# Prueba técnica. Ejecutivo Data Science. Coca-Cola FEMSA.

Bienvenido a la prueba técnica para el puesto de Ejecutivo Data Science en Coca-Cola FEMSA.

Esta prueba se ha diseñado para evaluar una variedad de habilidades y competencias relacionadas con el análisis de datos, la gestión de proyectos y el liderazgo de equipos. Además de las habilidades técnicas, como la programación en Python y la aplicación de técnicas avanzadas de Machine Learning, también estamos interesados en tu capacidad para comunicar resultados, administrar proyectos y equipos, y mantenerte al día con las últimas tendencias y avances en el campo de la analítica avanzada.

Para ayudarte a prepararte para esta prueba, hemos creado una guía de estudio que cubre los principales temas y habilidades que se evaluarán. Esta guía incluye temas como los fundamentos del análisis de datos, la programación y el manejo de datos en Python, las técnicas de Machine Learning, el análisis de series temporales, la gestión de proyectos de análisis de datos, la comunicación de resultados, los aspectos éticos y legales del análisis de datos, y las tendencias en análisis avanzado y tecnologías relacionadas.

Te animamos a que utilices esta guía para revisar y fortalecer tus conocimientos en estas áreas, y también para identificar cualquier área en la que podrían necesitar un estudio adicional. Recuerda, esta es una oportunidad para demostrar no solo lo que ya sabes, sino también tu capacidad para aprender, adaptarte y aplicar tus habilidades en nuevas y desafiantes situaciones.

Esta guía ya te fue entregada con anterioridad.

## Guía de estudio:

1. Fundamentos de Análisis de Datos:
    - Estadística básica y descriptiva: medidas de centralidad, dispersión y correlación.
    - Interpretación de gráficos y tablas.
    - Manejo de datos faltantes y atípicos.
    - Ingeniería de características y selección de características.
    - Control de calidad de datos.
2. Programación y Manejo de Datos:
    - Programación en Python: sintaxis básica, operaciones con listas, diccionarios,
    - ciclos, funciones, etc.
    - Uso de bibliotecas de análisis de datos como Pandas y Numpy.
    - Carga y manipulación de datos con Pandas: filtrado, agrupación, pivotación, etc.
    - Visualización de datos con Matplotlib, Seaborn u otras bibliotecas.
3. Machine Learning:
    - Conceptos básicos de Machine Learning: aprendizaje supervisado y no supervisado, overfitting, bias-variance trade-off, etc.- Modelos de regresión y clasificación: regresión lineal, árboles de decisión, random forest, SVM, etc.

- Técnicas de validación cruzada y ajuste de hiperparámetros.
- Evaluación de modelos: precisión, recall, F1 score, AUC-ROC, error cuadrático
- medio, etc.
- Uso de bibliotecas de Machine Learning como scikit-learn y spark MLLib.

4. Análisis de Series Temporales:
   - Componentes de una serie temporal: tendencia, estacionalidad, ruido.
   - Modelos de series temporales: AR, MA, ARIMA, etc.
   - Descomposición de series temporales.
5. Gestión de Proyectos de Análisis de Datos:
   - Definición de objetivos y métricas de éxito.
   - Diseño y planificación de proyectos de análisis de datos.
   - Gestión de recursos y plazos.
   - Monitorización y actualización de modelos.
6. Comunicación de Resultados:
   - Creación de informes y presentaciones claras y convincentes.
   - Traducción de hallazgos técnicos a recomendaciones de negocio.
   - Uso de visualizaciones de datos para apoyar las conclusiones.
7. Aspectos Éticos y Legales del Análisis de Datos:
   - Protección y privacidad de los datos.
   - Cumplimiento de regulaciones y normativas relevantes en materia de privacidad y protección de datos.
   - Uso ético de los datos y los modelos predictivos.
8. Tendencias en Análisis Avanzado y Tecnologías Relacionadas:
   - Técnicas avanzadas de Machine Learning: redes neuronales, algoritmos de ensemble, etc.
   - Herramientas y plataformas de Big Data: Hadoop, Spark, etc.
   - Inteligencia artificial y aprendizaje profundo.
   - Innovaciones recientes y tendencias futuras en el campo de la analítica avanzada.

