Challenge

Data Science I: Fundamentos para la Ciencia de Datos

Alumno: Javier López Malone
Prof.: Estefania Karina Susanj
Tutor: Rubén Baquel

Coderhouse - febrero de 2024

Elección de datasets potenciales

Consigna

Identificar 3 datasets que cumplan con las siguientes condiciones:
a) al menos 2000 filas
b) al menos 15 columnas.

Cargar los archivos correspondientes por medio de la librería Pandas.

Describir las variables potencialmente interesantes en cada archivo teniendo en cuenta el contexto del caso.

Elección de datasets

Dataset 1: LinkedIn Job Postings 2023

Fuente: LinkedIn Job Postings - 2023 | Kaggle

Cada día, miles de empresas y particulares recurren a LinkedIn en busca de talento. Este conjunto de datos contiene un registro casi completo de más de 33.000 ofertas de trabajo enumeradas en el transcurso de 2 días, con meses de diferencia. Cada publicación individual contiene 27 atributos, incluídos el título, la descripción del puesto, el salario, la ubicación, la URL de la aplicación y los tipos de trabajo (remoto, por contrato, etc.). Además cuenta con archivos separados que contienen los beneficios, las habilidades y las industrias asociadas con cada uno. La mayoría de los trabajos también están vinculados a una empresa, y todos se enumeran en otro archivo csv que contiene atributos como la descripción de la empresa, la ubicación de la sede, la cantidad de empleados y el número de seguidores.
Con tantos datos, el potencial de exploración de este conjunto de datos es enorme e incluye la exploración de los títulos, empresas y ubicaciones mejor remunerados; predecir salarios/beneficios a través de PNL; y examinar cómo las industrias y las empresas varían a través de sus ofertas y beneficios de pasantías. Las actualizaciones futuras permitirán una mayor exploración de las tendencias temporales, incluido el crecimiento de las empresas, la prevalencia de trabajos remotos y la demanda de puestos de trabajo individuales a lo largo del tiempo.

Elección de datasets

Dataset 2: Spotify Tracks Dataset

Fuente: Spotify Tracks Dataset | Kaggle

Es un conjunto de datos de pistas de Spotify en una variedad de 125 géneros diferentes.

Elección de datasets

Dataset 3: Real/fake job posting prediction

Fuente: Real / Fake Job Posting Prediction | Kaggle

Este conjunto de datos contiene 18.000 descripciones de puestos de trabajo, de las cuales unas 800 son falsas. Los datos constan tanto de información textual como de metainformación sobre los puestos de trabajo. El conjunto de datos se puede utilizar para crear modelos de clasificación que puedan conocer las descripciones de puestos que son fraudulentas.