Analytics para que todo el equipo esté al tanto del funcionamiento del modelo.

## Resultados

En relación al rediseño del proceso se registro de clientes, luego de las iteraciones basadas en la mejora del proceso en base a la especialización de roles, conceptos de BPM y las metodologías Design y Lean Thinking, se alcanzaron los siguientes resultados en el último modelamiento del proceso.

- Tiempo de flujo del proceso de registro de clientes (días): 12 días aproximadamente.
- Valor monetario de horas de trabajo de analistas comerciales en Gestión Comercial (UM): 365 UM aproximadamente.

Basándose en los resultados anteriores y en las mediciones realizadas durante el rediseño, se puede ver una reducción a 12 días del tiempo de flujo del proceso de registro de clientes, cumpliendo el objetivo SMART de reducción a dos semanas del registro de clientes (a falta de más mediciones), además del objetivo específico de agilización del proceso, rediseñando por completo el subprocesso de “Gestión Comercial” con un foco más estratégico transformándose en una “Verificación comercial de los datos”. Además, con respecto a las primeras mediciones del proceso se tuvo una reducción de 40 días en el tiempo de flujo del proceso y una reducción de 432 horas de trabajo de los analistas comerciales, significando un ahorro de \$5.600 unidades monetarias aprox., representando una reducción del 93% aproximadamente del costo empresa dedicados a los analistas comerciales.

Con respecto a la contactación de clientes del ciclo para la evaluación de la efectividad del proceso en comparación al ciclo del trimestre anterior, todavía está en desarrollo el ciclo actual, teniendo una diferencia porcentual actual del 7% con respecto al ciclo anterior, donde ya se visualiza una diferencia positiva. Todos los resultados anteriores, muestran cuantitativamente el aporte del desarrollo del proyecto al objetivo de agilización del registro de clientes, aportando al desarrollo ágil para la contactación de clientes donde se asegure la calidad de su información dentro del proceso.

En cuanto al objetivo de reducir a 0 la cantidad de clientes duplicados y existentes en el registro de Salesforce, con el desarrollo realizado en la aplicación del modelo de identificación y el filtrado en el armado de las bases de datos, ya no se presentan cuentas duplicadas en el registro, donde analizando la última implementación del modelo comentada anteriormente, no se presentaron duplicados, mientras que existentes se presentaron 2,6% de cuentas de usuarios, donde dicho porcentaje, según la información levantada sobre la conexión del CRM Salesforce y Bigquery son errores propios de la visualización de los datos de los usuarios en la base de datos de clientes de MELI, lo cual actualmente está siendo filtrado y analizado con el área encargada del CRM.

En relación a la identificación de clientes de la competencia, los cuales no pueden ser identificados mediante sus datos personales, se desarrolló el modelo de identificación de nombres de clientes, donde podemos ver su estructura en la siguiente figura.

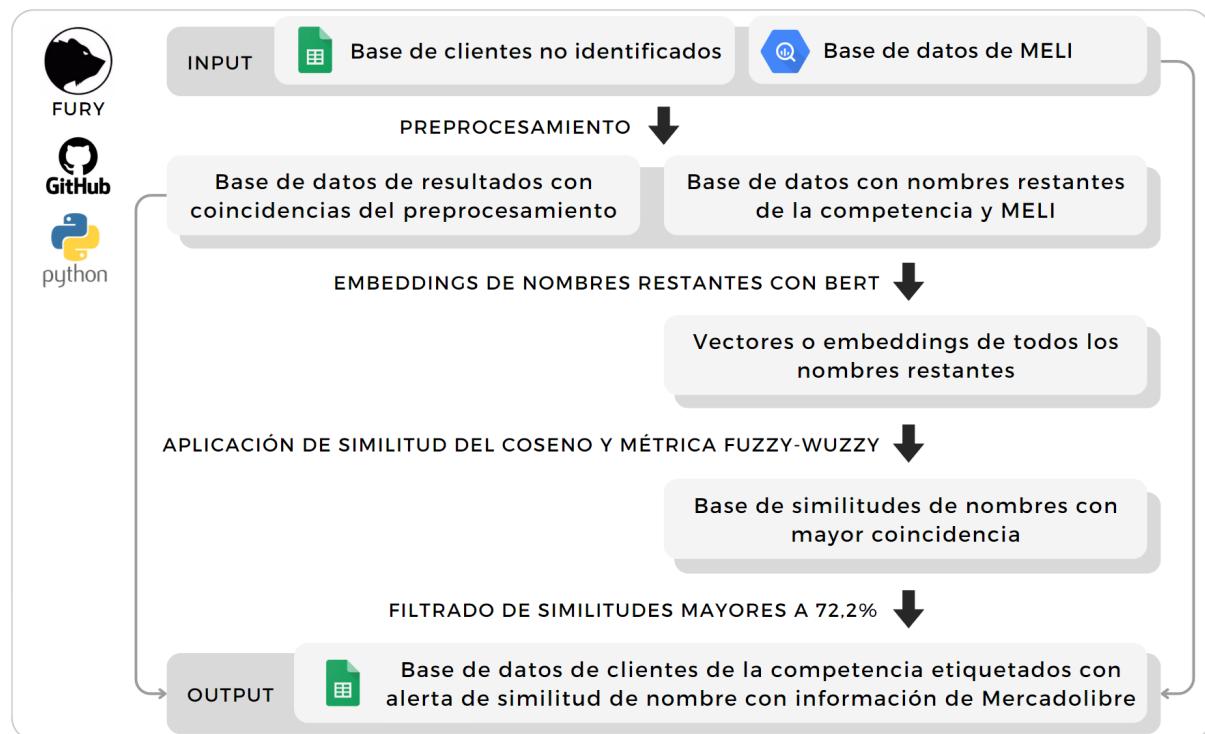


Figura 32: Diagrama del modelo de identificación

Como se aprecia en la figura 32, se ve el flujo del modelo desarrollado en python y alojado en Fury con conexión a Github para la subida de archivos por parte del área de BI & Analytics. Se pueden ver las bases de datos que entran en el input y todos los pasos desarrollados a lo largo de los sprints. Se consideran alertas de similitud en los clientes nuevos que presenten una coincidencia mayor al 72,2% con el nombre de Mercadolibre. Por último, se genera una base de datos de clientes con la etiqueta y se dispone junto con información de dichos clientes al proceso de registro de clientes.

En cuanto al rendimiento del desarrollo del modelo, se presentaron los siguientes resultados según sus medidas de desempeño. En primer lugar se tiene una “Exactitud” del 82,7% y “F1-Score” del 88,2%, siendo valores elevados en contraste con los modelos de embeddings basados en BERT, donde el modelo “multi-qa-MiniML-L6-cos-v1” utilizado por su rendimiento, para tareas de vectorización de texto cuenta con una exactitud media del 64,33% (Sentence-Transformers, 2023), logrando complementar dicho rendimiento con el preprocesamiento y métricas difusas, logrando un F1-score elevado para el primer desarrollo del algoritmo, basándose en un equilibrio entre los tipos de errores I y II, evitando contar con falsos positivos y falsos negativos en la predicción.

