"Año de la recuperación y consolidación de la economía peruana".

Escuela Profesional de Ciencia de la Computación
Topicos en Ciencia de Datos

Pipeline de Ciencia de Datos

Estudiantes:
Jharold Alonso Mayorga Villena

Docente :
ANA MARIA CUADROS VALDIVIA

1. Pipeline de Ciencia de Datos

Introducción

El análisis de las ofertas laborales presenta un desafío en la identificación de patrones que indiquen si una oferta es fraudulenta o legítima. Las variables presentes en el dataset, como el tipo de empleo, la longitud de las descripciones, los beneficios ofrecidos, la ubicación geográfica, y la presencia de preguntas para el candidato, pueden estar relacionadas con la probabilidad de fraude.

A continuación, exploraremos varias hipótesis clave para comprender mejor cómo ciertas características pueden influir en la probabilidad de fraude en las ofertas de trabajo.

—

Hipótesis 1

¿El tipo de empleo (tiempo completo, medio tiempo, contrato) tiene un impacto significativo en la probabilidad de que una oferta sea fraudulenta?

Contexto: Se cree que los trabajos temporales o freelance tienen una mayor probabilidad de ser fraudulentos debido a su naturaleza menos formal. Por el contrario, los trabajos full-time suelen estar asociados con una mayor estabilidad y por ende menor probabilidad de fraude.

Gráfico 1: Porcentaje de Fraude según Tipo de Empleo

A continuación, mostramos un gráfico de barras que compara la distribución de fraude entre los diferentes tipos de empleo (full-time, contract, part-time, etc.), destacando los porcentajes de fraude para cada tipo.

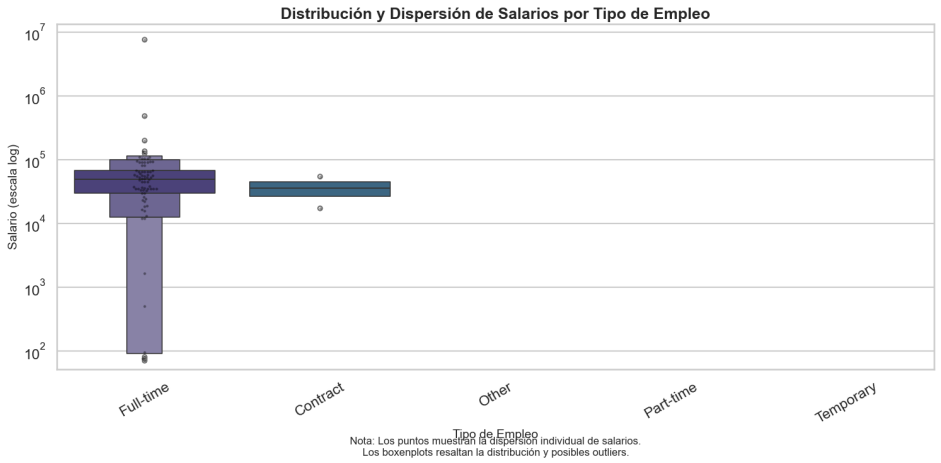

Figura 1: Porcentaje de Fraude según Tipo de Empleo

Explicación del gráfico: Este gráfico nos permite visualizar cómo se distribuye el fraude dentro de cada tipo de empleo. El porcentaje de fraude está claramente marcado en cada barra, lo que nos permite comparar la prevalencia de fraude según el tipo de trabajo. Los colores indican si la oferta es fraudulenta o legítima, permitiendo ver la relación.

—

Hipótesis 2

¿Las ofertas fraudulentas tienden a tener descripciones y requisitos más vagos o menos detallados en comparación con las ofertas legítimas?

Contexto: Las ofertas fraudulentas tienden a ser más vagas y menos detalladas. Los estafadores a menudo utilizan descripciones generales y ambiguas para atraer a víctimas sin comprometerse demasiado. Las ofertas legítimas, por otro lado, suelen ser más claras y específicas.

Gráfico 2: Proporción de Fraude según la Presencia de Preguntas

A continuación, mostramos un gráfico de barras apiladas que ilustra cómo la presencia de preguntas en una oferta está asociada con el fraude.

