PROYECTO

Presentación final aprendizaje estadístico

Presentado por:

Julio Avila
Guillermo Furlan

AGENDA

01 Información General del proyecto

02 Objetivos

03 Análisis exploratorio

04 Clasificacion

05 Modelo de preiccion

06 final

Contexto

Se tiene una base de datos que contiene información acerca de empleos relacionados a datos, se cuenta entre otras variables; el salario promedio, puesto, empresa donde se trabaja y lista de habilidades colocadas en el cv.

Objetivos

01

ELABORAR UN MODELO PARA PREDECIR EL RANGO SALARIAL PARA UN PUESTO EN CIENCIA DE DATOS

02

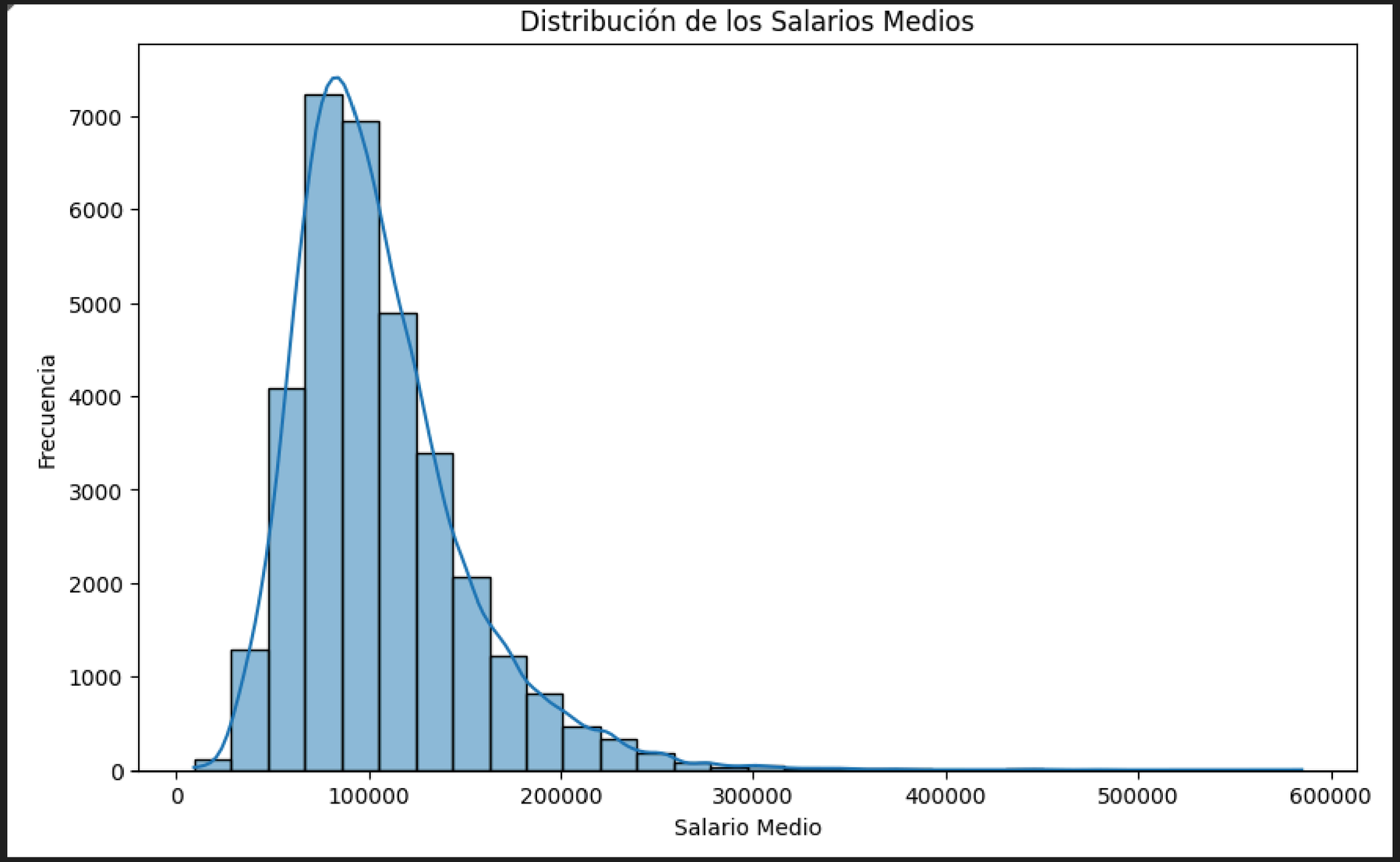
CLASIFICAR EL PERFIL DE UNA PERSONA BASADO EN SUS HBAILIDADES