Es importante destacar que estos valores fueron obtenidos con la base destinada al desarrollo, donde se tuvo un tiempo de flujo del modelo de 4 minutos, que en la implementación posterior a una base de clientes completa aumentó a 1,5 horas identificando un 66% de clientes de la base de clientes de la competencia no identificados por el armado de la base de datos, el cual actualmente está siendo revisado para determinar la efectividad de su identificación con el cálculo de las métricas de rendimiento de exactitud, precisión, sensibilidad, y así el F1-Score.

Desde un foco cualitativo el modelo de clientes no sólo logró reducir el tiempo de la tarea manual de identificación de clientes del área comercial agilizando el proceso anterior a la contactación con los clientes, sino también incorporar la participación del área de BI en el negocio aplicando en él técnicas de NLP, el cual es un campo que está en constante desarrollo en el rubro de la inteligencia artificial e innovación provenientes en estudios de data science, y puede tener diversos usos en el futuro donde se pueden aprovechar sus capacidades para otros casos de aplicación dentro del proyecto de Inteligencia Competitiva.

## **Conclusiones y discusión**

En cuanto a las conclusiones del proyecto, se tiene en primer lugar los puntos negativos a destacar, donde el desarrollo necesario para cumplir con el alcance de las iteraciones, si bien mediante el trabajo iterativo permitió refinar el desarrollo de las soluciones, también limitó las posibles mediciones y análisis posterior del rediseño del proceso implementado.

Lo anterior, finalmente afectó al alcance del proyecto en términos de capacidad y tiempo, sin embargo, al tener la posibilidad de continuar midiendo el proceso también se tiene la oportunidad de refinar la medición constante de resultados, con un foco mayor en la implementación final del proyecto. Por ejemplo, se podría aplicar nuevamente el Modelo de Madurez de Hammer y otros análisis estadísticos que profundicen el análisis del proceso implementado.

Por otro lado, dentro de los puntos positivos que agilizaron el desarrollo del proyecto, la conexión con el negocio propia de las metodologías Design Thinking, Lean Thinking y CRISP-DM, fue fundamental a la hora de aplicar los elementos encontrados en el estado del arte para el desarrollo de la solución, que en conjunto con un trabajo iterativo y coordinación con los involucrados se pudieron desarrollar cambios efectivos al proceso y poder desarrollar un modelo de identificación. Dentro de lo anterior, la definición del estado del arte fue fundamental a la hora de conectar el problema con la industria, levantando diversas alternativas, desde casos aplicados hasta el complemento con la literatura, como es el caso del rediseño del proceso con foco en la especialización y los conceptos de BPM o la aplicación de modelos de BERT en NLP.

Siguiendo lo anterior, si bien hubo dificultades para la búsqueda de soluciones para el problema de identificación de clientes, el rubro de NLP fue un descubrimiento fundamental que permitió ampliar la gama de procesamiento de datos con la metodología iterativa del proyecto. Todo lo anterior, permitió poder realizar cambios en el proceso y alcanzar el objetivo SMART planteado según las mediciones realizadas, pudiendo utilizar las métricas establecidas en el proyecto como guía para el desarrollo de la solución y del proyecto en su totalidad.

En términos del contexto de la empresa, tanto su enfoque colaborativo para reforzar el entendimiento del negocio mencionado durante el desarrollo de todo el proyecto, como sus capacidades, entornos y tecnologías establecidas en Mercadolibre, permitieron agilizar la implementación del modelo de identificación. Por ejemplo, la plataforma de Fury que entrega un desarrollo en máquinas virtuales potentes, conexiones para alertas y almacenamiento de la información, y el uso de tecnologías conocidas como Github y Jupyter.

Todo lo anterior, permitió desarrollar e implementar el proyecto de forma efectiva, aportando con sus resultados a uno de los proyectos más importantes dentro de VIS, siendo fundamental para los objetivos de negocio de Mercadolibre como líder del negocio de los mercados de vehículos e inmuebles.

## Bibliografía

1. Velázquez, A. (2023). ¿Qué es el diagrama de Pareto? QuestionPro. <https://www.questionpro.com/blog/es/diagrama-de-pareto/>
2. Rodriguez, J. (2023). Qué es el diagrama de Ishikawa, para qué sirve, cómo crearlo y ejemplos. Blog de HubSpot. <https://blog.hubspot.es/sales/diagrama-ishikawa>
3. Scikit-Learn. (s.f.). Ejemplo de Precisión y Recall. Scikit-Learn. [https://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_precision\\_recall.html](https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html)
4. Scikit-Learn. (s.f.). sklearn.metrics.precision\_score. Scikit-Learn Documentation. [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html)
5. Terreros, D. HubSpot. (2023). Design Thinking. HubSpot Blog. <https://blog.hubspot.es/marketing/design-thinking>
6. Instituto Agile. (2022). Lean Thinking: 5 Principios de Lean. Instituto Agile. <https://www.institutoagile.com/post/lean-thinking-5-principios-de-lean>
7. López, J. F. (2022). Medidas de dispersión. Economipedia. <https://economipedia.com/definiciones/medidas-de-dispersión.html>
8. Captación de clientes: ¿Qué es y cómo realizarla? (2022). Salesforce. <https://www.salesforce.com/mx/blog/2022/06/captacion-de-clientes.html>
9. Narvaez, M. (2023). ¿Qué es la segmentación de clientes? QuestionPro. <https://www.questionpro.com/blog/es/segmentacion-de-clientes/>
10. Yapo Data Archivos - El blog de Yapo. (s. f.). El Blog de Yapo. <https://blog.yapo.cl/category/yapodata/>
11. F. Fernandez, F. F. (2017, 27 abril). Estudio de Mercado. Google Books. Recuperado 4 de octubre de 2023, de [https://books.google.cl/books/about/Estudio\\_de\\_Mercado.html?id=yuskDwAAQBAJ&redir\\_esc=y](https://books.google.cl/books/about/Estudio_de_Mercado.html?id=yuskDwAAQBAJ&redir_esc=y)
12. Escudero, J. (2021, 4 junio). Estrategias para llegar a más clientes. Emprendedores. <https://emprendedores.es/gestion/estrategias-clientes-vender-mas/>
13. Carvajal, R. (2023, 30 abril). Frenetic: la compañía que acelera la electrificación en el mundo. La Razón. [https://www.larazon.es/economia/startups/frenetic-compania-que-acelera-electrificacion-mundo\\_20230430644de8b773ab380001e692d8.html](https://www.larazon.es/economia/startups/frenetic-compania-que-acelera-electrificacion-mundo_20230430644de8b773ab380001e692d8.html)
14. We-Prospect. (2022, 26 agosto). Especialización de roles y Estructura en el Departamento de Ventas - We-Prospect. We-Prospect. <https://www.we-prospect.com/especializacion-de-roles-y-estructura-en-el-departamento-de-ventas/>
15. Los RECI como documento de planificación de procesos – la broma. (2009, 16 enero). <http://www.labroma.org/blog/2009/01/16/los-recci-como-documento-de-planificacion-de-procesos/>
16. BPMN Specification - Business process model and notation. (s. f.). <https://www.bpmn.org/>
17. Tutorial de BPMN y BPMN 2.0. (s. f.). Lucidchart. <https://www.lucidchart.com/pages/es/bpmn-bpmn-20-tutorial>
18. Sitio de Bizagi. (2023, 11 septiembre). Bizagi - líder en automatización inteligente de Procesos. <https://www.bizagi.com>.

