Abstract
A continuación se desarrollará las características de una base de datos que corresponde a las ofertas laborales del año 2023 que tiene Linkedln.
Linkedln es una red social fundada en el año 2002 orientada al uso empresarial, a los negocios y al empleo. En esta red adhieren usuarios que revelan su experiencia laboral y destrezas poniéndose en contacto con miles de empresas y personas. A simple vista podemos distinguir dos actores fundamentales que interactúan en esta red: empleadores y empleados.
Cada usuario de esta red crea un perfil que le permite conectarse con otros usuarios, buscar empleo, publicar ofertas de trabajo, buscar candidatos para puestos laborales y promulgar contenido relacionado a las actividades.
Linkedln es una plataforma tecnológica que conecta a personas y empresas proporcionando herramientas para que los usuarios que la integran puedan encontrar empleo y/o a candidatos para los puestos laborales, reduciendo espacio y tiempo.

Motivación
El desarrollo laboral es una actividad productiva, colectiva, social e individual que genera un impacto en la vida de las personas y en los sistemas sociales y productivos.
A diferencia de varias décadas atrás, las actividades productivas actuales generan un mercado laboral heterogéneo. Esto quiere decir que no todos los trabajos son iguales y que hay más diversidad de habilidades y aptitudes en los trabajadores. Esta heterogeneidad impacta tanto en la oferta y demanda laboral, en consecuencia, si analizamos este data set podremos realizar algunas observaciones de lo que sucedió en esa red social durante el año 2023.

Audiencia
Este análisis encuentra orientado a ser utilizado a nivel táctico, es decir por lo mandos medios de la organización. Su utilización permitirá comprender las ofertas laborales, el salario, la localización, los tipos de contratos y trabajos, entre otras cosas.

Contexto comercial
La organización en cuestión es una empresa especializada en recursos humanos. Su objetivo principal es facilitar el contacto entre personas que buscan empleo y empresas que están en proceso de contratación. Actualmente, la compañía busca optimizar su red de conexiones entre la oferta y la demanda laboral.

Problema comercial
Dificultad para contactar entre personas que buscan trabajo y empresas que lo ofrecen.
Dificultad para recopilar suficientes datos sobre los salarios y los factores que influyen.

Objetivo Analítico y Contexto
Definición de objetivo
Se tiene la intención de entrenar un algoritmo para detectar cuáles son los factores que influyen en el tipo de salario. El objetivo es predecir el tipo de salario basándose en diferentes características. Se utilizará un data set que contiene 27 variables asociadas a Linkedln con propiedades de la composición del mercado laboral formal.

Contexto
Se tiene un dataset con registros de puestos laborales publicados en la plataforma Linkedln del año 2023. La idea es determinar qué características se deben tener en cuenta para predecir el tipo de salario. Este modelo puede ser utilizado para tener información precisa sobre el salario en relación al puesto laboral.

Pregunta principal
¿Cuál será el salario medio según el tipo de trabajo?
Preguntas secundarias
¿Cuáles son las categorías de empleo más solicitadas?
¿Cuáles son los principales puestos de trabajo con los salarios medios más altos?
¿Cuáles son los puestos de trabajo que recibieron mayor número de solicitudes?
¿Cuál es el salario promedio de los puestos de trabajo que recibieron la mayor cantidad de solicitudes?
¿Cuál es el porcentaje de ofertas de empleo patrocinadas?
¿Qué porcentaje de las visualizaciones totales corresponden a las ofertas de empleo patrocinadas?

Información del dataset

- Registro de puestos laborales publicados en el año 2023 en la plataforma LinkedIn.
- Variables que describen los puestos laborales, la cantidad de visualizaciones, aplicaciones, lugar de trabajo y sponsoreos, entre otras.

Distinción entre variables categóricas o numéricas
Para distinguirlas se analiza la información que anteriormente aportó el método df.info ()

VARIABLES CATEGÓRICAS --> object

- 2 title: puesto del trabajo
- 3 description: descripción del puesto de trabajo
- 7 pay_period: período de pago del salario (yearly, hourly, monthly)
- 8 formatted_work_type: tipo de trabajo (full-time, contract, part-time, temporary, intership, other, volunteer)
- 9 location: lugar de trabajo
- 14 job_posting_url: url a la oferta de empleo en una plataforma
- 15 application_url: url donde se pueden presentar las solicitude
- 16 application_type: tipo de proceso de solicitud (offsiteapply, complexonsiteapply, simpleonsiteapply)
- 19 formatted_experience_level: nivel de experiencia laboral (mid-senior level, entry level, associete, director, intership, executive)
- 20 skills_desc: descripción que detalla las habilidades requeridas para el trabajo
- 22 posting_domain:dominio del sitio web con aplicación
- 24 work_type: tipo de trabajo asociado al puesto de trabajo (full-time, contract, part-time, temporary, intership, other, volunteer)
- 25 currency: moneda en la que se proporciona el salario.
- 26 compensation_type: tipo de compensación por el trabajo