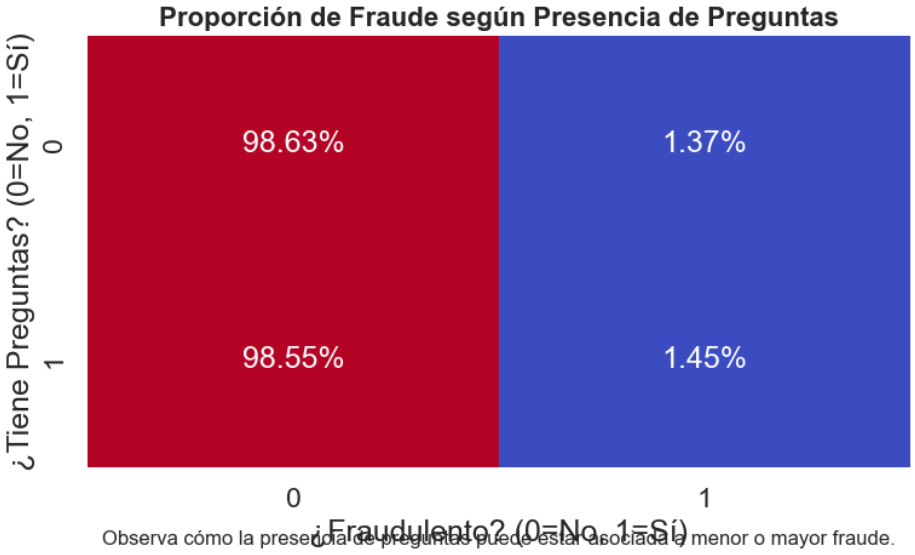

Figura 2: Proporción de Fraude según Presencia de Preguntas

Explicación del gráfico: Este gráfico muestra la relación entre las ofertas con y sin preguntas y el porcentaje de fraude. Las barras rojas y azules representan la presencia de fraude para cada grupo. El gráfico nos indica que las ofertas que incluyen preguntas tienen una proporción más baja de fraude, sugiriendo que la presencia de preguntas puede ser un indicador de una oferta más legítima.

—

Hipótesis 3

¿Existen patrones de fraude asociados con las ubicaciones geográficas de las ofertas de trabajo?

Contexto: Las ofertas de trabajo con ubicaciones poco claras o ambiguas, como las que ofrecen trabajos remotos o no tienen una dirección física, podrían ser más propensas al fraude. Las ofertas ubicadas en regiones verificables o en ciudades conocidas tienen una mayor probabilidad de ser legítimas.

Gráfico 3: Distribución y Dispersion de Salarios por Tipo de Empleo

A continuación, se muestra un gráfico boxplot para observar cómo se distribuyen los salarios por tipo de empleo. Esto nos ayudará a entender si existen diferencias salariales según la ubicación geográfica y el tipo de trabajo.

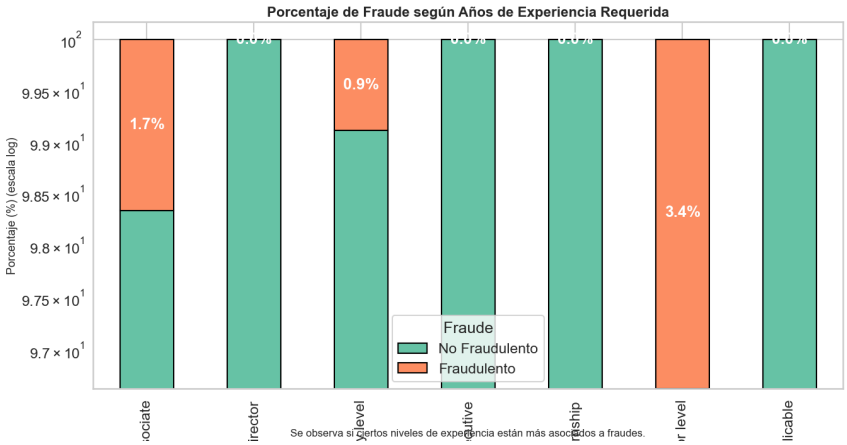

Figura 3: Distribución y Dispersion de Salarios por Tipo de Empleo

Explicación del gráfico: Este gráfico de boxplot muestra cómo se distribuyen los salarios según el tipo de empleo. Los puntos fuera de los bigotes del gráfico representan valores atípicos, lo que podría indicar ofertas extremadamente altas o bajas que podrían ser fraudulentas. El análisis de estos valores nos ayuda a comprender si los salarios están relacionados con la ubicación o el tipo de empleo.