ANÁLISIS EXPLORATORIO

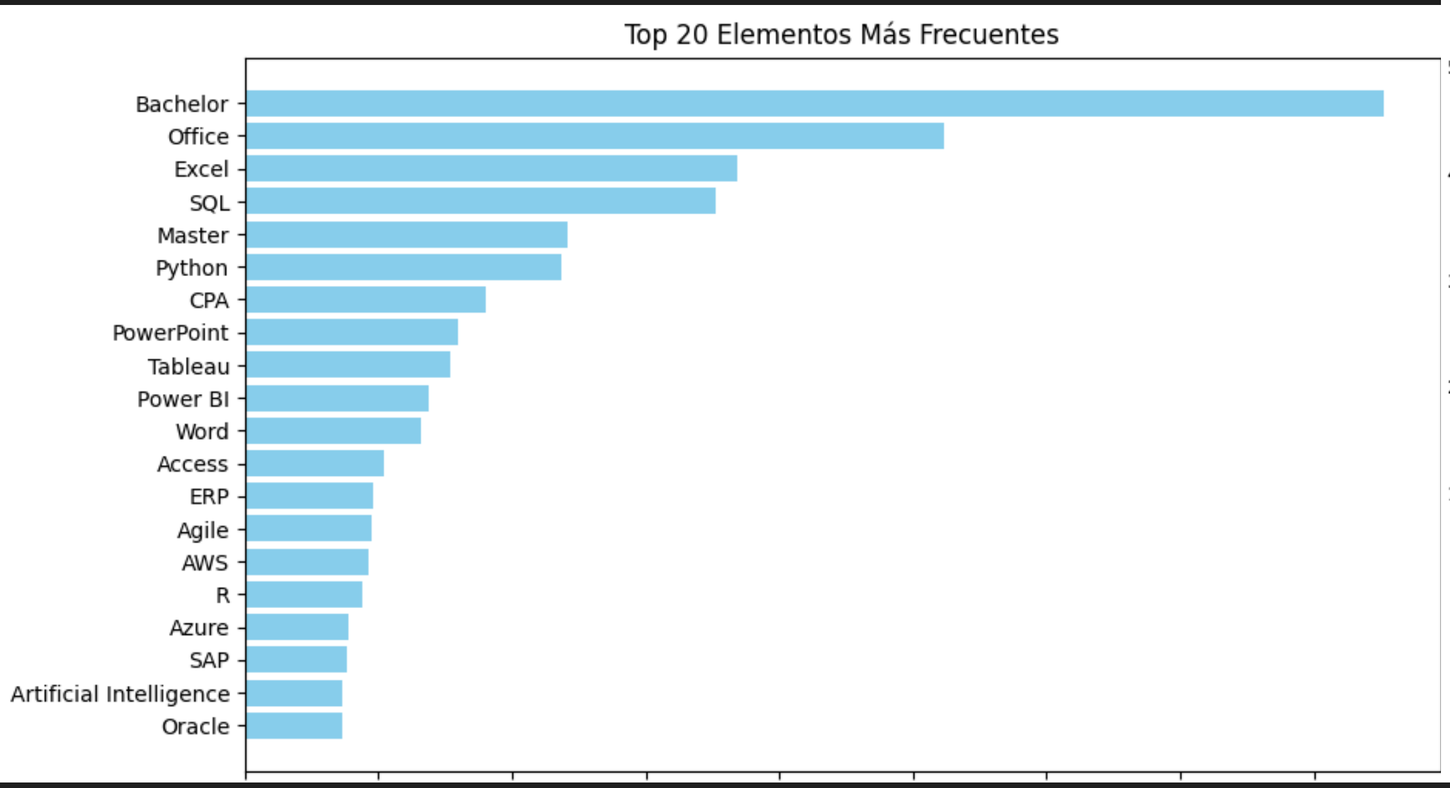
Resultado

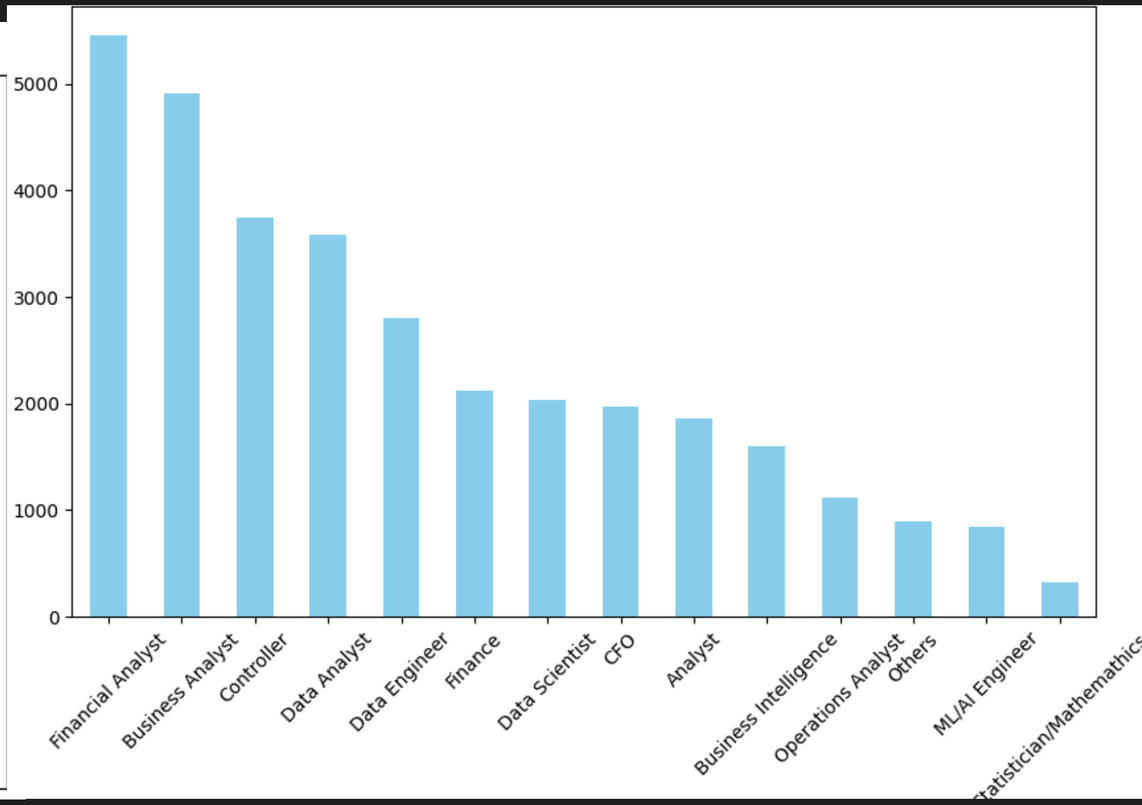
#	Column	Non-Null Count	Dtype
0	ID	33248 non-null	object
1	Job	33248 non-null	object
2	Jobs_Group	33248 non-null	object
3	Profile	12141 non-null	object
4	Remote	13929 non-null	object
5	Company	33239 non-null	object
6	Location	33235 non-null	object
7	City	29424 non-null	object
8	State	30136 non-null	object
9	Frecuency_Salary	33248 non-null	object
10	Mean_Salary	33248 non-null	float64
11	Skills	33248 non-null	object
12	Sector	26034 non-null	object
13	Sector_Group	26034 non-null	object
14	Revenue	14930 non-null	object
15	Employee	20449 non-null	object
16	Company_Score	24486 non-null	float64
17	Reviews	24486 non-null	float64
18	Director	12463 non-null	object
19	Director_Score	11324 non-null	float64
20	URL	17215 non-null	object

Resultado

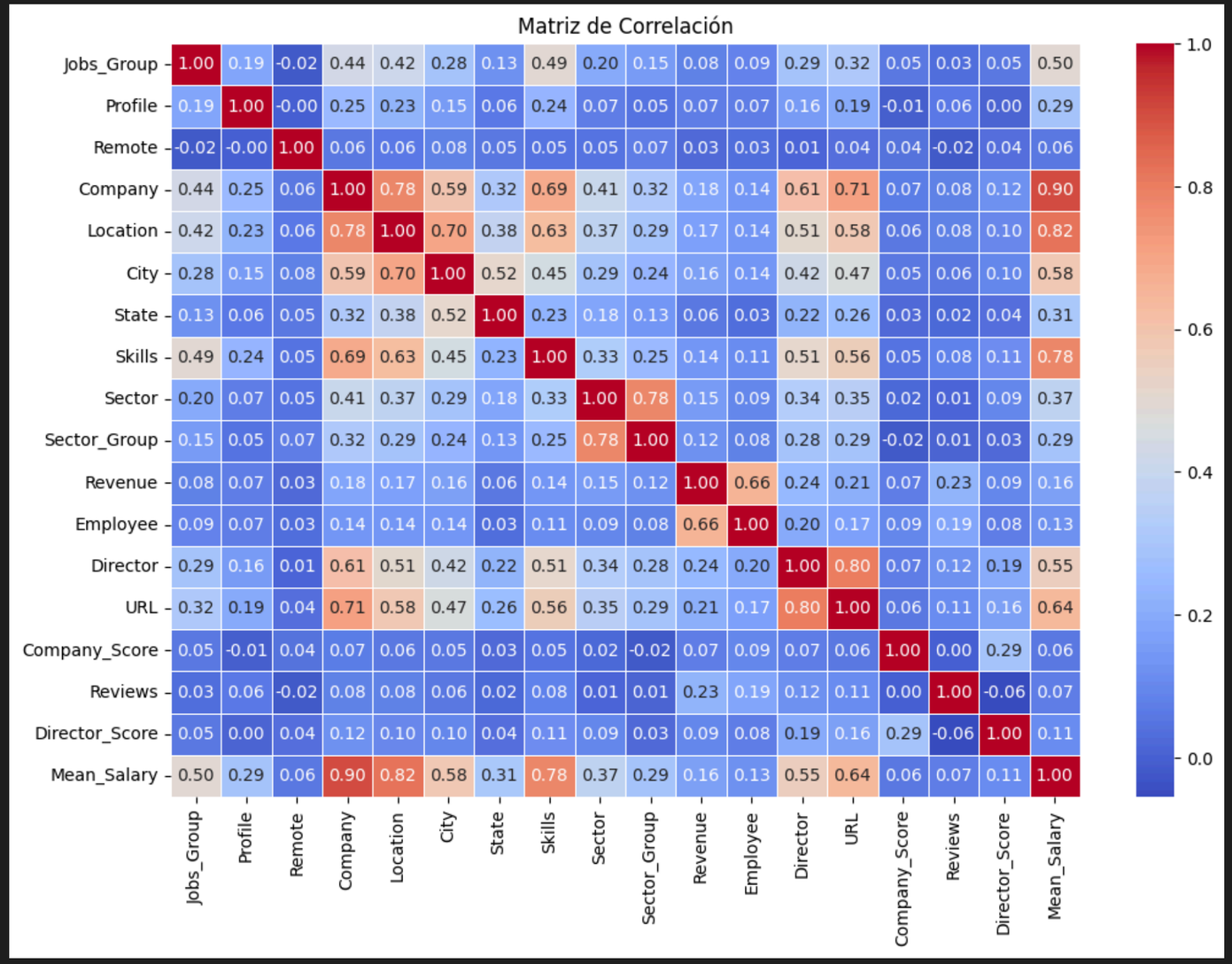

Distribución de los Salarios Medios
Frecuencia
Salario Medio
0
1000
2000
3000
4000
5000
6000
7000
0
100000
200000
300000
400000
500000
600000