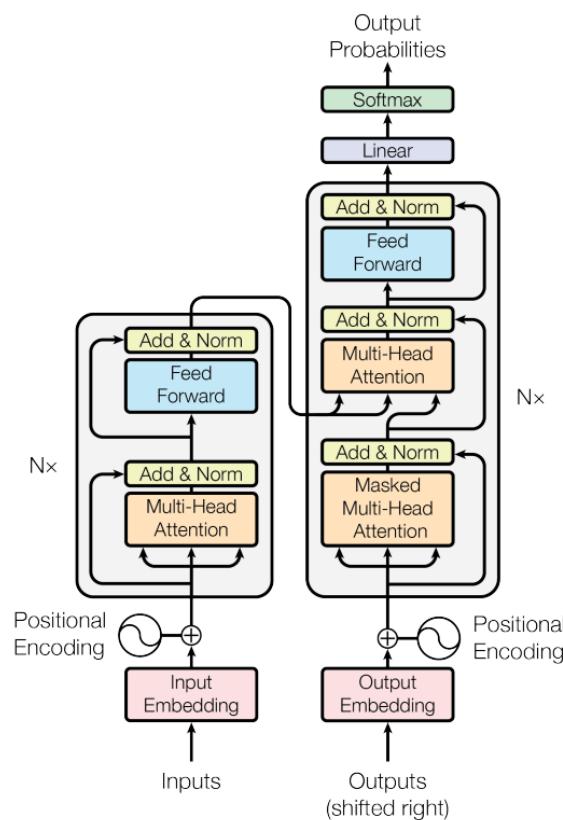
19. Kirchmer, Mathias & Franz, Peter. (2016). The Process of Process Management. [https://www.researchgate.net/publication/294091113\\_The\\_Process\\_of\\_Process\\_Management](https://www.researchgate.net/publication/294091113_The_Process_of_Process_Management)
20. Rucker, B., & Freund, J. (2019). Real-Life BPMN (4th Edition): Includes an Introduction to DMN (4.a ed.). Independently Published.
21. Nijhuis, M. (2022, 5 marzo). Company name matching - DNB — Data Science Hub - Medium. Medium. <https://medium.com/dnb-data-science-hub/company-name-matching-6a6330710334>
22. GraphEverywhere, E. (2019). Algoritmo de similitud de coseno. GraphEverywhere. <https://www.grapheverywhere.com/algoritmo-de-similitud-de-coseno/#:~:text=La%20similitud%20del%20coseno%20es>
23. Thornton, C. (2021, 15 diciembre). HMNI: Fuzzy name matching with Machine Learning | Towards Data Science. Medium. <https://towardsdatascience.com/fuzzy-name-matching-with-machine-learning-f09895dce7b4>
24. Babel Street. (2023, 28 julio). Name Matching Algorithms - Rosette Text Analytics. Rosette Text Analytics. <https://www.rosette.com/name-matching-algorithms/>
25. Team, T. A. (2020). Text mining in Python: steps and examples. Towards AI. <https://towardsai.net/p/artificial-intelligence/text-mining-in-python-steps-and-examples-78b3f8fd913b>
26. Incorporaciones. (s. f.). Google for Developers. <https://developers.google.com/machine-learning/crash-course/embeddings/video-lecture?hl=es-419>
27. ¿Qué es el procesamiento de lenguaje natural? - Explicación del procesamiento de Lenguaje Natural - AWS. (s. f.). Amazon Web Services, Inc. <https://aws.amazon.com/es/what-is/nlp/#:~:text=tareas%20de%20NLP%3F,-%C2%BFQu%C3%A9%20es%20la%20NLP%3F,y%20comprender%20el%20lenguaje%20humano.>
28. ¿Qué es una red neuronal? - Explicación de las redes neuronales artificiales - AWS. (s. f.). Amazon Web Services, Inc. <https://aws.amazon.com/es/what-is/neural-network/>
29. De La Cigoña, J. R. F. (2023, 13 noviembre). Tasa Interna de Retorno (TIR): ¿Qué es y cómo se calcula? Sage Advice España. <https://www.sage.com/es-es/blog/tasa-interna-de-retorno-tir-que-es-y-como-se-calcula/>
30. Unir, V. (2022, 18 julio). El modelo CAPM: ¿Cómo calcular la tasa de retorno de un activo financiero? UNIR. <https://www.unir.net/empresa/revista/modelo-capm/#:~:text=El%20modelo%20CAPM%20siglas%20de,en%20funci%C3%B3n%20del%20riesgo%20asumido.>
31. elEconomista.es. (2023, 22 noviembre). Riesgo del VAN : Qué es - Diccionario de Economía. <https://www.eleconomista.es/diccionario-de-economia/riesgo-del-van>
32. Guidi, L. R. (2023, 30 agosto). Valuación de Mercado Libre Inc. (MELI) [2do trimestre 2023]. <https://www.linkedin.com/pulse/valuaci%C3%B3n-de-mercado-libre-inc-meli-2do-trimestre-2023-guidi/?originalSubdomain=es>
33. Comprensión de la matriz de confusión y cómo implementarla en Python. (2023, 1 septiembre). DataSource.ai. <https://www.datasource.ai/es/data-science-articles/comprehension-de-la-matriz-de-confusion-y-como-implementarla-en-python>

34. Briggs, J. (2022, 7 enero). NLP Similarity Metrics | Towards Data science. Medium. <https://towardsdatascience.com/similarity-metrics-in-nlp-acc0777e234c>
35. Semantic textual similarity — Sentence-Transformers documentation. (2023). <https://www.sbert.net/examples/training/sts/README.html>
36. Pretrained Models — Sentence-Transformers documentation. (2023). [https://www.sbert.net/docs/pretrained\\_models.html](https://www.sbert.net/docs/pretrained_models.html)
37. Fuzzywuzzy. (2020, 13 febrero). PyPI. <https://pypi.org/project/fuzzywuzzy/>
38. Dot CSV. (2021, 27 septiembre). ¿Qué es un TRANSFORMER? La red neuronal que lo cambió TODO! [Vídeo]. YouTube. <https://www.youtube.com/watch?v=aL-EmKuB078>
39. Vaswani, A. (2017, 12 junio). Attention is all you need. arXiv.org. <https://arxiv.org/abs/1706.03762>
40. CodeEmporium. (2020, 4 mayo). BERT Neural Network - EXPLAINED! [Vídeo]. YouTube. <https://www.youtube.com/watch?v=xI0HHN5XKD0>
41. Briggs, J. (2022a, enero 6). Sentence similarity with BERT | towards data science. Medium. <https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1>

## Anexos

### 1. Arquitectura del modelo de transformers

En la siguiente figura se muestra la arquitectura que sigue el modelo BERT, donde a la izquierda se encuentra el paso de encode para realizar la vectorización o embedding de palabras, utilizando el positional encoding que permite trabajar simultáneamente y no perder el contexto de los nombres ingresados en el embedding.



