



# ANALISIS

PROYECTOS DE  
INVESTIGACIÓN

JHAROLD ALONSO MAYORGA VILLENA

2025

.....

# CONTENIDO

01

Detección de fraudes en  
anuncios de empleo usando  
Machine Learning

.....

# INTRODUCCIÓN (PROBLEMA)

Contexto:

---

Contexto:

- El 86% de los fraudes en empleo buscan el robo de identidad (Vidros et al., 2016).
- El coste estimado de estos fraudes es de \$500M anuales solo en EE.UU. (FTC, 2021).

Problema:

- Los modelos actuales de detección de fraudes son binarios (fraude/no fraude) y carecen de interpretabilidad (Vidros et al., 2017).
- Existe una falta de granularidad para distinguir entre los distintos tipos de fraude (ej.: estafas piramidales vs. suplantación corporativa).

# OBJETIVOS DEL ESTUDIO

## Clasificación multiclasificación:

### Tipo 1: Robo de identidad

Donde los anuncios fraudulentos solicitan información personal como nombre completo, dirección, número de teléfono, etc.

### Tipo 2: Suplantación corporativa

Anuncios que se hacen pasar por empresas legítimas, utilizando su nombre y marca para ganar confianza y obtener información sensible.

### Tipo 3: Esquemas piramidales

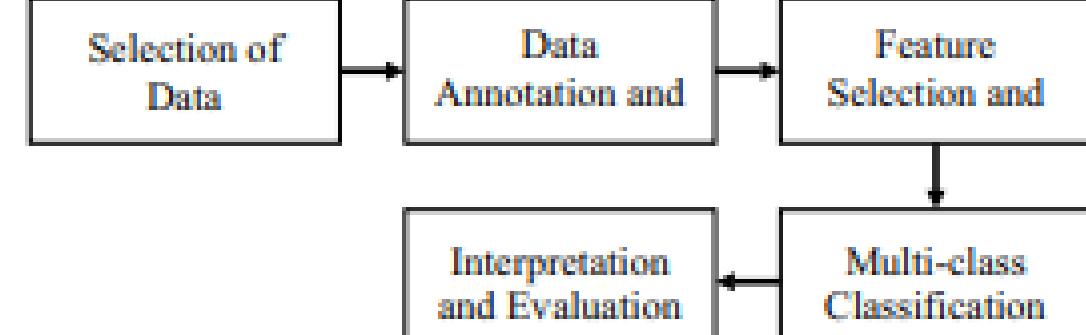
También conocidos como marketing multinivel, donde el anuncio promueve un sistema de reclutamiento que depende de la participación de más personas para obtener ingresos.

# METODOLOGÍA (DATASET Y PREPROCESAMIENTO)

- El dataset utilizado en este estudio es EMSCAD, un conjunto de datos de 17,880 anuncios de empleo, de los cuales 866 son fraudulentos. El dataset tiene un desbalanceo significativo, con 556 anuncios de tipo 1 (robo de identidad), 234 de tipo 2 (suplantación corporativa), y solo 72 de tipo 3 (esquemas piramidales). Para abordar este problema de desbalanceo, se aplicó una técnica de upsampling en los anuncios de tipo 3, lo que permitió balancear las clases.

## Preprocesamiento:

- Los campos de texto como company\_profile, description y requirements se fusionaron en una sola columna llamada "text" para simplificar el proceso. Luego, se realizó una limpieza del texto que incluyó la eliminación de stopwords, tokenización, eliminación de números y puntuación repetida (Figura 5). Esta limpieza ayuda a que el modelo se enfoque en las palabras relevantes para la detección de fraude.

