



## Challenge Data Science 2022

Agosto, 2022

¡Hola! Agradecemos mucho tu interés por formar parte del equipo de Genomawork! Como siguiente etapa de tu postulación te invitamos a realizar un pequeño ejercicio de ciencia de datos.

### 1 Objetivo general

El objetivo de este ejercicio es observar cómo utilizas diversas técnicas de ciencia de datos para enfrentarte a un problema real con objetivos y preguntas de investigación a nivel de negocio.

### 2 Objetivos específicos

En particular, mediante este ejercicio queremos evaluar tu:

- Manejo de datos (leer archivos .csv, filtrarlos, limpiarlos, etc.)
- Enfoque científico (planteamiento de hipótesis y validación o refutación con datos)
- Uso de estadística en problemas reales
- Uso de herramientas de visualización de datos
- Técnicas de elección y entrenamiento de modelos simples de machine learning
- Capacidad de sintetizar los principales resultados en conclusiones enfocadas a un problema de negocio

### 3 Instrucciones

Imagina que nuestro cliente Stark Industries se ha planteado el objetivo de mejorar sus contrataciones. Dentro de sus principales problemas, Stark Industries identifica la rotación de personal, es decir, los colaboradores que abandonan la compañía. Stark Industries ha recolectado algunos datos para poder comprender mejor el fenómeno de la rotación de personal y se te ha encargado a ti extraer valor de estos datos. Como mínimo, se plantean las siguientes preguntas:

1. ¿Qué variables presentan alguna relación con la rotación? ¿Qué tipo de relación presentan? ¿Es significativa?
2. ¿Podemos crear un modelo que permita identificar a un empleado con alto riesgo de rotación? ¿Cuáles son los pros y los contras de un modelo de este tipo?
3. ¿Cómo podemos medir el desempeño de un modelo de estas características?
4. ¿Qué otros tipos de datos pueden ser valiosos para alimentar a estos modelos?

5. ¿Podríamos generar un modelo que identifique a las personas con alto riesgo de rotación antes de contratarlas? ¿En qué se diferenciaría este modelo del anterior?
6. ¿Qué recomendarías a Stark Industries para atacar este problema?

**Entrega:** Cuando hayas completado el challenge, envíalo al correo [pablo@genomawork.com](mailto:pablo@genomawork.com) con copia a [bryanc@genomawork.com](mailto:bryanc@genomawork.com).

### 3.1 Requisitos

- Debes entregar un archivo comprimido con todos los archivos necesarios
- Tu trabajo debe estar desarrollado en Python 3.6 o superior
- Si hay que ejecutar tu código, incluye un readme
- Puedes ocupar cualquier librería

### 3.2 Recomendaciones

- Utiliza Jupyter notebook para que puedas alternar entre código y texto.
- Entrega el Jupyter notebook con las celdas ya ejecutadas.
- Organiza tu trabajo en secciones según distintos objetivos. Plantea claramente cuales son las preguntas o hipótesis y concluye a partir de los datos.
- Dirige las conclusiones finales a un interlocutor no-técnico. Incluye pasos futuros o recomendaciones, además de los principales hallazgos. Enfoca los resultados en conclusiones valiosas para el negocio detrás de los datos.

## 4 Datos

Los datos que utilizarás son provenientes de datos relacionados con rotación de empleados (employee turnover). El dataset original es público y está disponible en Kaggle en el siguiente link: <https://www.kaggle.com/davinwijaya/employee-turnover>

Para facilitar un poco el manejo de la base, puedes usar la siguiente traducción de columnas:

independ	->	agreeableness
selfcontrol	->	conscientiousness
anxiety	->	neuroticism
novator	->	openness

## 5 Próximos pasos

En caso de que tu solución cumpla con los objetivos descritos, coordinaremos una entrevista, cuyo objetivo será presentar los principales hallazgos de este ejercicio enfocado a una audiencia no-técnica. Nos interesa ver cómo puedes comunicar ideas de manera simple, pero precisa y respaldada en tu trabajo.