## 2. Flujo de ingresos y costos de los escenarios pesimista y optimista del proyecto.

### 2.1. Escenario pesimista

Flujo de ingresos y costos para evaluación económica del proyecto (unidades monetarias)					
Ingresos y costos / Periodos	Trimestre 0	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
<b>Ingresos</b>					
Ahorro de costos de empresa de horas de trabajo	-	\$3.640	\$3.640	\$3.640	\$3.640
<b>Costos de mantención</b>					
Costo de plan de entorno de programación Fury	-	-\$510	-\$510	-\$510	-\$510
Costo empresa de horas de trabajo para revisión de alertas de Fury	-	-\$113	-\$113	-\$113	-\$113
Costo empresa de revisión retrospectiva del proceso de registro de clientes	-	-\$100	-\$100	-\$100	-\$100
<b>Inversión fija</b>					
Costo de datos para MVP del modelo	-\$1.540				
Costo empresa por horas de trabajo del desarrollador del proyecto	-\$162				
Costo empresa por horas de trabajo por personal involucrado en el proyecto	-\$117				
Costo empresa por horas de trabajo para capacitación del proceso	-\$821				
Costo empresa por horas de trabajo para capacitación del modelo	-\$332				
<b>Flujo total</b>	<b>-\$2.972</b>	<b>\$2.917</b>	<b>\$2.917</b>	<b>\$2.917</b>	<b>\$2.917</b>

### 2.2. Escenario optimista

Flujo de ingresos y costos para evaluación económica del proyecto (unidades monetarias)					
Ingresos y costos / Periodos	Trimestre 0	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
<b>Ingresos</b>					
Ahorro de costos de empresa de horas de trabajo	-	\$5.395	\$5.395	\$5.395	\$5.395
<b>Costos de mantención</b>					
Costo de plan de entorno de programación Fury	-	-\$510	-\$510	-\$510	-\$510
Costo empresa de horas de trabajo para revisión de alertas de Fury	-	-\$113	-\$113	-\$113	-\$113
Costo empresa de revisión retrospectiva del proceso de registro de clientes	-	-\$100	-\$100	-\$100	-\$100
<b>Inversión fija</b>					
Costo de datos para MVP del modelo	-\$1.540				
Costo empresa por horas de trabajo del desarrollador del proyecto	-\$162				
Costo empresa por horas de trabajo por personal involucrado en el proyecto	-\$117				
Costo empresa por horas de trabajo para capacitación del proceso	-\$821				
Costo empresa por horas de trabajo para capacitación del modelo	-\$332				
<b>Flujo total</b>	<b>-\$2.972</b>	<b>\$4.672</b>	<b>\$4.672</b>	<b>\$4.672</b>	<b>\$4.672</b>

### 3. Detalle de tareas de la Carta Gantt del proyecto

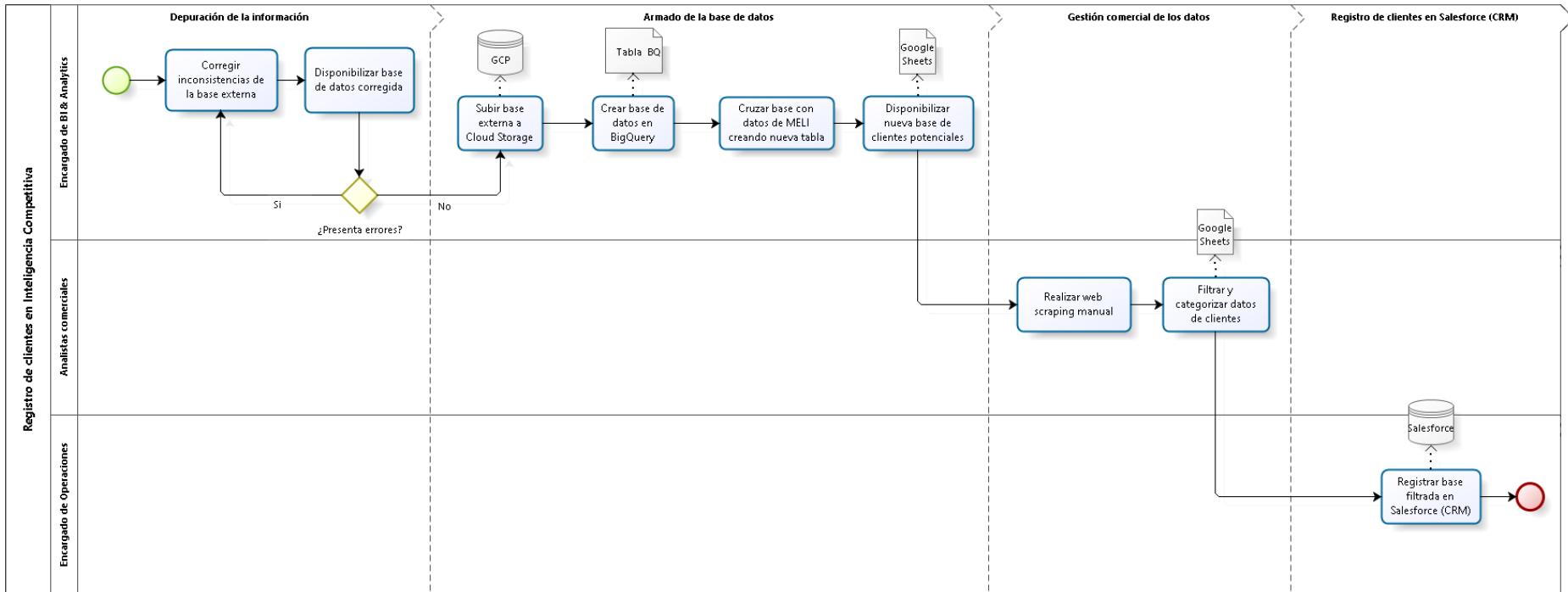
ETAPAS	DETALLE DE TAREAS
0	<b>Documentación</b>
1	<b>Levantamiento del problema</b> <ul style="list-style-type: none"> <li><b>Entendimiento del negocio</b></li> <li>Entendimiento del contexto y la necesidades del usuario</li> <li>Definición del problema y sus procesos asociados</li> <li>Definición del alcance y de los objetivos</li> <li>Revisión del estado del arte y alternativas de solución</li> </ul>
2	<b>Desarrollo de la solución</b> <ul style="list-style-type: none"> <li><b>Recolección de datos</b> <ul style="list-style-type: none"> <li>Definición de fuentes de datos</li> <li>Disponibilización y recolección de datos</li> </ul> </li> <li><b>Preprocesamiento de los datos</b> <ul style="list-style-type: none"> <li>Limpieza de datos faltantes o duplicados con text meaning</li> <li>Estandarización de formato de datos</li> <li>Chequeo de consistencia de formato</li> </ul> </li> <li><b>Sprints semanales de modelamiento</b> <ul style="list-style-type: none"> <li>Insights del comportamiento de datos</li> <li>Desarrollo del modelo y aplicación de algoritmos</li> <li>Construcción del flujo de los datos</li> <li>Aplicación de métricas y test estadísticos para comparación</li> <li>Evaluación de performance y recopilar feedback</li> </ul> </li> <li><b>Rediseño de procesos</b> <ul style="list-style-type: none"> <li>Reuniones para coordinación con área comercial</li> <li>Fijar intereses comerciales de los datos</li> <li>Definir nuevos roles y tiempos de los nuevos procesos</li> <li>Documentar nuevos procesos en base a la coordinación</li> </ul> </li> </ul>
3	<b>Despliegue de la solución</b> <ul style="list-style-type: none"> <li><b>Automatización de la solución</b> <ul style="list-style-type: none"> <li>Definición de entregable con área operativa y comercial</li> <li>Automatización de disponibilización de resultados en un job</li> </ul> </li> <li><b>Reingeniería de procesos</b> <ul style="list-style-type: none"> <li>Consolidación del nuevo proceso con la solución desarrollada</li> <li>Definición de roles para uso del job</li> </ul> </li> <li><b>Implementación</b> <ul style="list-style-type: none"> <li>Generar documentación</li> <li>Capacitaciones</li> </ul> </li> <li><b>Marcha blanca</b></li> </ul>