Resultado


Top 20 Elementos Más Frecuentes
Bachelor
Office
Excel
SQL
Master
Python
CPA
PowerPoint
Tableau
Power BI
Word
Access
ERP
Agile
AWS
R
Azure
SAP
Artificial Intelligence
Oracle


5000
4000
3000
2000
1000
0
Financial Analyst
Business Analyst
Controller
Data Analyst
Data Engineer
Finance
Data Scientist
CFO
Analyst
Business Intelligence
Operations Analyst
Others
ML/AI Engineer

Resultado


Matriz de Correlación
Jobs_Group 1.00 0.19 -0.02 0.44 0.42 0.28 0.13 0.49 0.20 0.15 0.08 0.09 0.29 0.32 0.05 0.03 0.05 0.50
Profile 0.19 1.00 -0.00 0.25 0.23 0.15 0.06 0.24 0.07 0.05 0.07 0.07 0.16 0.19 -0.01 0.06 0.00 0.29
Remote -0.02 -0.00 1.00 0.06 0.06 0.08 0.05 0.05 0.05 0.07 0.03 0.03 0.01 0.04 0.04 -0.02 0.04 0.06
Company 0.44 0.25 0.06 1.00 0.78 0.59 0.32 0.69 0.41 0.32 0.18 0.14 0.61 0.71 0.07 0.08 0.12 0.90
Location 0.42 0.23 0.06 0.78 1.00 0.70 0.38 0.63 0.37 0.29 0.17 0.14 0.51 0.58 0.06 0.08 0.10 0.82
City 0.28 0.15 0.08 0.59 0.70 1.00 0.52 0.45 0.29 0.24 0.16 0.14 0.42 0.47 0.05 0.06 0.10 0.58
State 0.13 0.06 0.05 0.32 0.38 0.52 1.00 0.23 0.18 0.13 0.06 0.03 0.22 0.26 0.03 0.02 0.04 0.31
Skills 0.49 0.24 0.05 0.69 0.63 0.45 0.23 1.00 0.33 0.25 0.14 0.11 0.51 0.56 0.05 0.08 0.11 0.78
Sector 0.20 0.07 0.05 0.41 0.37 0.29 0.18 0.33 1.00 0.78 0.15 0.09 0.34 0.35 0.02 0.01 0.09 0.37
Sector_Group 0.15 0.05 0.07 0.32 0.29 0.24 0.13 0.25 0.78 1.00 0.12 0.08 0.28 0.29 -0.02 0.01 0.03 0.29
Revenue 0.08 0.07 0.03 0.18 0.17 0.16 0.06 0.14 0.15 0.12 1.00 0.66 0.24 0.21 0.07 0.23 0.09 0.16
Employee 0.09 0.07 0.03 0.14 0.14 0.14 0.03 0.11 0.09 0.08 0.66 1.00 0.20 0.17 0.09 0.19 0.08 0.13
Director 0.29 0.16 0.01 0.61 0.51 0.42 0.22 0.51 0.34 0.28 0.24 0.20 1.00 0.80 0.07 0.12 0.19 0.55
URL 0.32 0.19 0.04 0.71 0.58 0.47 0.26 0.56 0.35 0.29 0.21 0.17 0.80 1.00 0.06 0.11 0.16 0.64
Company_Score 0.05 -0.01 0.04 0.07 0.06 0.05 0.03 0.05 0.02 -0.02 0.07 0.09 0.07 0.06 1.00 0.00 0.29 0.06
Reviews 0.03 0.06 -0.02 0.08 0.08 0.06 0.02 0.08 0.01 0.01 0.23 0.19 0.12 0.11 0.00 1.00 -0.06 0.07
Director_Score 0.05 0.00 0.04 0.12 0.10 0.10 0.04 0.11 0.09 0.03 0.09 0.08 0.19 0.16 0.29 -0.06 1.00 0.11
Mean_Salary 0.50 0.29 0.06 0.90 0.82 0.58 0.31 0.78 0.37 0.29 0.16 0.13 0.55 0.64 0.06 0.07 0.11 1.00
Jobs_Group
Profile
Remote
Company
Location
City
State
Skills
Sector
Sector_Group
Revenue
Employee
Director
URL
Company_Score
Reviews
Director_Score
Mean_Salary
1.0
0.8
0.6
0.4
0.2
0.0