VARIABLES NUMÉRICAS --> float64 e inter64

- 0 job_id: el ID del trabajo tal y como lo define LinkedIn (https://www.linkedin.com/jobs/view/ *job_id*)
- 1 company_id: identificador de la empresa asociada a la oferta de empleo (se asigna a empresas.csv)
- 4 max_salary: salario máximo
- 5 med_salary: salario medio
- 6 min_salary: salario mínimo
- 10 applies: números de solicitudes que se presentaron
- 11 original_listed_time: hora orginal que se postuló el trabajo
- 12 remote_allowed: si el trabajo es remoto

- 13 views: número de veces que se vieron las ofertas de empleo
- 17 expiry: fecha u hora de caducidad de la oferta de empleo
- 18 closed_time: hora de cerrar la oferta de empleo
- 21 listed_time: hora en la que se publicó el trabajo
- 23 sponsored: si la oferta de empleo está patrocinada o promocionada

ABC - Análisis del dataset

Tamaño del dataset: 15886 registros por 27 columnas

Data columns (total 27 columns):

#	Column	Non-Null Count	Dtype	Suma de nulos
0	job_id	15886 non-null	int64	0
1	company_id	15520 non-null	float64	366
2	title	15886 non-null	object	0
3	description	15885 non-null	object	1
4	max_salary	5521 non-null	float64	10365
5	med_salary	981 non-null	float64	14905
6	min_salary	5521 non-null	float64	10365
7	pay_period	6502 non-null	object	9384
8	formatted_work_type	15886 non-null	object	0
9	location	15886 non-null	object	0
10	applies	8700 non-null	float64	7186
11	original_listed_time	15886 non-null	float64	0
12	remote_allowed	2340 non-null	float64	13546
13	views	13123 non-null	float64	2763
14	job_posting_url	15886 non-null	object	0
15	application_url	9795 non-null	object	6091
16	application_type	15886 non-null	object	0
17	expiry	15886 non-null	float64	0
18	closed_time	928 non-null	float64	14958
19	formatted_experience_level	10984 non-null	object	4902
20	skills_desc	144 non-null	object	15742
21	listed_time	15886 non-null	float64	0
22	posting_domain	9044 non-null	object	6842
23	sponsored	15886 non-null	int64	0
24	work_type	15886 non-null	object	0
25	currency	6502 non-null	object	9384
26	compensation_type	6502 non-null	object	9384

dtypes: float64(11), int64(2), object(14)

Tratamiento de Datos Nulos

1) Los datos nulos en las variables normales los reemplazo con la media
2) Los datos nulos en las variables no normales los reemplazo con la mediana
3) Los datos nulos en las variables categóricas los reemplazo con la moda

Correlación entre algunas variables

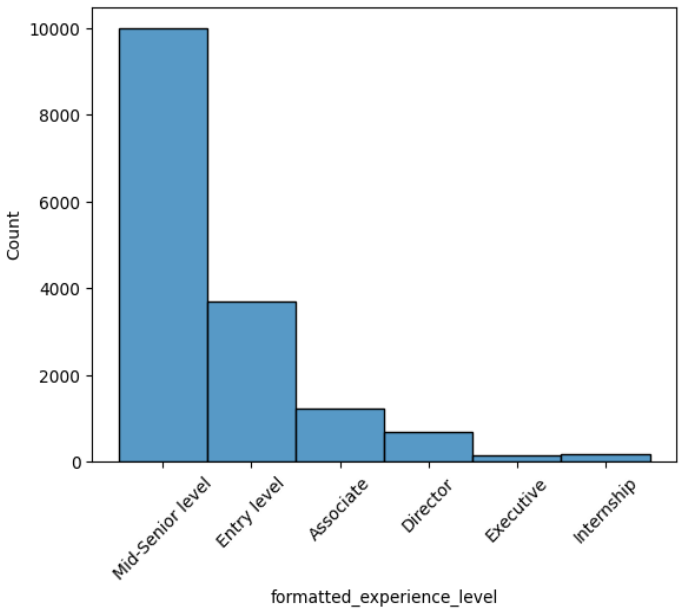
application_url - posting_domain (0.87)
max_salary - min_salary (0.9)
applies - views (o.83)

Se debe eliminar alguna de estas variables porque no se puede hacer la regresión si hay algunas variables que dependen de otras. Entonces las variables que se eliminaran son: posting_domain, max_salary, mid_salary y views.
Los gráficos de distribuciones muestran que en las variables con distribución no normal hay demasiados datos atípicos. Por eso, se toma la decisión de continuar con el modelo sin

reemplazar los outliers de las variables_no_normales. Los valores atípicos pueden indicar variabilidad en los datos o errores de medición.
La variable que quiero predecir es “med_salary” que contiene muchos valores atípicos y podría ser que estos valores atípicos representen una variabilidad real en los salarios que es importante para el modelo predictivo. Eliminar estos valores atípicos podría simplificar demasiado el modelo y hacer que sea menos preciso para predecir salarios que son inusualmente altos o bajos.