#### 4. Modelo de madurez de procesos de Hammer

		P-1	P-2
<b>Diseño</b>	Propósito	El proceso no se ha diseñado de punta a cabo. Los ejecutivos utilizan el diseño que venía rigiendo como contexto para la mejora del desempeño funcional.	El proceso se ha rediseñado completamente para mejorar su desempeño.
	Contexto	Se han identificado los insumos, productos, proveedores y clientes del proceso.	Las necesidades de los clientes del proceso son conocidas y hay acuerdo sobre ellas.
	Documentación	La documentación del proceso es principalmente funcional, pero identifica las interconexiones entre las organizaciones involucradas en ejecutar el proceso.	Hay documentación completa del diseño del proceso.
<b>Ejecutores</b>	Conocimiento	Los ejecutores pueden dar nombre al proceso que ejecutan e identificar los indicadores clave de su desempeño.	Los ejecutores pueden describir el flujo global del proceso; cómo su trabajo afecta a los clientes, a otros empleados del proceso y el desempeño del proceso; y los niveles de desempeño reales y requeridos.
	Destrezas	Los ejecutores son diestros en técnicas de resolución de problemas y de mejora de procesos.	Los ejecutores son diestros en trabajo en equipo y en gestionarse personalmente
	Conducta	Los ejecutores profesan cierta lealtad al proceso pero deben máxima lealtad a su función.	Los ejecutores tratan de seguir el diseño del proceso, ejecutarlo correctamente y trabajar en formas que permitan a otras personas que ejecutan el proceso hacer eficazmente su trabajo.
<b>Responsable</b>	Identidad	El responsable del proceso es una persona o grupo encargado informalmente de mejorar el desempeño del proceso.	Los líderes de la empresa han creado un papel oficial de responsable del proceso y han colocado en ese puesto a un alto ejecutivo con influencia y credibilidad.
	Actividades	El responsable identifica y documenta el proceso, lo comunica a todos los ejecutores y patrocina pequeños proyectos de cambio.	El responsable comunica las metas del proceso y una visión de su futuro, patrocina esfuerzos de rediseño y mejora, planifica su implementación y se asegura de que se cumpla el diseño del proceso.
	Autoridad	El responsable hace lobby por el proceso, pero solamente puede alentar a los ejecutivos funcionales a hacer cambios.	El responsable puede reunir a un equipo de rediseño de proceso e implementar el nuevo diseño y tiene cierto control sobre el presupuesto de tecnología para el proceso.
<b>Infraestructura</b>	Sistemas de información	El proceso es apoyado por sistemas fragmentados de TI.	El proceso es apoyado por un sistema de TI creado a partir de componentes funcionales.
	Sistemas de recursos humanos	Los ejecutivos funcionales recompensan el logro de excelencia funcional y la resolución de problemas funcionales en un contexto de proceso.	El diseño del proceso impulsa los roles, las descripciones de cargo y los perfiles de competencias. La capacitación se basa en documentación de proceso.
<b>Indicadores</b>	Definición	El proceso tiene ciertos indicadores básicos de costo y calidad .	El proceso tiene indicadores de extremo a extremo derivados de los requerimientos de los clientes.
	Usos	Los ejecutivos usan los indicadores del proceso para monitorear su desempeño, identificar las causas fundamentales de desempeño defectuoso e impulsar mejoras funcionales.	Los ejecutivos usan los indicadores del proceso para comparar su desempeño con los benchmarks, el desempeño mejor en su clase y las necesidades de los clientes, y para fijar objetivos de desempeño.

P-3	P-4	P-1	P-2	P-3	P-4
El proceso se ha diseñado para ajustarse a otros procesos de la empresa y a sus sistemas de TI a fin de optimizar el desempeño de la empresa.	El proceso se ha diseñado para ajustarse a los procesos de los clientes y los proveedores a fin de optimizar el desempeño interempresa.				
El responsable del proceso y los responsables de los otros procesos con los que interactúa el proceso han definido sus expectativas mutuas de desempeño.	El responsable del proceso y los responsables de los procesos de los clientes y proveedores con los que interactúa el proceso han definido sus expectativas mutuas de desempeño.				
La documentación del proceso describe las interacciones del proceso con otros procesos, y sus expectativas respecto a éstos, y vincula al proceso con el sistema y con la arquitectura de datos de la empresa.	Una representación electrónica del diseño del proceso apoya su desempeño y gestión, y permite analizar los cambios ambientales y las reconfiguraciones de proceso.				
Los ejecutores están familiarizados tanto con los conceptos fundamentales de negocios como con los impulsos del desempeño de la empresa, y pueden describir cómo afecta su trabajo a otros procesos y al desempeño de la empresa.	Los ejecutores están familiarizados con las tendencias en el sector de la empresa y pueden describir cómo afecta su trabajo al desempeño interempresa.				
Los ejecutores son diestros en la toma de decisiones de negocios.	Los ejecutores tienen capacidades de gestión e implementación del cambio.				
Los ejecutores se esfuerzan por asegurarse de que el proceso entregue los resultados necesarios para lograr las metas de la empresa.	Los ejecutores buscan señales de que el proceso debería cambiar y proponen mejoras al proceso.				
El responsable da máxima prioridad al proceso en términos de asignación de tiempo, preocupación y metas personales.	El responsable es miembro de la unidad de más alto rango en la toma de decisiones de la empresa.				
El responsable colabora con otros responsables de proceso para integrar procesos y lograr las metas de la empresa.	El responsable desarrolla un plan estratégico de extensión del proceso, participa en planificación estratégica a nivel de empresa y colabora con sus contrapartes que trabajan donde clientes y proveedores para patrocinar iniciativas interempresa de rediseño de proceso.				
El responsable controla los sistemas de TI que apoyan el proceso y cualquier proyecto que cambie el proceso, y tiene cierta influencia sobre las asignaciones y evaluaciones de personal así como sobre el presupuesto del proyecto.	El responsable controla el presupuesto del proceso y ejerce fuerte influencia sobre las asignaciones y la evaluación de personal.				
El proceso es apoyado por un sistema integrado de TI, diseñado teniendo en mente el proceso y adhiriendo a los estándares de la empresa.	El proceso es apoyado por un sistema de TI con arquitectura modular, que se adhiere a los estándares del sector para la comunicación interempresa.				
Los sistemas de contratación, desarrollo, reconocimiento y recompensa enfatizan las necesidades y los resultados del proceso, y los equilibran con las necesidades de la empresa.	Los sistemas de contratación, desarrollo, recompensa y reconocimiento refuerzan la importancia de la colaboración intra e interempresarial, el aprendizaje personal y el cambio organizacional.				
Los indicadores del proceso, así como los indicadores entre procesos, se han derivado de las metas estratégicas de la empresa.	Los indicadores del proceso se han derivado de metas interempresariales.				
Los ejecutivos presentan los indicadores a los ejecutores de proceso para motivar y crear conciencia. Usan tableros basados en indicadores para la gestión cotidiana del proceso.	Los ejecutivos revisan y actualizan regularmente los indicadores y objetivos del proceso y los usan al planificar la estrategia de la empresa.				