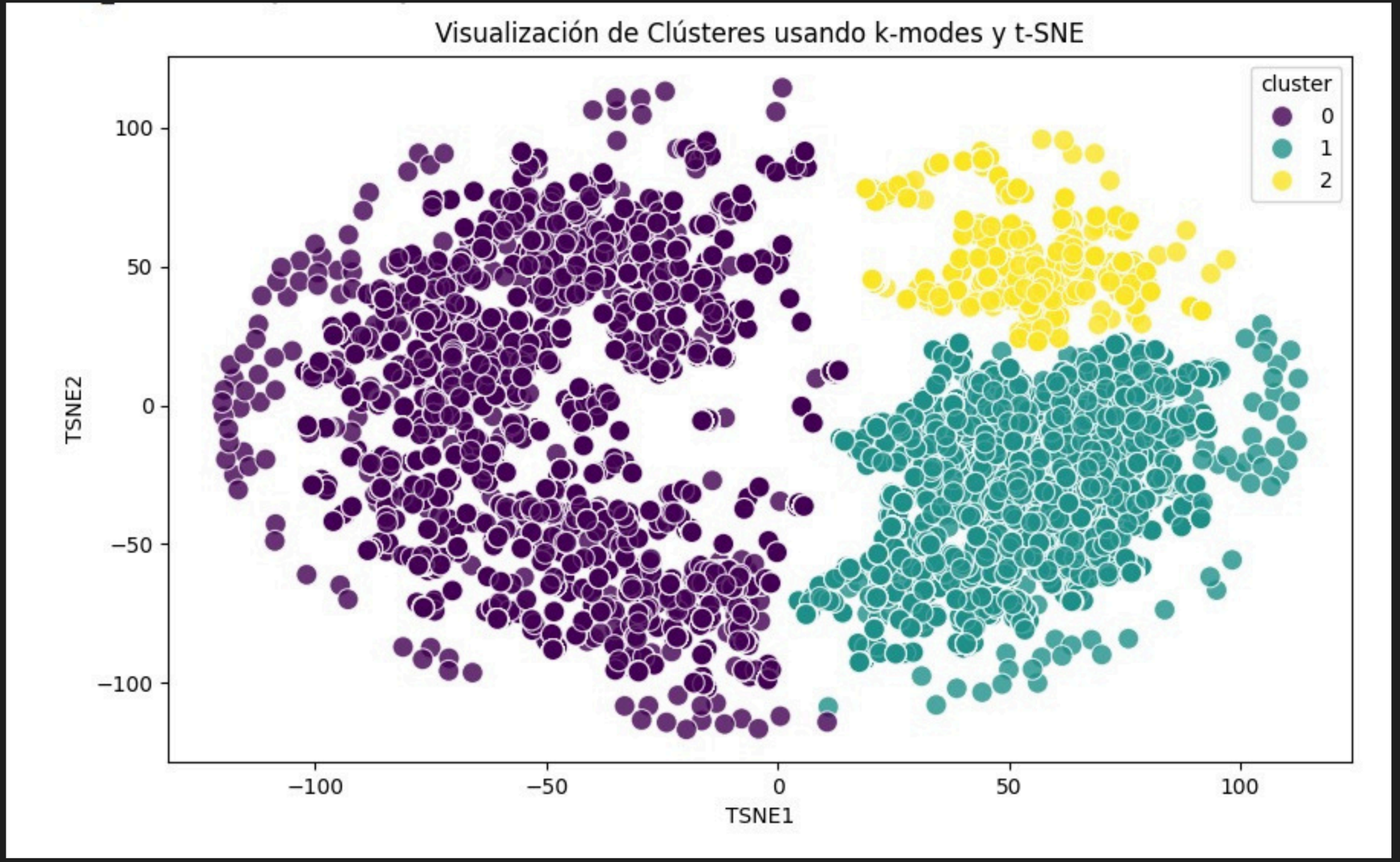
CLASIFICACIÓN

Tecnicas usadas

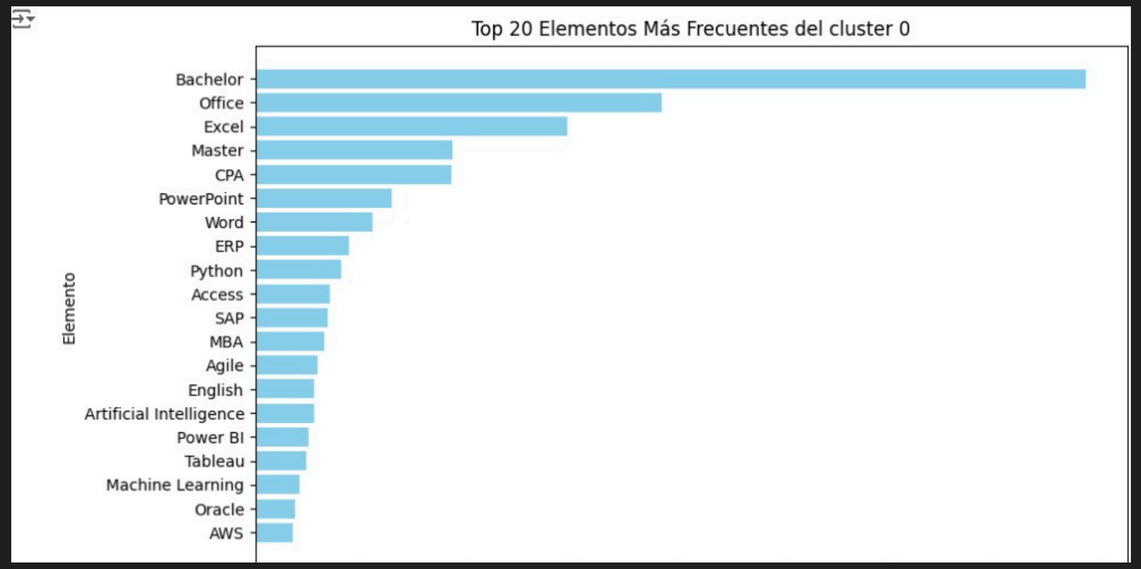
- K-MODES: el algoritmo K-Modes es una variante del algoritmo K-Means diseñada específicamente para datos categóricos. Usando la metrica Hamming

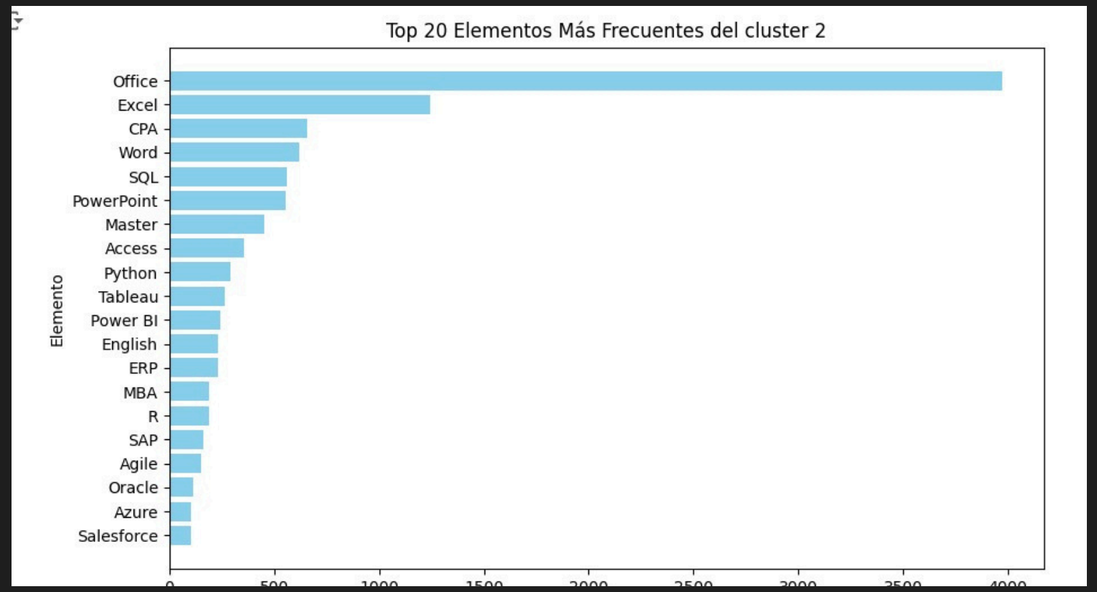
- TSNE:t-distributed Stochastic Neighbor Embedding, es un algoritmo de reducción de dimensionalidad no lineal

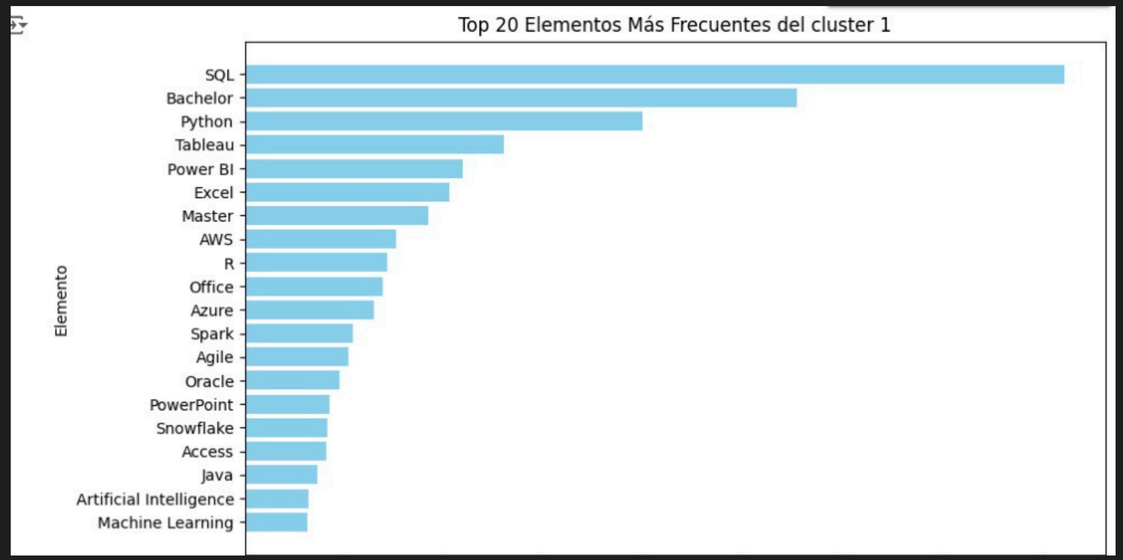
Resultado


Visualización de Clústeres usando k-modes y t-SNE
cluster
0
1
2
100
50
0
−50
−100
TSNE2
−100
−50
0
50
100
TSNE1