2. Preguntas para el análisis del dataset

- **¿Qué problema identifican en el dataset?**
  - El dataset presenta una desproporción considerable entre las ofertas laborales fraudulentas (1) y las no fraudulentas (0). Este desbalance de clases es común en muchos datasets y plantea un desafío significativo al momento de entrenar modelos de clasificación. Si no se maneja adecuadamente, el modelo podría estar sesgado hacia la clase mayoritaria, es decir, las ofertas no fraudulentas, lo que reduciría su capacidad para detectar correctamente las ofertas fraudulentas.

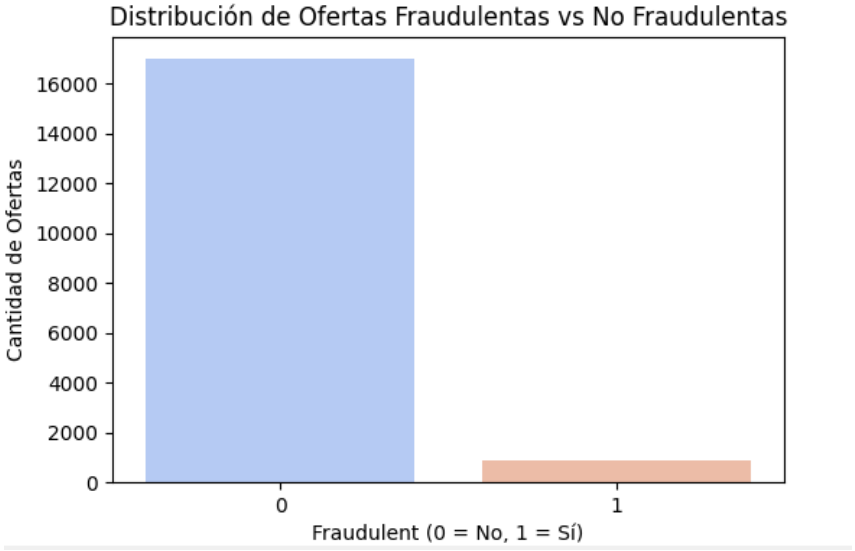

Figura 4: Imagen explicativa del problema de desproporción de clases.

- **¿Qué dificultades presenta la falta de datos?**
  - En varias columnas clave como salary_range, department, required_experience, entre otras, se presentan valores faltantes. Estos vacíos pueden interferir con la creación de modelos predictivos precisos, ya que la información faltante podría contener patrones importantes para detectar fraudes.
  - La falta de datos puede distorsionar los resultados del análisis. Para tratar estos problemas, se pueden aplicar métodos como la imputación de valores (por ejemplo, media, mediana o el uso de algoritmos de imputación avanzados) o bien, eliminar las filas que contengan demasiados valores faltantes si no afectan el análisis general.