## 5. Modelamiento BPMN de la situación As-Is del registro de clientes en Bizagi



## 6. Código en python del preprocessamiento desarrollado

### Preprocesamiento

```
# Función para preprocesar el nombre del cliente estandarizando el texto
def preprocessar(_lead):
    comunes=['propiedades','negocios','inmobiliario','inmobiliaria','inmobiliarias','inmobiliarios','servicios']
    lead = _lead.lower().strip() # eliminamos minúsculas y espacios laterales
    lead = re.sub(' +', ' ', lead) # quitamos múltiples espacios
    lead = unicodedata.normalize('NFD', lead).encode('ascii', 'ignore').decode("utf-8") # elimina caract. esp.
    lead = re.sub(r'\W\s+', '', lead) # elimina caracteres '.', ',' ...
    lead = lead.replace('_', ' ') # reemplaza los _ por espacios
    lead = eval("lead"+''.join([".replace(\""+str(i)+"\",'')" for i in comunes]))
    lead = basename(lead) # elimina elementos comunes nombres de empresas tipo "Ltd"
    return lead

# Función para crear una lista con todos los nombres estandarizados
def crear_lista_preprocesados(df):
    lista = []
    for i in range(df.count()):
        lista.append(preprocessar(df[i]))
    return lista

# Función para hacer el match exacto en el preprocessamiento recuperando los ids
def match_preprocesamiento(lista_leads,lista_meli,cust_id):
    cruce_ids = [None for i in lista_leads]
    for i in range(len(lista_leads)):
        for j in range(len(lista_meli)):
            if(lista_leads[i] == lista_meli[j]):
                cruce_ids[i] = cust_id[j]
    return cruce_ids

# Preprocesamos ambas bases de datos
leads_preprocesados = crear_lista_preprocesados(nombres_leads)
meli_preprocesados = crear_lista_preprocesados(nombres_meli['nombre_meli'])

# Recuperamos los ids del cruce exacto del preprocessamiento y creamos un dataframe con el cruce
ids = match_preprocesamiento(leads_preprocesados,meli_preprocesados,nombres_meli['cust_id'])
df_cruce = pd.DataFrame({'nombres_lead': nombres_leads, 'cust_id_cruce': ids}).merge(nombres_meli, left_on='cust_id_cruce', right_on='cust_id')
df_cruce = df_cruce.where(pd.notnull(df_cruce), None)
df_cruce = df_cruce.loc[_df_cruce['cust_id_cruce'].notnull(),:]
df_cruce = df_cruce.assign(similarity=1)
df_cruce = df_cruce.drop(columns=['cust_id_cruce']).reset_index(drop=True)

# Creamos un dataframe con los restantes y preprocesamos los restantes
df_restantes = _df_cruce.loc[_df_cruce['cust_id_cruce'].isnull(),'nombres_lead'].reset_index(drop=True)
restantes_preprocesados = crear_lista_preprocesados(df_restantes)
cant_restantes = len(restantes_preprocesados)

# Creamos una lista con los nombres restantes de la competencia y los nombres de MELI
nombres = restantes_preprocesados+meli_preprocesados

# Revisamos la cantidad de nombres identificados
porcentaje_identificados = (1-(cant_restantes/len(nombres_leads)))
print('% de nombres identificados en el preprocessamiento: '+str(porcentaje_identificados))

% de nombres identificados en el preprocessamiento: 0.19515389652914206
```

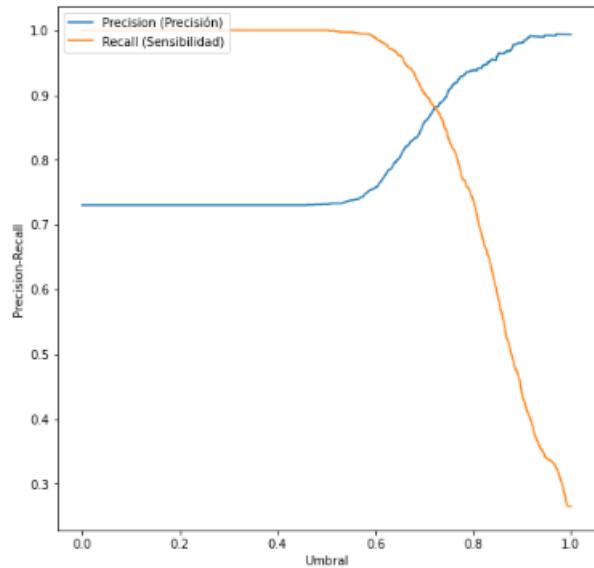
## 7. Código en python de la definición del umbral para la predicción del modelo

```

def fijar_umbral(prob, y_true):
    umbral=0
    y_test=[1 if x==1 else 0 for x in y_true ]
    y_pred=np.zeros(len(prob))
    precision=[]
    recall=[]
    min_dist=1
    x=np.arange(0.0, 1.001, 0.001)
    for i in x:
        y_pred[prob>=i]=1
        y_pred[prob<i]=0
        precision_1=precision_score(y_test, y_pred)
        recall_1=recall_score(y_test, y_pred)
        precision.append(precision_1)
        recall.append(recall_1)
    for i in range(len(x)):
        dist=np.absolute(precision[i]-recall[i])
        if (dist<=min_dist and precision[i]>0 and recall[i]>0):
            min_dist=dist
            umbral=x[i]
    plt.figure(figsize=(8,8))
    plt.plot(x, precision)
    plt.plot(x, recall)
    plt.xlabel("Umbral")
    plt.ylabel("Precision-Recall")
    plt.legend(['Precision (Precisión)', 'Recall (Sensibilidad)'], loc='upper left')
    plt.show()
    print("Umbral: "+str(umbral))
    return umbral