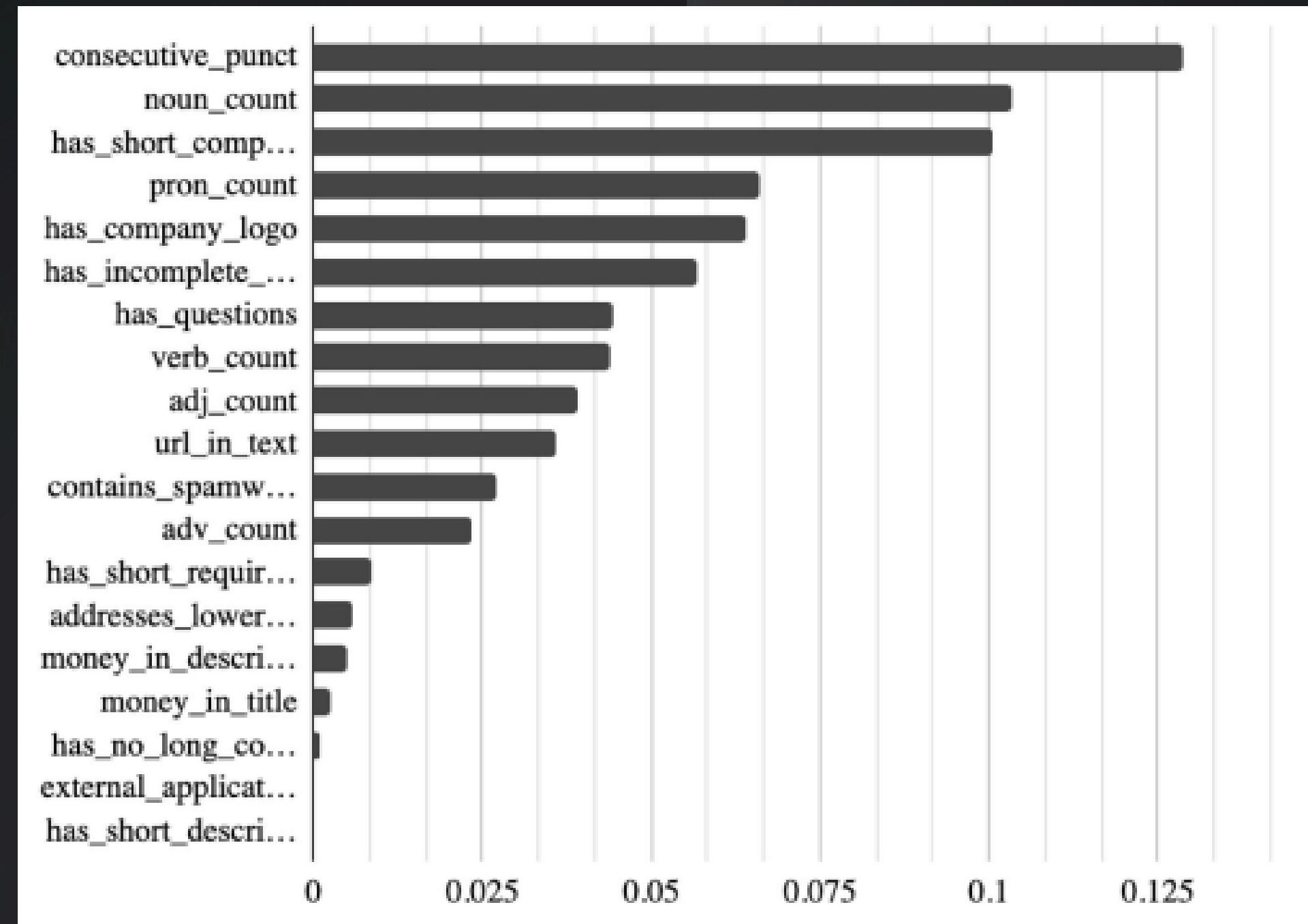


Category	Name	Description
Linguistic	contains_spamwords	Job text contains a spam word such as 'online', 'extra', 'cash'
	consecutive_punct	Number of consecutive punctuation in the job text
	money_in_title	Title contains money symbols
	money_in_description	Description contains money symbols
Contextual	url_in_text	Text contains an e-mail address, phone number or link to an external website
	external_application	Text contains phrases such as 'apply at' or 'send resume'
	addresses_lower_education	Text contains phrases such as 'High School' or 'No degree'
	has_incomplete_extra_attributes	Attributes such as industry, function, required education or employment type are empty
	has_no_company_profile	Profile is empty
	has_short_company_profile	Profile is less than 10 words
	has_no_long_company_profile	Profile is more than 10 words but less than 100 words
	has_short_description	Description is less than 10 words
	has_short_requirements	Requirements are less than 10 words
Metadata	Telecommuting	Job marked as a telecommuting job
	has_no_company_logo	No company logo
	has_no_questions	Screening questions are missing

# CARACTERÍSTICAS CLAVE

Top 5 Features (Figura 5):

1. consecutive\_punct: La puntuación repetida, como en el caso de anuncios que dicen "¡¡Gana dinero!!".
2. noun\_count: Alta frecuencia de sustantivos, especialmente en anuncios de tipo 2 (suplantación corporativa).
3. has\_short\_company\_profile: Perfiles de empresa con menos de 10 palabras, indicativo de robos de identidad (Tipo 1).
4. has\_company\_logo=0: La ausencia de logotipo de la empresa, que se observa en los tipos 1 y 3.
5. money\_in\_title: La presencia de símbolos monetarios, común en los anuncios de tipo 3 (esquemas piramidales).