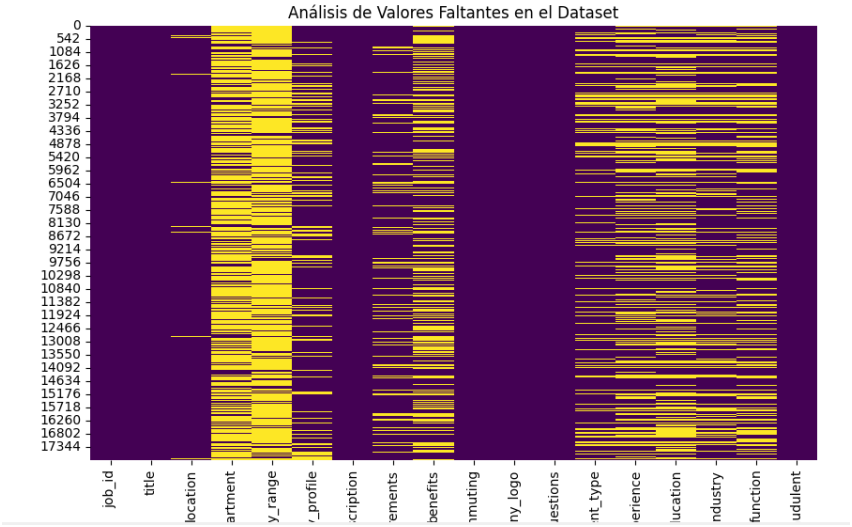

Figura 5: Visualización de los datos faltantes en el dataset.

- **¿Cómo se manejan las inconsistencias en las ubicaciones?**
  - La columna `location` contiene formatos no estandarizados, por ejemplo, algunas ofertas tienen la ubicación como "New York, NY", mientras que otras solo indican "NYC". Estas inconsistencias pueden dificultar el análisis geográfico y la agrupación de los datos.
  - Para resolver este problema, se pueden utilizar técnicas de expresiones regulares para estandarizar los formatos de ubicación, o bien, utilizar una librería como Geopy para convertir las ubicaciones a coordenadas estándar (latitud, longitud) y poder agruparlas de manera eficiente.

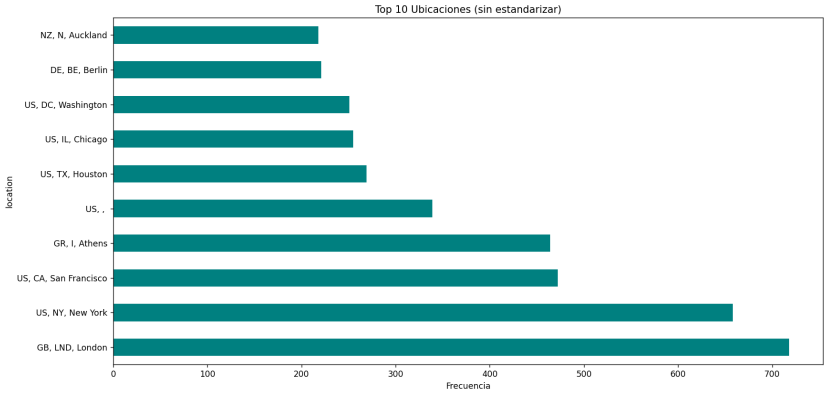

Figura 6: Inconsistencias en los formatos de ubicación.

- **¿Qué descubrimientos se hicieron al analizar los datos?**
  - Al analizar los datos, descubrimos que las ofertas de empleo categorizadas como Ïnternshipƨ "Freelance"tienen una probabilidad significativamente más alta de ser fraudulentas. Esto sugiere que los trabajos temporales o con un compromiso a corto plazo son más propensos a ser fraudulentos debido a su naturaleza menos formal.
  - Además, las ofertas que requieren menos experiencia o no especifican claramente los requisitos también tienden a ser fraudulentas. Esto refleja que los estafadores buscan candidatos con menos formación y experiencia para que no investiguen a fondo la oferta.

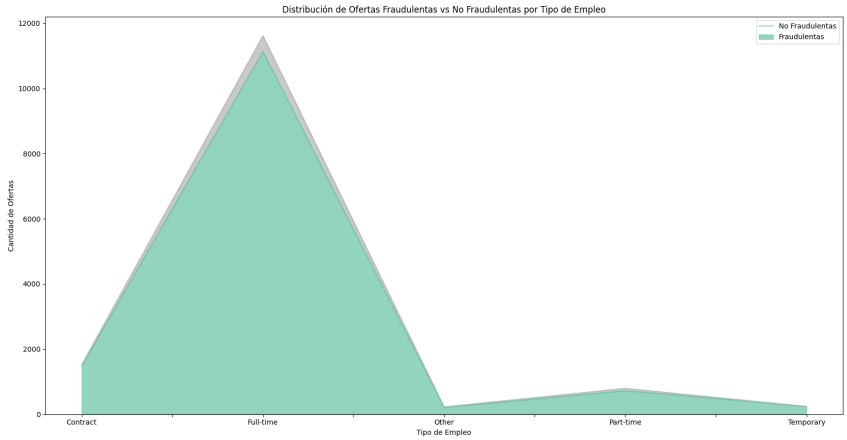

Figura 7: Distribución de la probabilidad de fraude según el tipo de empleo.