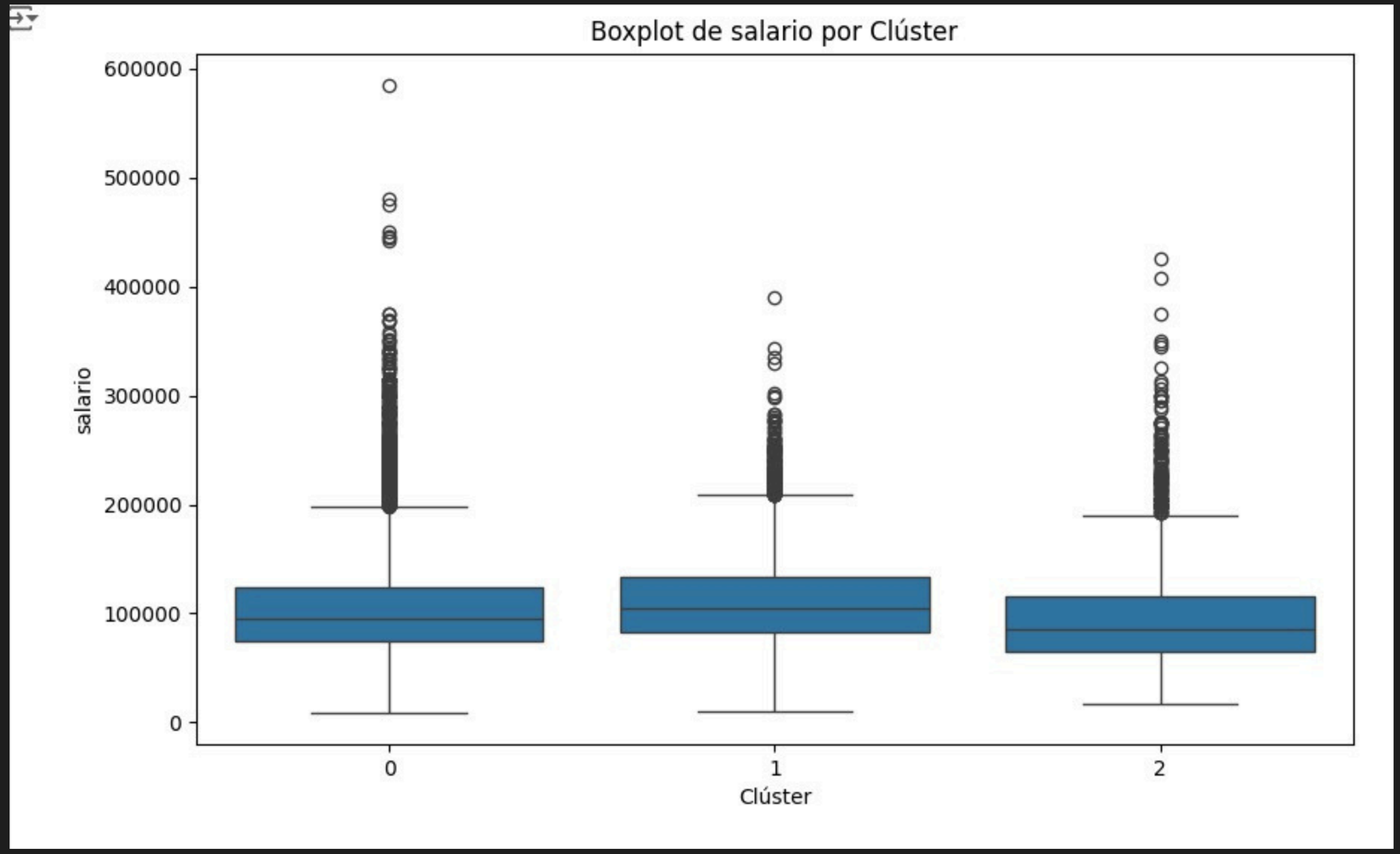
Habilidades



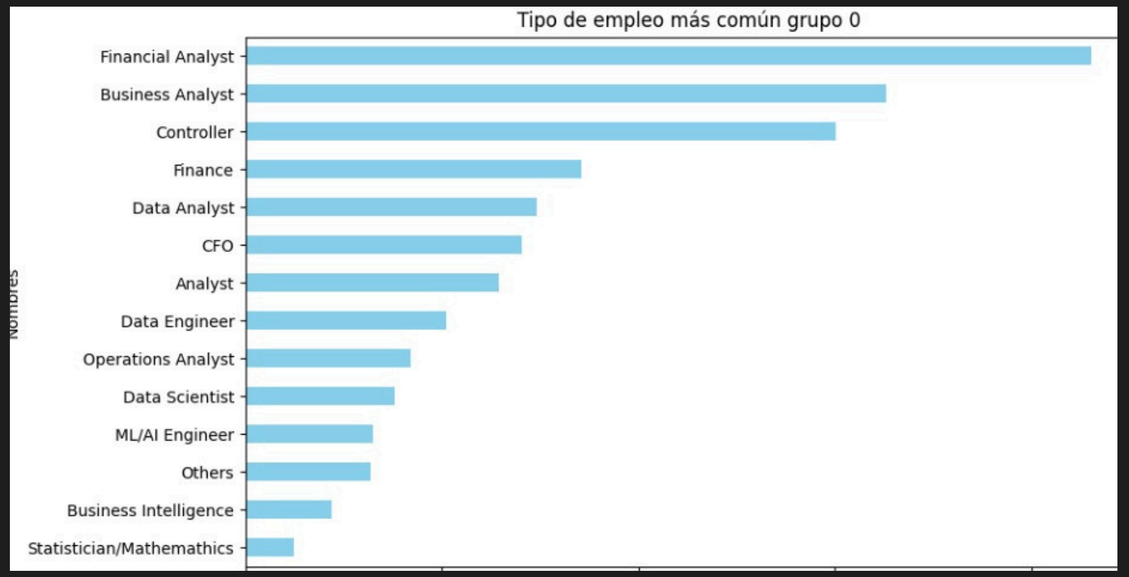


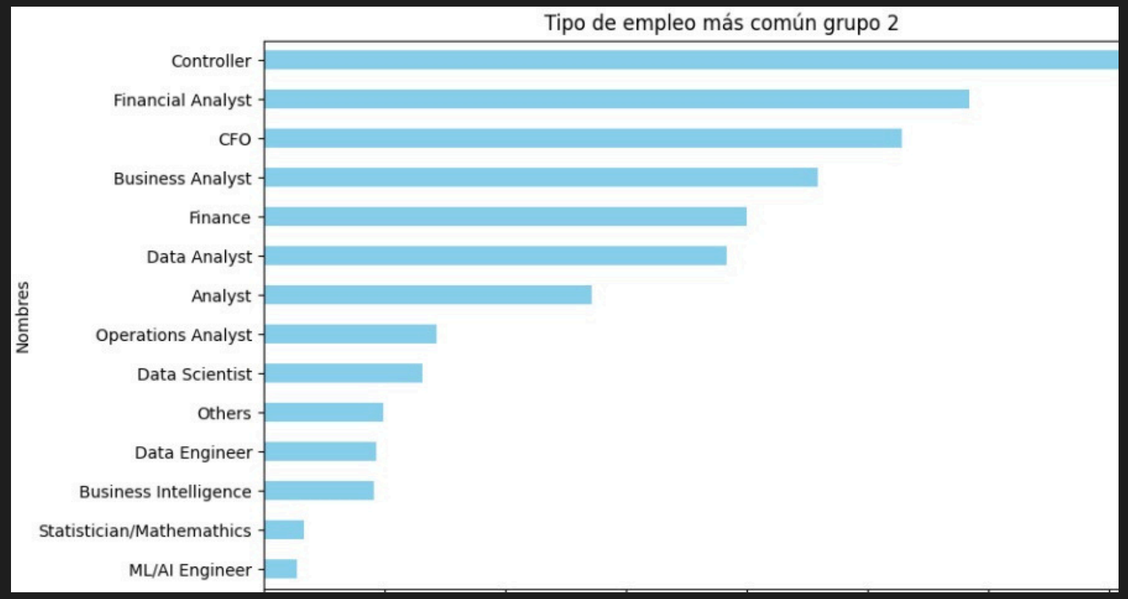


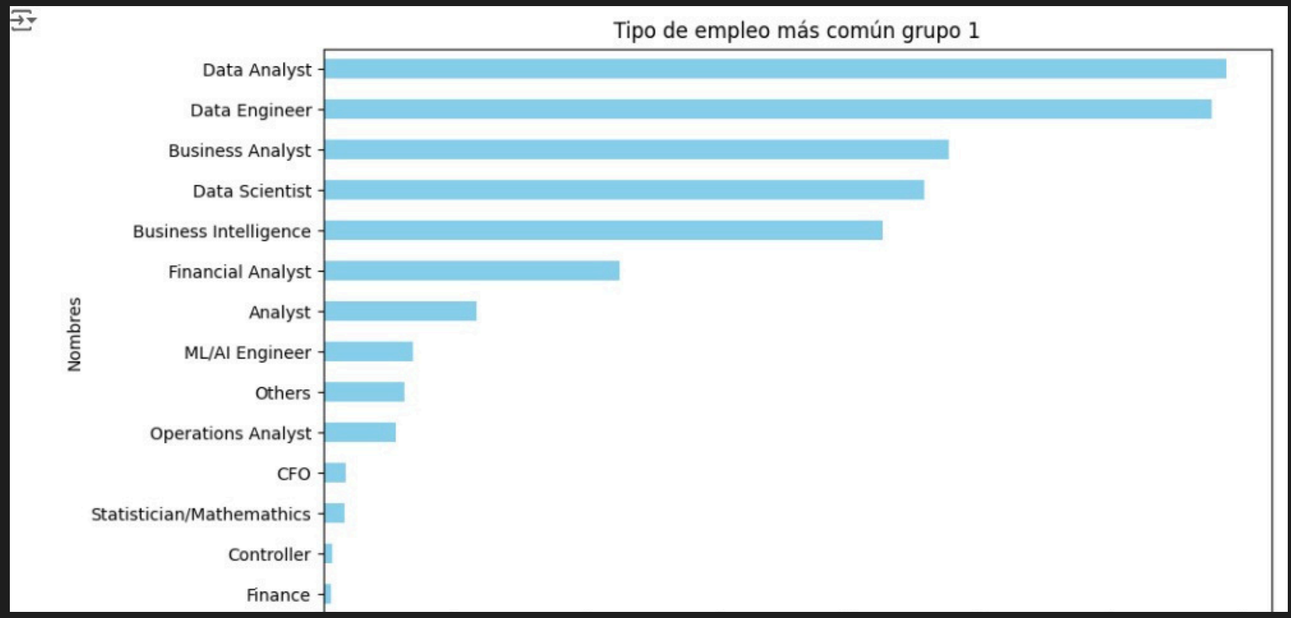
Salario Medio



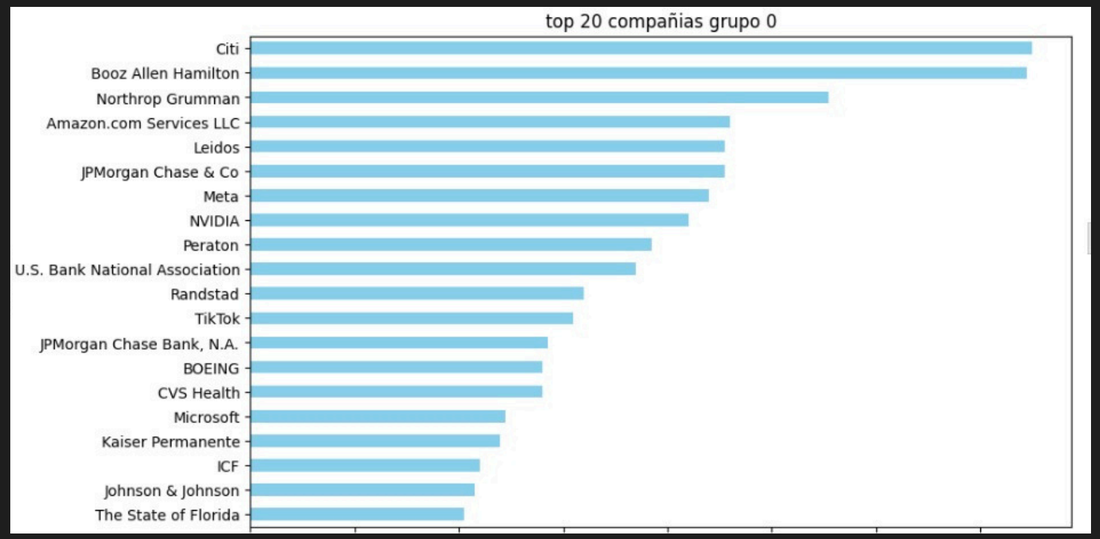
Tipo de trabajo

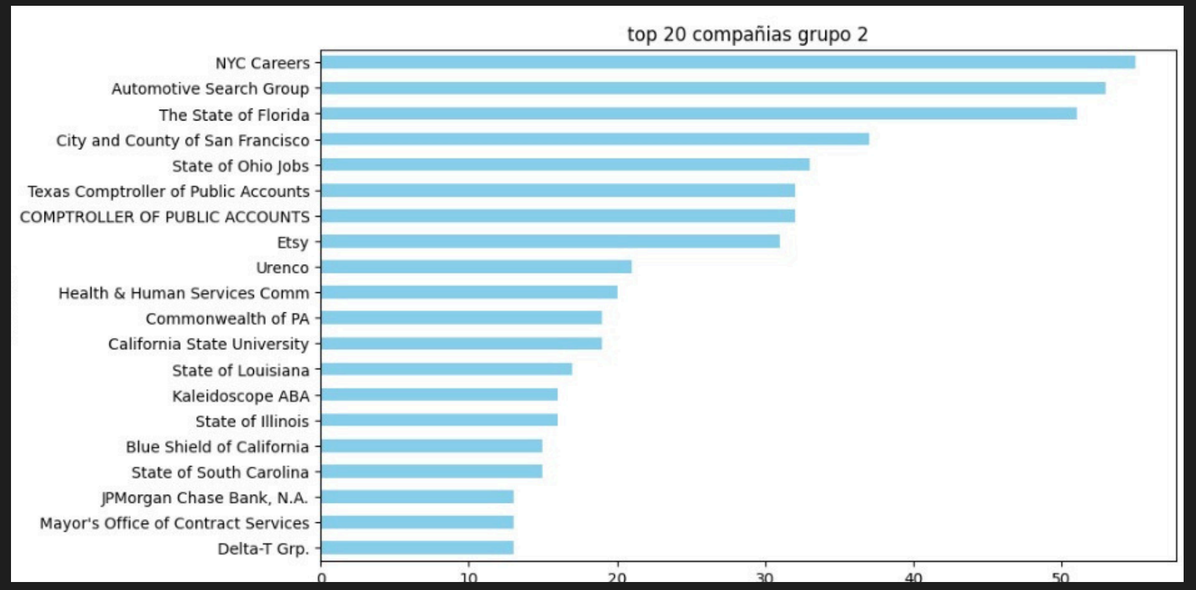


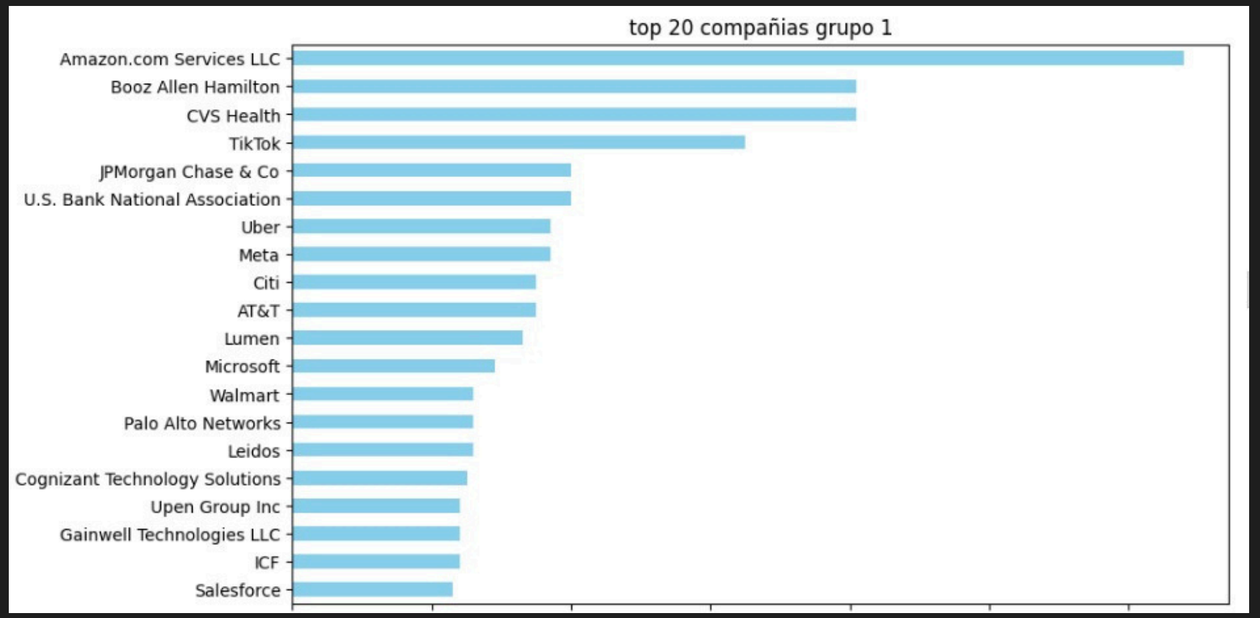




Empresas


top 20 compañias grupo 0
Citi
Booz Allen Hamilton
Northrop Grumman
Amazon.com Services LLC
Leidos
JPMorgan Chase & Co
Meta
NVIDIA
Peraton