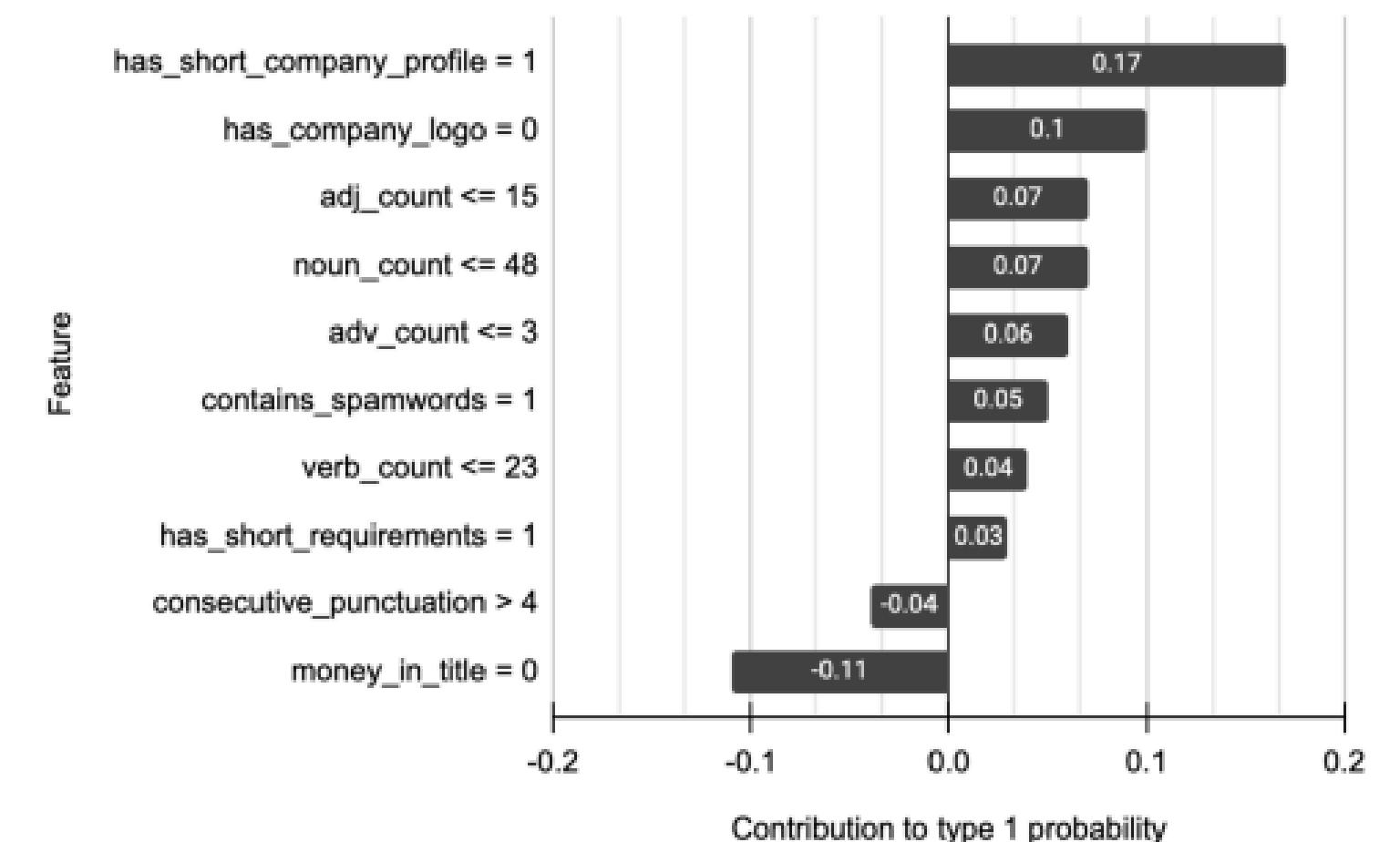
# INTERPRETABILIDAD (LIME)

Ejemplo de predicción para Robo de identidad (Tipo 1):

- Features decisivas:
  - has\_short\_company\_profile=1 (+0.15 peso).
  - has\_company\_logo=0 (+0.12 peso).
  - money\_in\_title=0 (-0.05 peso).