- **¿Cómo influyen los rangos salariales en la probabilidad de fraude?**
  - Las ofertas fraudulentas suelen ofrecer salarios extremadamente altos o bajos sin detalles claros. Los estafadores suelen atraer a las víctimas ofreciendo grandes sumas de dinero, pero sin especificar los beneficios o las condiciones del trabajo.
  - Al analizar la distribución salarial, encontramos que las ofertas fraudulentas presentan una mayor dispersión en los salarios, lo que refuerza la idea de que las promesas de salarios altos son una táctica común de los fraudes.

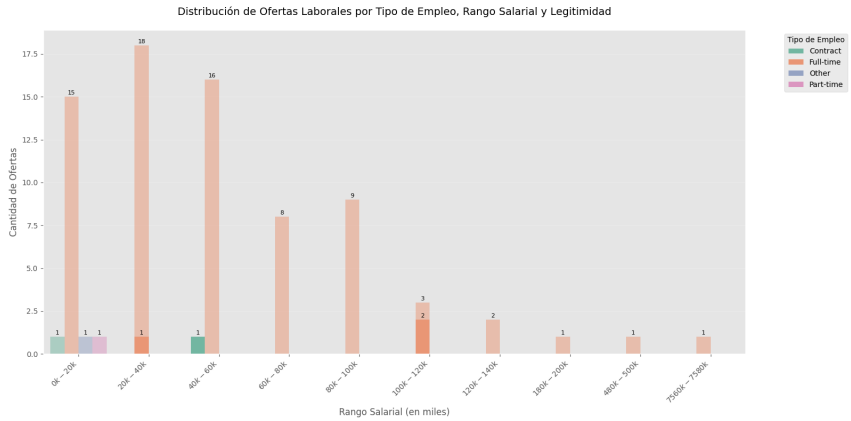

Figura 8: Distribución de salarios en ofertas fraudulentas y legítimas.

- **¿Por qué los anuncios fraudulentos tienden a tener descripciones más cortas?**

  - Los anuncios fraudulentos generalmente tienen descripciones más cortas en promedio, con una longitud de aproximadamente 500 caracteres, mientras que los anuncios legítimos suelen tener descripciones más largas, cercanas a 1000 caracteres. Esta brecha en la longitud puede ser una señal de que los estafadores no proporcionan suficientes detalles sobre el trabajo para evitar el escrutinio.

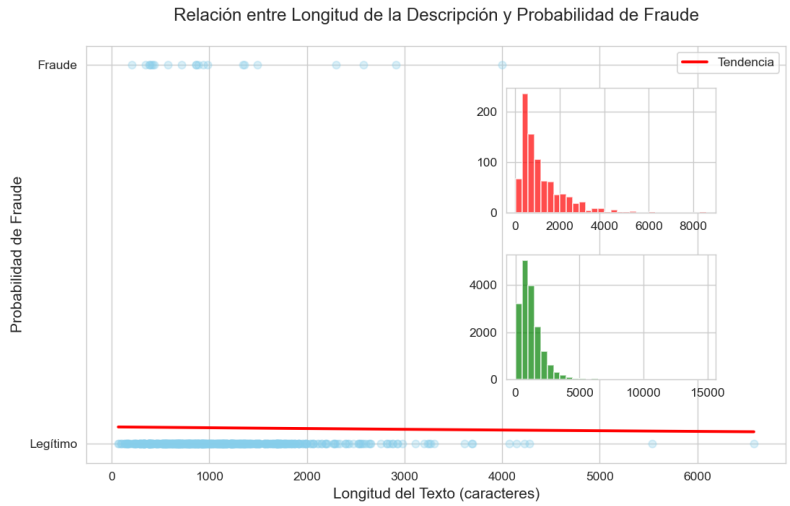

Figura 9: Comparación de longitud de texto entre ofertas fraudulentas y legítimas.