U.S. Bank National Association
Randstad
TikTok
JPMorgan Chase Bank, N.A.
BOEING
CVS Health
Microsoft
Kaiser Permanente
ICF
Johnson & Johnson
The State of Florida


top 20 compañias grupo 2
NYC Careers
Automotive Search Group
The State of Florida
City and County of San Francisco
State of Ohio Jobs
Texas Comptroller of Public Accounts
COMPTROLLER OF PUBLIC ACCOUNTS
Etsy
Urenco
Health & Human Services Comm
Commonwealth of PA
California State University
State of Louisiana
Kaleidoscope ABA
State of Illinois
Blue Shield of California
State of South Carolina
JPMorgan Chase Bank, N.A.
Mayor's Office of Contract Services
Delta-T Grp.
0
10
20
30
40
50

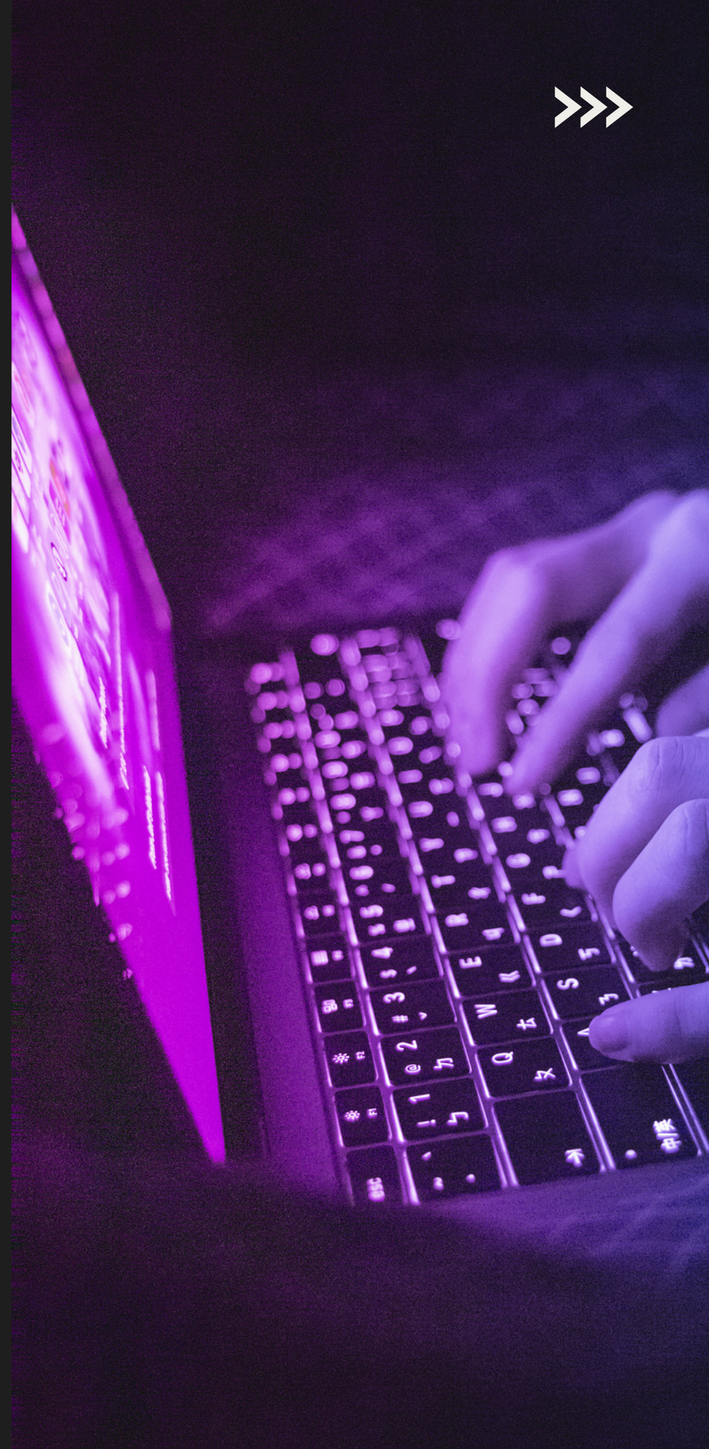

top 20 compañias grupo 1
Amazon.com Services LLC
Booz Allen Hamilton
CVS Health
TikTok
JPMorgan Chase & Co
U.S. Bank National Association
Uber
Meta
Citi
AT&T
Lumen
Microsoft
Walmart
Palo Alto Networks
Leidos
Cognizant Technology Solutions
Upen Group Inc
Gainwell Technologies LLC
ICF
Salesforce