Explicación: Los anuncios con perfiles cortos de empresa y ausencia de logotipo son típicos de fraudes de robo de identidad, lo que facilita la predicción mediante modelos interpretables.

**Fig. 7** Adapted LIME output for an example instance of class type 1



# LIMITACIONES

---

- Dataset antiguo: El dataset EMSCAD fue recopilado entre 2012 y 2014, lo que podría hacer que los resultados no reflejen las tácticas actuales de los estafadores.
- Anonimización: La anonimización de los datos alteró los significados lingüísticos originales, lo que podría afectar la calidad de las predicciones.
- Clase Tipo 2 pequeña: El número limitado de ejemplos para la suplantación corporativa (Tipo 2) podría haber causado sobreajuste y dificultades en la predicción precisa de estos casos.

# FUTURO Y CONCLUSIONES

---

- Se espera explorar el fine-tuning de modelos avanzados como ELECTRA y GPT-3.5, lo que podría mejorar aún más la detección de fraudes.
- Es crucial actualizar el dataset con anuncios de empleo más recientes de plataformas como LinkedIn e Indeed.

# CONTEXTO

- El problema de las ofertas laborales fraudulentas: Con el auge de las plataformas digitales, las ofertas laborales fraudulentas se han convertido en una amenaza creciente. Estas ofertas, que pretenden ser legítimas, buscan engañar a los usuarios, a menudo para obtener datos sensibles o dinero de manera ilícita.
- **Impacto de los Fraudes:**
- Pérdidas económicas: Las ofertas fraudulentas generan pérdidas significativas no solo para los usuarios que caen en el fraude, sino también para las plataformas de empleo que enfrentan costos de moderación.
- **Riesgos de seguridad:** El 34% de los fraudes laborales están relacionados con el robo de identidad, poniendo en riesgo la seguridad de los datos personales de los candidatos.
- 
- **Jóvenes (18-35 años):** 58% de los casos de fraude afectan a personas jóvenes que suelen ser más propensas a caer en engaños debido a su inexperiencia en el mercado laboral.
- **Personas con baja cualificación educativa:** Aproximadamente 72% de los fraudes laborales afectan a personas con menor preparación educativa, quienes son más susceptibles a ofertas engañosas que prometen empleos fáciles.
- **Desempleados de larga duración:** 41% de los fraudes afectan a personas que llevan un largo tiempo sin empleo, quienes buscan oportunidades desesperadamente y, por lo tanto, son más vulnerables a las estafas.



Software engineer in Austin, Texas  
2,702 results

Software Engineer  
CDS Global, Inc.  
Austin, TX, US  
Preferred skills and experience. Proven successful experience making decisions on variety of tasks requiring discretion, judgment,...

Sr. Software Engineer in Test  
UFCU  
Austin, TX, US  
Eight years of quality assurance experience; may substitute related job qualifications such as production/customer support or busi...

Sr. Software Engineer  
Main Street Hub  
Austin, TX, US  
The Engineering Team is on a mission to build the absolute best software in the local space. 7+ years in software development ...

Experienced Installation Software Engineer  
Emerson Automation Solutions  
Austin, TX, US  
The software is responsible for installing new software modules and...

**Job description**

**Job**

- 6 applicants
- Mid-Senior level

**Company**

- 1001-5000 employees
- Outsourcing/Offshoring

**Connections**

You have 0 connections at this company. Add >

**Seniority Level**  
Mid-Senior level

**Industry**  
Outsourcing/Offshoring

**Employment Type**  
Full-time

**Job Functions**  
Engineering

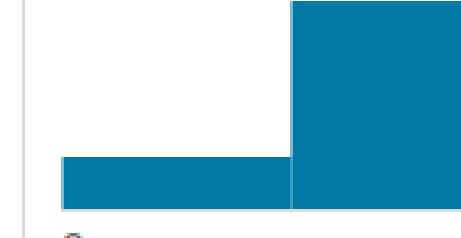
# DATASET

job_id	title	location	department	salary_range
Unique Job ID	The title of the job ad entry.	Geographical location of the job ad.	Corporate department (e.g. sales).	Indicative salary range (e.g. \$50,000-\$60,000)
1 	English Teacher A... 2% Customer Service ... 1% Other (17423) 97%	GB, LND, London 4% US, NY, New York 4% Other (16504) 92%	[null] 65% Sales 3% Other (5782) 32%	[null] 84% 0-0 1% Other (2726) 15%
10	Customer Service Associate - Part Time	US, AZ, Phoenix		
11	ASP.net Developer Job opportunity at United States, New Jersey	US, NJ, Jersey City		100000-120000
12	Talent Sourcer (6 months fixed-term contract)	GB, LND, London	HR	