```

```
umbral=fijar_umbral(df_resultados["similarity"],df_resultados["flag"])
```



## 8. Notebook del modelo desarrollado de identificación de clientes

### Modelo de identificación de nombres de sellers (MINS) para Inteligencia Competitiva

Creado por Cristóbal Salas R. - Área de BI & Analytics de VIS MELI

Fecha de creación: 25 de septiembre 2023

Última fecha de actualización: 26 de noviembre 2023

#### Descripción

En este notebook se desarrolla un modelo de identificación de nombres de sellers que lleguen en bases de datos con datos no estructurados con el objetivo de poder realizar un matching con nombres de sellers de MELI y detectar coincidencias. Para esto se desarrolla un modelo que aplique un preprocesamiento para la estandarización de texto, aplicando posteriormente un modelo de embedding basado en el transformer BERT (vectorización de texto) y métricas de distancias para encontrar una probabilidad de coincidencia y, bajo un umbral, levantar una alarma de "Posible Seller MELI" y tomar esto en cuenta en el análisis de sellers en Inteligencia Competitiva.

#### Módulos de Python utilizados

- Numpy
- Pandas
- Unicodedata
- Re
- Cleanco
- Scikit-learn (sklearn)
- Sentences-Transformer
- Matplotlib (pyplot)

#### Instalación de módulos no contenidos en Fury por defecto

```
!pip install cleanco
!pip install fuzzywuzzy
!pip install python-Levenshtein
!pip install transformers
```

#### Carga de módulos/librerías

```
import numpy as np
import pandas as pd
import unicodedata
import re
from cleanco import basename
from sentence_transformers import SentenceTransformer
from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score, confusion_matrix
from sklearn.metrics.pairwise import cosine_similarity, euclidean_distances
import matplotlib.pyplot as plt
```

#### Carga de bases de datos de Mercadolibre y la competencia

```
# Bases de datos de MELI y nuevos clientes
base_meli = pd.read_csv("base-meli.csv")
base_meli.rename(columns={'NOMBRE CUENTA':'nombre_meli','CUS_CUST_ID':'cust_id'},inplace=True)
base_leads = pd.read_csv("ic-base-competencia.csv").drop_duplicates().reset_index(drop=True)
base_leads.rename(columns={'SELLER_NAME':'cust_name'},inplace=True)

# Lista de todos los nombres
nombres = pd.concat([base_leads['cust_name'],base_meli['nombre_meli']]).reset_index(drop=True)

# df_nombres
nombres_leads = base_leads['cust_name'].drop_duplicates().reset_index(drop=True)
nombres_meli = base_meli[['cust_id', 'nombre_meli']].drop_duplicates().reset_index(drop=True)

print('Nombres de meli: '+str(len(nombres_meli))+'\nNombres leads: '+str(len(nombres_leads)))
```

## Preprocesamiento

```
# Función para preprocesar el nombre del cliente estandarizando el texto
def preprocesar(_lead):
    comunes=["propiedades","negocios","inmobiliario","inmobiliaria","inmobiliarias","inmobiliarios","servicios"]
    lead = _lead.lower().strip() # eliminamos minúsculas y espacios laterales
    lead = re.sub(' +', ' ', lead) # quitamos múltiples espacios
    lead = unicodedata.normalize('NFD', lead).encode('ascii', 'ignore').decode("utf-8") # elimina caract. esp.
    lead = re.sub(r'[^w\s]', '', lead) # elimina caracteres ., , ...
    lead = lead.replace('_', ' ') # reemplazo los _ por espacios
    lead = eval("lead"+''.join([".replace(\""+str(i)+"\", '')" for i in comunes]))
    lead = basename(lead) # elimina elementos comunes nombres de empresas tipo "ltd"
    return lead

# Función para crear una lista con todos los nombres estandarizados
def crear_lista_preprocesados(df):
    lista = []
    for i in range(df.count()):
        lista.append(preprocesar(df[i]))
    return lista

# Función para hacer el match exacto en el preprocesamiento recuperando los ids
def match_preprocesamiento(lista_leads,lista_meli,cust_id):
    cruce_ids = [None for i in lista_leads]
    for i in range(len(lista_leads)):
        for j in range(len(lista_meli)):
            if(lista_leads[i] == lista_meli[j]):
                cruce_ids[i] = cust_id[j]
    return cruce_ids

# Preprocesamos ambas bases de datos
leads_preprocesados = crear_lista_preprocesados(nombres_leads)
meli_preprocesados = crear_lista_preprocesados(nombres_meli['nombre_meli'])

# Recuperamos los ids del cruce exacto del preprocesamiento y creamos un dataframe con el cruce
ids = match_preprocesamiento(leads_preprocesados,meli_preprocesados,nombres_meli['cust_id'])
df_cruce = pd.DataFrame({'nombres_lead': nombres_leads, 'cust_id_cruce': ids}).merge(nombres_meli, left_on='cust_id_cruce', right_on='nombre_meli')
df_cruce = df_cruce.where(pd.notnull(df_cruce), None)
df_cruce = df_cruce.loc[df_cruce['cust_id_cruce'].notnull(),:]
df_cruce = df_cruce.assign(similarity=1)
df_cruce = df_cruce.drop(columns=['cust_id_cruce']).reset_index(drop=True)

# Creamos un dataframe con los restantes y preprocesamos los restantes
df_restantes = df_cruce.loc[df_cruce['cust_id_cruce'].isnull(),'nombres_lead'].reset_index(drop=True)
restantes_preprocesados = crear_lista_preprocesados(df_restantes)
cant_restantes = len(restantes_preprocesados)

# Creamos una lista con los nombres restantes de la competencia y los nombres de MELI
nombres = restantes_preprocesados+meli_preprocesados

# Revisamos la cantidad de nombres identificados
porcentaje_identificados = (1-(cant_restantes/len(nombres_leads)))
print('% de nombres identificados en el preprocesamiento: '+str(porcentaje_identificados))
```

## Embedding con BERT

```
modelo = SentenceTransformer('multi-qa-MiniLM-L6-cos-v1')
nombres_embeddings = modelo.encode(nombres)
```