MODELO DE PREDICCIÓN

Modelos utilizados

Para realizar este proyecto, se hicieron las predicciones con varios diferentes modelos, los cuales fueron:

- Random Forest
- Decision Tree
- AdaBoost
- Gradient Boosting
- Linear Regression
- XGBoost
- SVM



Resutados

Después de implementar dichos modelos, se concluyó que el mejor modelo para este problema es el Random Forest, aunque la regresión lineal tuvo rendimiento similar.

```
Mejor modelo: Random Forest
Métricas:
RMSE: 28692.82
R2 Score: 0.55
```

```
Random Forest:
RMSE: 28692.82
R2 Score: 0.55

Decision Tree:
RMSE: 35982.70
R2 Score: 0.29

AdaBoost:
RMSE: 40026.16
R2 Score: 0.12

Gradient Boosting:
RMSE: 31792.15
R2 Score: 0.44
```

```
Linear Regression:
RMSE: 29701.97
R2 Score: 0.51

XGBoost:
RMSE: 31841.47
R2 Score: 0.44

SVM:
RMSE: 38789.30
R2 Score: 0.17
```

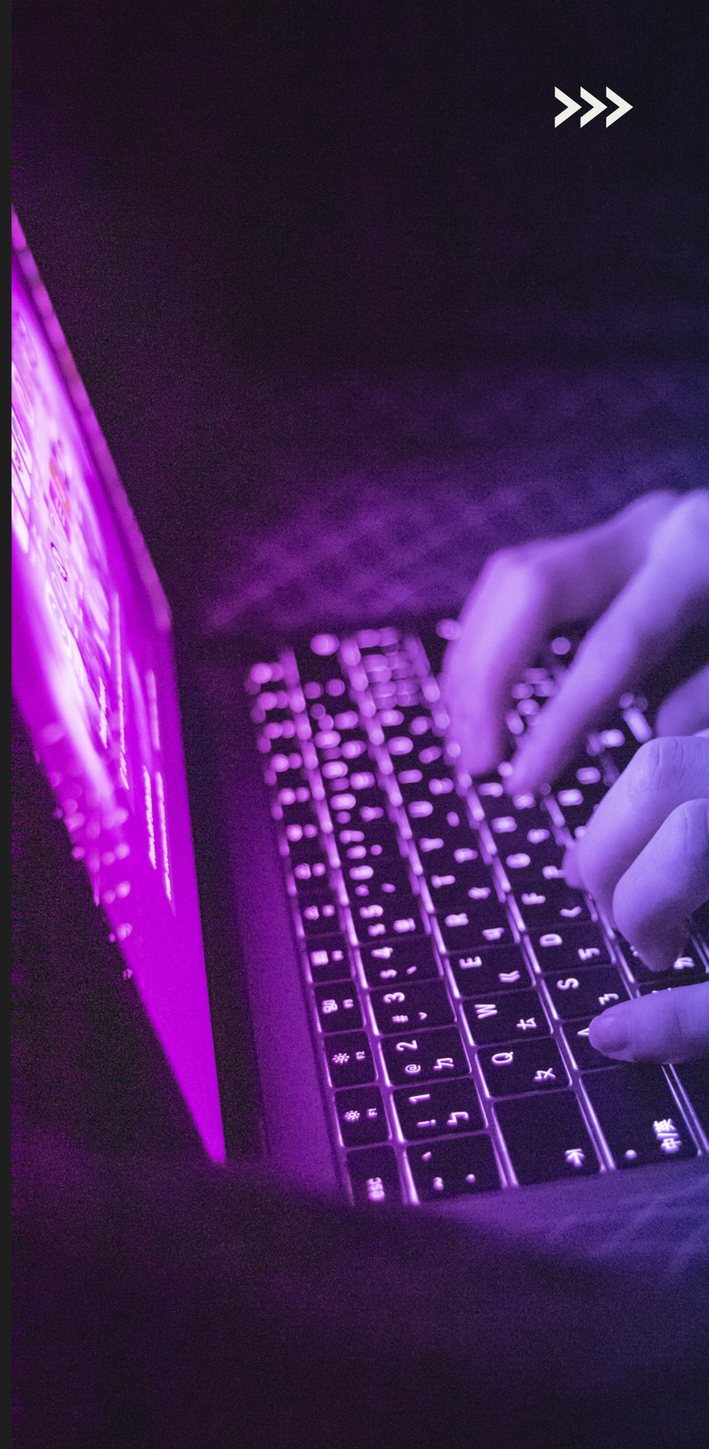
Adicionalmente, se quiso analizar si los resultados que se estaban prediciendo se encontraban en un rango aceptable para cada grupo de trabajo, por lo que se tomó la media de cada grupo y se hizo un rango con los cuartiles 1 y 3 para comparar si las predicciones estaban muy alejadas en un panorama más general y se obtuvieron resultados positivos.

Matriz de conteo:

Jobs_Group	Within_Range	Out_of_Range	Lower_Salary	Upper_Salary	Percentage
Controller	388	124	85000.00	127500.00	75.78
Financial Analyst	442	230	71241.75	103500.00	65.77
Business Intelligence	144	52	87045.25	130000.00	73.47
Business Analyst	342	176	75000.00	110000.00	66.02
Data Engineer	238	106	108503.50	160000.00	69.19
Analyst	139	69	67500.00	98777.81	66.83
ML/AI Engineer	84	30	136000.00	203625.00	73.68
Data Analyst	271	137	71000.00	112500.00	66.42
CFO	202	61	101362.50	175000.00	76.81
Data Scientist	172	100	112500.88	171050.25	63.24
Others	64	27	85000.00	137401.50	70.33
Operations Analyst	89	40	68363.00	107500.00	68.99
Finance	189	87	72325.00	110000.00	68.48
Statistician/Mathemathics	28	17	80000.00	134500.00	62.22

Modelos utilizados

Debido a que los modelos funcinaban correctamente, se consideró que la presición aún era muy baja y se buscó una implementación más potente. Por lo que se optó por un modelo stacked que tomó en cuenta el random forest y la regresión lineal



Resutados

Los resultados de las predicciones de este modelo mejoraron los modelos anteriores, obteniendo un R^2 score de 0.6

```
Modelo Stacked:
RMSE: 26803.81
R2 Score: 0.60
```

CONCLUSIONES Y RECOMENDACIOENS

Conclusiones

- ES POSIBLE HACER PREDICCIONES DE SALARIOS SEGÚN DIFERENTES CARACTERÍSTICAS DE LA PERSONA
- EN ESTADOS UNIDOS, HAY UNA CORRELACIÓN FUERTE ENTRE LA LOCACIÓN Y EL SALARIO

Recomendaciones

- UTILIZAR MAYOR CANTIDAD DE DATOS PARA PODER MEJORAR LA PRESICIÓN DE LOS MODELOS
- UTILIZAR UNA BASE DE DATOS CON MENOS VALORES FALTANTES, YA QUE AUNQUE SE MANEJEN ESOS CASOS DE FORMA ESPECIAL, PUEDEN TENER INCIDENCIA EN EL RESULTADO FINAL DE LOS MODELOS

MUCHAS GRACIAS