

# Data Analytics Engineer Challenge

## Advanced Analytics & Machine Learning Commerce

*Bienvenido al Challenge de Data Analytics Engineer de Mercado Libre. Agradecemos tu interés por formar parte de nuestro equipo, y nos gustaría que te enfrentarás a los desafíos con los que trabajamos en nuestro día a día. Por medio de esta prueba nos gustaría conocer cómo te desenvuelves en las diferentes etapas de la solución de un problema de negocio, desde el entendimiento del problema, el planteamiento de preguntas, la obtención y exploración de la información, el planteamiento y creación de la solución analítica y la presentación de tus resultados. ¡Te deseamos mucha suerte!*

---

## 1. Instrucciones Generales

El objetivo de esta prueba es conocer la manera en que abordarías un caso de negocio, desde el entendimiento del problema hasta la creación de una solución analítica y su eventual escalabilidad. No es necesario que tengas conocimiento particular del negocio, pues lo que queremos evaluar es la forma en que te desenvuelves en cada fase del ciclo de Data Engineering y cómo incorporas herramientas de Inteligencia Artificial Generativa (GenAI) de manera estratégica y técnica.

Te presentaremos un caso de negocio base y dos caminos opcionales de extensión para que elijas uno de ellos, todos con foco en el uso de GenAI.

Durante el desarrollo deberás:

- Diseñar una solución analítica escalable que permita resolver el caso base.
- Extender la solución utilizando herramientas o enfoques de GenAI, según el camino elegido.
- Documentar todo tu trabajo en un repositorio público de GitHub (código, notebooks, resultados y presentación).
- Preparar una presentación de máximo **20 minutos**, donde compartas el proceso, análisis, soluciones, resultados y próximos pasos.

Este repositorio debe poder ser clonado y utilizado para replicar tu código de extracción de información, análisis exploratorios y modelos. Te recomendamos utilizar **Python en Jupyter Notebooks** para claridad y organización.

---

## 2. Caso de Negocio Base: Clusterización de Sellers

A continuación te presentamos un problema de negocio con el que nos enfrentamos día a día. Debes plantear una solución analítica de clusterización que te permita generar valor tanto a la unidad de negocio como a Mercado Libre en general.

El equipo comercial quiere realizar estrategias focalizadas para los sellers, pero en este momento no existe una clasificación que permita identificar a aquellos que tienen un buen perfil y son relevantes para el negocio.

**¿Cómo podrías ayudar al equipo comercial a identificar estos sellers y generar segmentaciones útiles?**

Deberás:

- Realizar un análisis exploratorio para entender los patrones de comportamiento.
  - Aplicar técnicas de **clusterización** para segmentar a los sellers.
  - Evaluar la calidad y utilidad de los clusters obtenidos.
  - Proponer formas concretas de cómo estos clusters pueden utilizarse en decisiones comerciales o estrategias personalizadas.
- 

## 3. Extensión con GenAI (elegir UNA de las siguientes)

Una vez resuelto el caso base, deberás elegir **una de las siguientes opciones** para extender tu solución incorporando capacidades de **Inteligencia Artificial Generativa**:

### **A** Opción A – *Clasificador Semántico con Embeddings Generados por LLM*

Entrena un modelo de clasificación que, utilizando **embeddings generados por un LLM** (por ejemplo, text-embedding-3-small, sentence-transformers, etc.), permita asignar un nuevo seller a uno de los clústeres predefinidos.

**Objetivo:** explorar cómo los LLM pueden utilizarse como herramienta de representación vectorial para construir clasificadores eficientes, útiles en tareas de onboarding o scoring comercial.

---

### **B** Opción B – *Recomendador Generativo de Estrategias Comerciales*

Diseña un componente generativo que, dado un seller, proponga una **estrategia comercial personalizada**, basándose en los datos del seller y su pertenencia a un clúster (por ejemplo: campañas de cuotas, incentivos, mejoras en catálogo, etc.).

Podés utilizar un agente basado en LLM (LangChain, pandas2prompt, etc.) para combinar datos estructurados con generación de lenguaje natural.

**Objetivo:** mostrar cómo las soluciones generativas pueden asistir a equipos comerciales en la toma de decisiones personalizadas a escala.

---

## 4. Extracción de Información

Se te compartirá un URL de Google Drive donde podrás descargar un CSV con información de ítems.

<https://drive.google.com/file/d/1Vh7ttgm9t86AFd6BEIRJummjSki3AI--/view?usp=sharing>

---

## 5. Entregables Esperados

Una vez planteado y resuelto tu caso de negocio debes realizar la socialización de tus resultados.

### 1. Repositorio público en GitHub con:

- Código de extracción y análisis.
- Notebooks bien documentados y ejecutables.
- Modelos implementados (clustering + extensión GenAI).
- Archivo README.md con instrucciones de ejecución y contexto del trabajo.

### 2. Presentación (20 minutos) que incluya:

- **Problema abordado y motivación:** ¿Qué problema se quiso resolver con el desarrollo de la solución?
- **Hipótesis, fuentes de datos y análisis exploratorio:** ¿Qué información buscaste y utilizaste para el desarrollo de la solución? Incluye las hipótesis que te llevaron a descargar dicha información.
- **Explicación técnica de tu solución:** Danos un contexto de la solución que escogiste, y cómo la abordaste. ¿Qué metodologías aplicaste? ¿Qué métricas de evaluación utilizaste?
- **Resultados obtenidos y métricas relevantes:** Muestra los insights más relevantes que hayas obtenido de tu análisis exploratorio. ¿Cómo te guiaron a la hora de plantear la solución?
- **Implementación del módulo de GenAI seleccionado:** ¿Cuál opción escogiste, por qué? ¿Cómo se comporta? ¿Cómo soluciona o ayuda a solucionar el problema de negocio?
- **Conclusiones y próximos pasos:** Concluye y plantea próximos pasos que harías después de entregar esta solución. Asegúrate de incluir una reflexión sobre cómo los resultados obtenidos pueden ser interpretados y aprovechados por perfiles de negocio.

Es importante que estos análisis estén respaldados por cifras, y que estas puedan ser consultadas y replicadas en los notebooks del repositorio de Github.

***Por favor envía tu repositorio de Github y el material (PowerPoint, prezi, infografía, canva, etc..) con el que apoyarás tu presentación/sustentación.***

---

## 6. ¿Dudas?

Ante cualquier duda técnica o funcional, no lo pienses dos veces y contáctanos por medio de correo electrónico o los canales que te hayamos proporcionado cuando te enviamos esta prueba.