Es importante recordar que esta es una guía general y puede que no cubra todos los aspectos que se evaluarán en la prueba. Es recomendable revisar la descripción del puesto y la prueba en detalle para identificar cualquier tema adicional que pueda ser relevante.

## Caso de Estudio - Predicción de demanda de ventas de Productos

Para este caso de estudio, utilizarás el conjunto de datos de las ventas minoristas de Rossmann, disponible en Kaggle. El dataset incluye información sobre las ventas diarias de las tiendas de Rossmann en diferentes regiones, así como datos sobre promociones, competencia y otros factores.

Link al dataset: https://www.kaggle.com/c/rossmann-store-sales/data

El objetivo del caso de estudio es construir un modelo de machine learning que pueda predecir con precisión las ventas futuras en las tiendas.

Puedes usar cualquier librería de Python o técnica de modelado para dar respuesta a las siguientes preguntas.

## Parte 1: Análisis exploratorio de los datos (EDA)

1. Carga los datos en Python y realiza una inspección inicial. ¿Qué tipos de datos hay en el conjunto de datos? ¿Hay algún valor perdido o atípico?

2. Realiza un análisis estadístico inicial de los datos. ¿Cuál es el rango de las ventas? ¿Cuántas tiendas y productos diferentes hay en los datos?

3. ¿Cómo se distribuyen las ventas por tiendas? ¿Existen tiendas con comportamientos atípicos? ¿Qué podrían implicar estas anomalías?

4. Visualiza las ventas a lo largo del tiempo para un par de tiendas. ¿Se observa alguna tendencia o patrón recurrente?

5. Con base en el análisis exploratorio, identifica las principales características que podrían afectar las ventas. Justifica tus elecciones.

## **P**arte 2: Preprocesamiento de los datos y Ingeniería de Características

6. ¿Cómo tratarías los valores perdidos y atípicos en los datos? Explica tu enfoque y las razones de tu elección.

7. Realiza la ingeniería de características necesaria para mejorar el rendimiento del modelo. ¿Qué nuevas características podrías crear y por qué crees que serán útiles?

8. ¿Cómo garantizarías la calidad de los datos durante este proceso?

## Parte 3: Modelado de Machine Learning

9. Selecciona un modelo de aprendizaje automático adecuado para esta tarea de predicción. ¿Por qué elegiste este modelo? ¿Qué ventajas ofrece para este tipo de problemas?

10. Divide los datos en un conjunto de entrenamiento y un conjunto de pruebas. ¿Cómo te asegurarías de que la división refleja cualquier estructura temporal en los datos?

11. Entrena el modelo en el conjunto de entrenamiento. Ajusta los hiperparámetros del modelo para mejorar su rendimiento. ¿Cómo optimizarías este proceso?

## Parte 4: Evaluación del Modelo y Optimización

12. Evalúa el rendimiento del modelo en el conjunto de prueba. ¿Cómo de preciso es el modelo? ¿Cómo varía el rendimiento a lo largo del tiempo y entre diferentes tiendas?

13. ¿Qué métricas utilizarías para evaluar el modelo? ¿Por qué elegiste estas métricas?

14. ¿Cómo mejorarías el modelo si tuvieras más tiempo o recursos? ¿Qué otros modelos o técnicas podrías probar?

## Parte 5: Interpretación y Comunicación de Resultados

15. ¿Qué características son las más importantes para la predicción de las ventas? ¿Cómo interpretas estos resultados?

16. Basándote en los resultados del análisis, ¿qué acciones recomendarías a la dirección de la empresa para aumentar las ventas?

17. ¿Cómo te asegurarías de que los resultados del análisis sean utilizados eficazmente por los demás departamentos de la empresa?

18. Supongamos que tienes que presentar los resultados del análisis a un público no técnico. ¿Cómo comunicarías tus hallazgos de una manera clara y convincente?

19. ¿Qué estrategias implementarías para mantener el modelo actualizado y garantizar que continúe proporcionando predicciones precisas a medida que llegan nuevos datos?