## Determinación de similitud para coincidencias de nombres

```
def sigmoid(x):
    return 1 / (1 + np.exp(-x))

def estandarizar(x):
    promedio = np.mean(x)
    desviacion_estandar = np.std(x)
    return (x - promedio) / desviacion_estandar

def producto_punto(embeddings1,embeddings2):
    producto_punto = np.dot(embeddings1,np.transpose(embeddings2))
    prod_punto_estandarizado = estandarizar(producto_punto)
    producto_punto_norm = sigmoid(prod_punto_estandarizado)
    return producto_punto_norm

def sim_euclidiana(embeddings1,embeddings2):
    distancia = euclidean_distances(embeddings1,embeddings2)
    return 1/np.exp(distancia)
```

```
from fuzzywuzzy import fuzz

matriz_fuzzy = []
for i in restantes_preprocesados:
    matriz_fuzzy_temp = []
    for j in meli_preprocesados:
        partial_ratio = fuzz.partial_ratio(i,j)
        matriz_fuzzy_temp.append(partial_ratio/100)
    matriz_fuzzy.append(matriz_fuzzy_temp)

fuzzy_array=np.array(matriz_fuzzy)

similitud = []
similitud_coseno = []
restantes_embeddings=nombres_embeddings[:cant_restantes]
meli_embeddings=nombres_embeddings[cant_restantes:]

similitud_coseno = cosine_similarity(restantes_embeddings,meli_embeddings)
similitud_fuzzy_cos=np.array([(matriz_fuzzy[i][j]*similitud_coseno[i][j])/2 for j in range(len(fuzzy_array[i]))] for i in range(len(fuzzy_array)))
similitud = similitud_fuzzy_cos

similitud_maxima=np.max(similitud, axis=1)
index_cruce=np.argmax(similitud, axis=1)

df_restantes_sim=pd.DataFrame(df_restantes).reset_index(drop=True)
df_restantes_sim["cust_id"]=[nombres_meli.loc[i, "cust_id"] for i in index_cruce]
df_restantes_sim["nombre_meli"]=[nombres_meli.loc[i, "nombre_meli"] for i in index_cruce]
df_restantes_sim["similarity"]=similitud_maxima

df_cruce_sim=pd.concat([df_cruce, df_restantes_sim]).reset_index(drop=True)
```

## Resultados

```
# Definimos las predicciones
df_resultados = df_cruce_sim
df_resultados["predicción"] = np.where(df_resultados["similarity"] >= 0.722, 1, 0)
# Podemos revisar el % de clientes identificados
total_identificados = df_resultados['predicción'].mean()

# Podemos exportar los resultados uniéndolos a las bases iniciales
df_resultados = df_resultados.merge(base_leads, left_on='cust_name', right_on='cust_name', how='left')
df_resultados = df_resultados.merge(base_meli, left_on='cust_id', right_on='cust_id', how='left')
df_resultados.to_csv('base-competencia-resultados.csv')
```

## 9. Levantamiento de máquina virtual en Fury para desarrollo del modelo

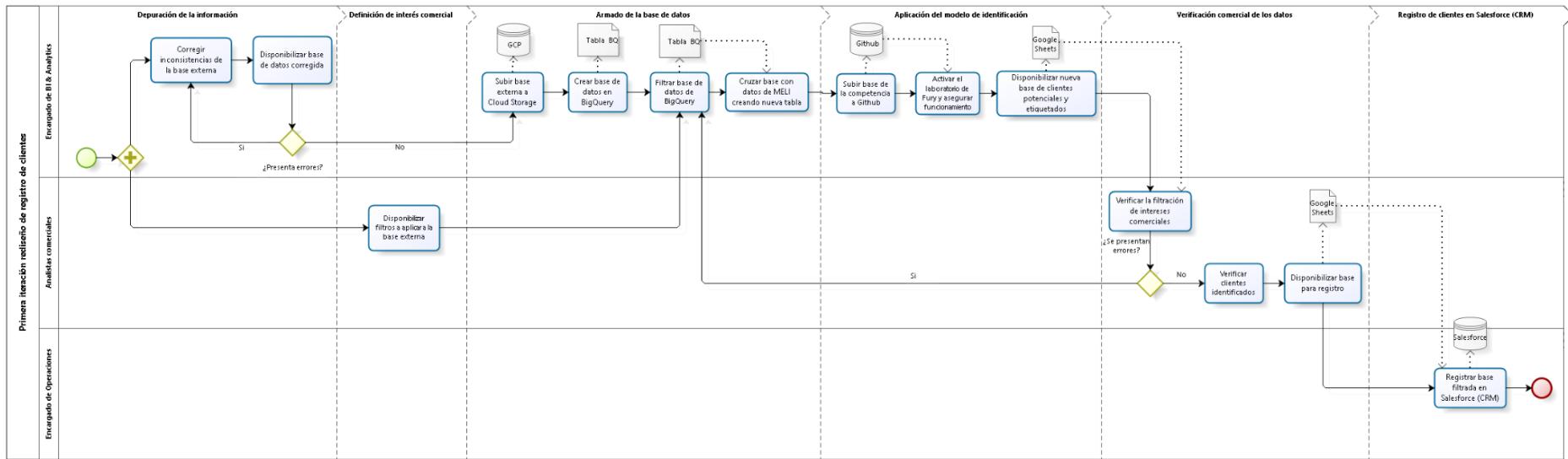
The screenshot shows the AWS Lambda console interface. On the left, there is a sidebar with various icons and a search bar. The main area is titled "Labs > Lab detail". A table displays a single lab entry:

ic-matching-names		Activate	Delete	Inactivate	View lab
Created by	ext_crissala	Creation date	2023-08-03 12:08:45	Activation time	2023-10-17 16:32:47
					Inactivation at 0:7:39

Below the table, there is a section titled "Flavor" with the following details:

Size	AWS instance type	vCPU	Memory	Storage
Medium	m5.2xlarge	8	32 GB	250 GB

## 10. Modelamiento BPMN del proceso rediseñado del registro de clientes



Powered by  
bizagi  
Modeler

## 11. Email de comunicado del relevamiento del nuevo proceso de registro de clientes

 Cristobal Salas [View profile](#) [Email](#) [Edit](#) [Share](#)

Un correo electrónico enviado el 20/03/2018 10:10:00 con el asunto "Relevamiento del nuevo proceso de registro de clientes" contiene el siguiente contenido:

**Relevamiento del nuevo proceso de registro de clientes**

1. Desarrollar los sistemas de recolección de datos para la implementación del nuevo sistema de registro de clientes.

2. Desarrollar una nueva estrategia de captación de datos para el nuevo sistema de registro de clientes.

3. Desarrollar un sistema de validación de datos para el nuevo sistema de registro de clientes.

**Resumen del nuevo proceso**

1. "Nuevo cliente" de 980 980 00, con el costo de desarrollo de 1.000 1000 de un nuevo cliente en la medida de su implementación.

2. "Nuevo cliente" de 980 980 00, porque es necesario un nuevo cliente.

**Relevamiento del nuevo proceso de registro de clientes**

1. Datos confidenciales, sensibles y delicados se devuelven con todos los sistemas para facilitar la protección de privacidad. Ningún dato se devolverá sin su consentimiento.

2. El seguimiento del desarrollo, al igual que el costo del desarrollo y costo del proceso para rendir reportes de información se pondrá en conocimiento en [Nuevo cliente](#).

**Nota adicional**

1. Proporcionar el costo estimado de 900 1000 para iniciar con 1 persona para revisar la base de datos existente, posteriormente se pondrá en conocimiento la respuesta.