

## Prueba Técnica: Científico de Datos

### Instrucciones

1. Desarrollar una solución para el requerimiento descrito en la sección “Problema”
2. Entregar el código de la solución en forma de un repositorio de Github; incluir instrucciones de instalación
3. Proveer una solución que pueda ejecutarse *offline*
4. Utilizar el lenguaje de programación Python  $\geq 3.7$
5. Agregar los comentarios que se consideren pertinentes
6. Agendar una llamada para sustentar la solución

### Problema

Contexto: una empresa de transporte sufre de una alta rotación de conductores. Para intentar identificar el problema, se desarrolló una App móvil por medio de la cual se registran las quejas de los conductores. Dichas quejas están etiquetadas según el área: finanzas, recursos humanos y operaciones (experiencia al realizar los viajes, evaluación del coordinador responsable del conductor).

Requerimientos:

1. Crear una base de datos sintética que emule el registro de los eventos capturados de la App; deberá de incluir por lo menos la etiqueta y el mensaje registrado por el usuario; proveer la base de datos en formato .csv; tamaño mínimo de eventos (registros): 3000
2. Crear un *script* (jupyter notebook) donde se plasme todo el proceso
3. Crear un modelo que permita identificar candidatos con una alta probabilidad de retirarse de la empresa

### Sugerencias

1. Agregar las variables adicionales que se consideren de utilidad; intentar recrear un escenario real (Tip: utilizar variables que hagan sentido que puedan ser capturadas. Es posible asumir que se tiene información socioeconómica y relacionada a antecedentes laborales del conductor que fue recolectada durante la contratación)
2. Opcional: utilizar AI generativa para crear los mensajes de texto proporcionados por el conductor/usuario de la App (Tip: utilizar LM Studio para no incurrir en gastos). Llegado el caso, utilizar NLP
3. Con relación a la información de viajes, también se puede asumir el registro de los viajes asignados al conductor, que incluye la duración de los viajes, distancia recorrida (nota: entre más distancia recorrida, mayor ganancia), características de las regiones geográficas (e.g. zonas de riesgo)
4. Sustentar los mecanismos utilizados para generar los datos sintéticos; se espera un análisis multivariado, por lo tanto, la correlación entre las variables deberá de hacer sentido
5. Visualizar los resultados (*data storytelling*)
6. Se espera el uso de GenAI (e.g. copilot) para el desarrollo del *script*, si es el caso, se deberá sustentar el cómo y el porqué

### Criterios de Evaluación

1. El autor siguió las instrucciones
2. Fue posible ejecutar la solución sin intervención adicional por parte del autor
3. La solución resuelve el problema de forma satisfactoria
4. La documentación es adecuada y suficiente
5. El autor demuestra dominio de la solución y sustenta correctamente las decisiones tomadas

Este documento es confidencial y/o puede contener información privilegiada. Si usted no es su destinatario o no es una persona autorizada por éste para acceder a sus documentos, no debe abrir, usar, publicar, distribuir, copiar, utilizar, revelar o tomar acción alguna basada en este documento, su contenido, archivo adjunto o cualquier información aquí incluida de cualquier modo. Si usted ha recibido este documento por error, favor de notificar al remitente de inmediato y eliminar este documento y sus anexos completamente. Cualquier opinión contenida en este documento es responsabilidad única y exclusiva del autor del mismo, a menos que lo contrario se indique claramente y por escrito. La Sociedad y subsidiarias no aceptan responsabilidad por cualquier error, omisión, corrupción o virus en el contenido de este documento o cualquier archivo adjunto.

This document is confidential and/or may contain privileged information. If you are not the recipient or a person authorized by him/her to access his/her documents, you must not open, use, publish, distribute, copy, use, disclose or take any action based on this document, its content, attachment or any information included here in any way. If you have received this document by mistake, please notify the sender immediately and delete this document and its attachments completely. Any opinion contained in this document is the sole and exclusive responsibility of the author, unless otherwise stated clearly and in writing. The Company and subsidiaries do not accept responsibility for any error, omission, corruption or virus in the content of this document or any attachment.