- **¿Cómo afectan las palabras clave a la detección de fraudes?**
  - Las ofertas fraudulentas suelen utilizar palabras clave genéricas, como "fácil", rápido dinero", "sin experiencia requerida", que incrementan la probabilidad de fraude. Estas palabras clave atraen a personas que buscan oportunidades sin investigar mucho sobre la oferta.
  - Un análisis de wordcloud de las palabras más comunes en ofertas fraudulentas muestra una prevalencia de términos vagos y atractivos para captar la atención de los candidatos.

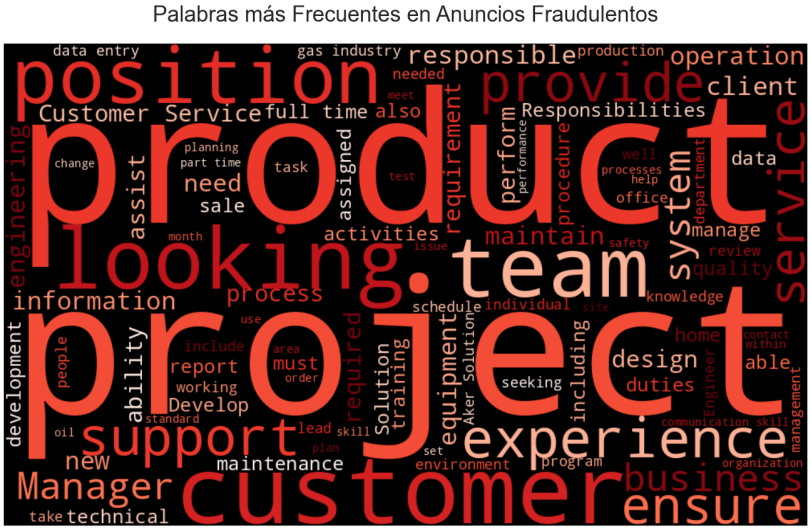

Figura 10: Wordcloud con palabras clave de ofertas fraudulentas.

- **¿Cómo influye el tipo de empleo y los beneficios ofrecidos en la percepción de fraude?**

  - Las ofertas sin beneficios suelen ser percibidas como más sospechosas. Los estafadores a menudo no mencionan beneficios o no especifican los detalles para evitar ser descubiertos. Además, las ofertas de empleo temporal o freelance tienden a ser vistas como más fraudulentas debido a la falta de estabilidad en estos trabajos.

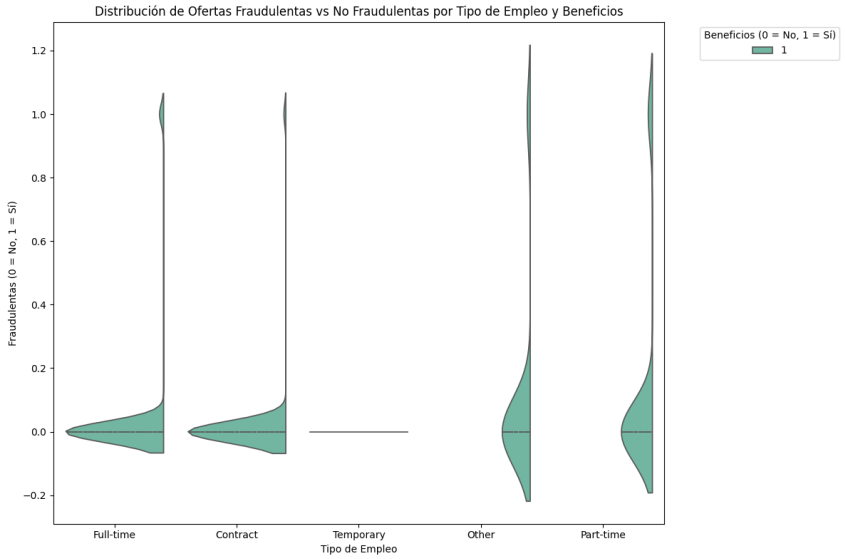

Figura 11: Distribución de ofertas fraudulentas y legítimas según tipo de empleo y beneficios.