# DATASET

company_profile	description	requirements	benefits	# telecommuting
A brief company description.	The details description of the job ad.	Enlisted requirements for the job opening.	Enlisted offered benefits by the employer.	True for telecommuting positions.
[null] 19% We help teachers ... 4% Other (13846) 77%	Play with kids, get ... 2% Play with kids, get... 0% Other (17435) 98%	[null] 15% University degree ... 2% Other (14776) 83%	[null] 40% See job description 4% Other (9948) 56%	
Management using best of breed ...	Health report doc	required.Additional Tools:HP BSM Application...		
Novitex Enterprise Solutions, formerly Pitney Bowes Management Services, delivers innovative documen...	The Customer Service Associate will be based in Phoenix, AZ. The right candidate will be an integral...	Minimum Requirements:Minimum of 6 months customer service related experience requiredHigh school dip...		0
	Position : #URL_86fd830a95a64e2 b30ceed829e63fd384c2 89e4f01e3c93608b42a8 4f6e662dd# DeveloperJob Locat...	Position : #URL_86fd830a95a64e2 b30ceed829e63fd384c2 89e4f01e3c93608b42a8 4f6e662dd# DeveloperJob Locat...	Benefits - FullBonus Eligible - YesInterview Travel Reimbursed - Yes	0
Want to build a 21st century financial service?We're convinced that that there is a need for innovat...	TransferWise is the clever new way to move money between countries. Co-founded by Skype's first empl...	We're looking for someone who:Proven track record in sourcing across marketing, banking & buildi...	You will join one of Europe's most hotly tipped startups with plenty of opportunities to grow and th...	0

# DATASET

# telecommuting	# has_company_logo	# has_questions	△ employment_type	△ required_experie...
True for telecommuting positions.	True if company logo is present.	True if screening questions are present.	Full-type, Part-time, Contract, etc.	Executive, Entry level, Intern, etc.
			Full-time [null] Other (2789)	65% 19% 16%
0	1	0	Part-time	Entry level
0	0	0	Full-time	Mid-Senior level
0	1	0		

# DATASET

	A required_experience	A required_education	A industry	A function	# fraudulent
	Executive, Entry level, Intern, etc.	Doctorate, Master's Degree, Bachelor, etc.	Automotive, IT, Health care, Real estate, etc.	Consulting, Engineering, Research, Sales etc.	target - Classification attribute.
	[null] 39% Mid-Senior level 21% Other (7021) 39%	[null] 45% Bachelor's Degree 29% Other (4630) 26%	[null] 27% Information Tech... 10% Other (11243) 63%	[null] 36% Information Tech... 10% Other (9676) 54%	
	Entry level	High School or equivalent	Financial Services	Customer Service	0
	Mid-Senior level	Bachelor's Degree	Information Technology and Services	Information Technology	0
					0

# Problema

El problema central de este análisis es la gran cantidad de datos mixtos provenientes de diversas fuentes (textos, numéricos, categóricos, etc.) que dificulta un análisis profundo y eficiente debido a su estructura compleja.

- 1. Características del dataset:** El dataset está compuesto por datos mixtos (textos, datos numéricos, categóricos) que incluyen información sobre ofertas de trabajo, como salario, tipo de empleo, ubicación, requisitos, entre otros.
- 2. Desafíos del análisis:** Los datos no están estructurados de manera convencional, lo que hace difícil aplicar técnicas tradicionales de análisis de datos.
- 3. Consecuencias:** No se realiza un análisis exhaustivo de este tipo de datos, lo que limita la capacidad de detectar patrones y hacer predicciones efectivas.

# Objetivo

Desarrollar una herramienta visual de detección de fraudes laborales basado en el análisis de datos multivariados y un sistema de scoring interpretable, que permita identificar patrones fraudulentos de manera eficaz y comprensible para los usuarios.

## Objetivos Específicos:

1. Realizar un análisis exploratorio de los datos para identificar patrones preliminares en las variables.
2. Implementar un sistema de scoring interpretable que permita clasificar las ofertas en fraudulentas y no fraudulentas basándose en patrones detectados.
3. Correlacionar variables que influyen en la probabilidad de fraude, como tipo de empleo, salario, ubicación, y longitud de las descripciones.
4. Visualizar los resultados de manera interactiva, permitiendo al usuario explorar diferentes escenarios y detectar patrones de fraude.

# GRACIAS