EDA: Exploratory Data Analysis

“Categorías de empleo más solicitadas”



Mid-Senior level 62%

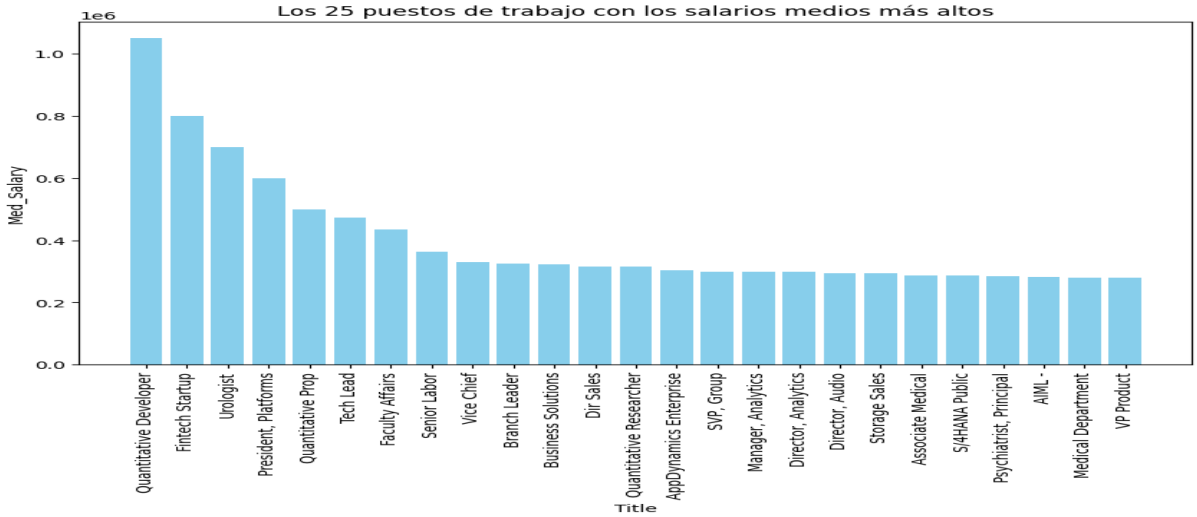
Entry level 23%

Associate 7%

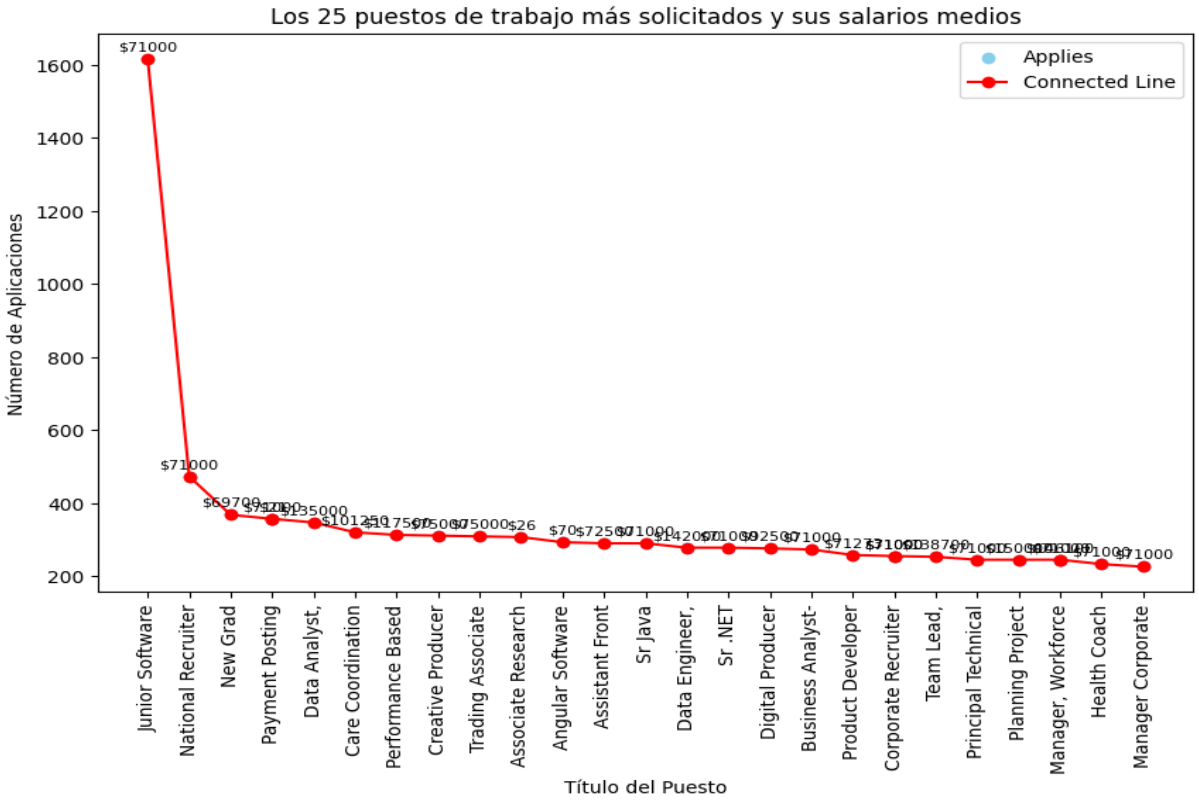
Director 4%

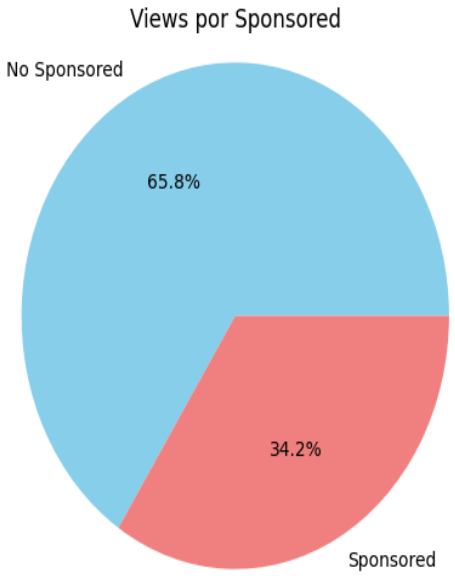
Internship 1%

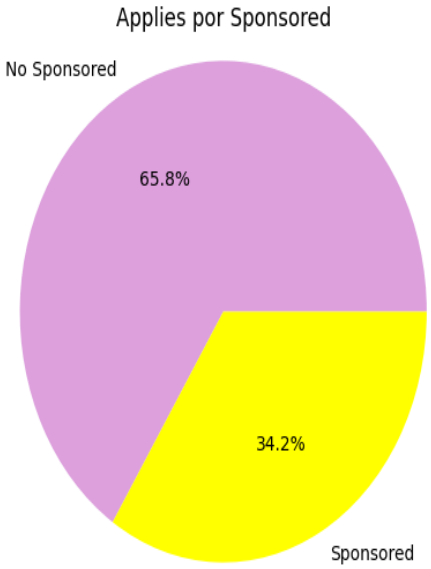
“Los 25 puestos de trabajo con los salarios medios más altos”



“Los 25 puestos de trabajo más solicitados y sus salarios medios”







INSIGHTS

- En las ofertas de empleo publicadas en LinkedIn hay un preredominio de niveles medios a altos: El 62% de las posiciones son de nivel medio a alto (Mid-Senior level). Esto sugiere que hay una alta demanda de profesionales con experiencia y habilidades avanzadas.

- Oportunidades para principiantes: Aproximadamente una cuarta parte de las posiciones (23%) son de Entry level. Esto indica que también hay oportunidades para aquellas personas que están comenzando sus carreras o que buscan entrar en un nuevo campo.

- Hay pocas posiciones de liderazgo y pasantías: Las posiciones de nivel de Director y las pasantías representan una pequeña proporción de las posiciones (4% y 1% respectivamente). Esto podría sugerir que las oportunidades para avanzar al nivel de liderazgo o para obtener experiencia como pasante son limitadas.

- Que las ofertas de empleo sean mayoritariamente no patrocinadas puede surgir por dos factores. Por un lado, los usuarios de LinkedIn suelen ser muy activos en la búsqueda de empleo y pueden encontrar estas ofertas de empleo no patrocinadas a través de búsquedas regulares o recomendaciones personalizadas. Por otro lado, las empresas pueden tener estrategias otras efectivas que les permiten atraer vistas y aplicaciones sin necesidad de patrocinio.

- Aunque el patrocinio puede aumentar la visibilidad de una oferta de empleo, no es el único factor que determina el número de vistas o aplicaciones que recibe. La calidad de la oferta de empleo, la reputación de la empresa y la eficacia de su estrategia de marketing también son factores importantes.

RECOMENDACIONES

- En base a lo observado sería relevante crear un algoritmo de machine learning que nos ayude a predecir la variable objetivo que es el tipo de salario medio por el tipo de trabajo o por el tipo de nivel de experiencia.