



# MACHINE LEARNING

IRONHACK



# ÍNDICE



1

**Descripción**  
*explicación  
proyecto*

5

**Conclusión**  
*Breve descripción  
aquí*

2

**Organización**  
*entender los datos*

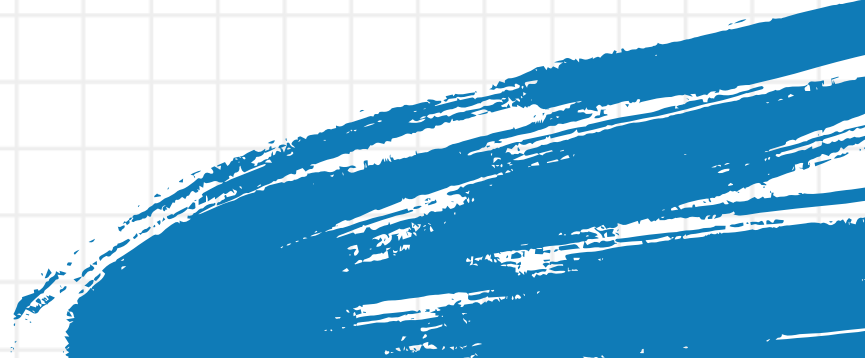
3

**Limpieza**  
*Procesar el  
dataset*



4

**Elección modelo**  
*modelo machine  
learning*



# INFORMACIÓN IMPORTANTE

## DESCRIPCIÓN

EL proyecto trata de hacer un modelo predictivo en un dataset de Salaries, entrenar diferentes modelos y buscar el más óptimo y que mejor explique la variable salary

## ORGANIZACIÓN

Realizo un estudio de los outliers para entender un poco mejor el dataset. También analizo los máximos y mínimos. Realizo la media de los salarios y a través de los Cuartiles establezco unos límites para los valores.

## LIMPIEZA

las columnas "Salary" y "Salary Currency". Estas columnas se encuentran recogidas en "Salary\_in\_usd". Por lo que decidí eliminarlas.

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
0	2022	SE	FT	Data Engineer	140250	US	100	US	M
1	2022	SE	FT	Data Engineer	135000	US	100	US	M
2	2021	MI	FT	BI Data Analyst	100000	US	100	US	M
3	2021	MI	CT	ML Engineer	270000	US	100	US	L
4	2021	MI	FT	Data Engineer	26005	RO	0	US	L

```
df_test.head()
```

	work_year	experience_level	employment_type	job_title	employee_residence	remote_ratio	company_location	company_size
0	2020	SE	FT	Machine Learning Scientist	JP	0	JP	S
1	2020	MI	FT	Lead Data Analyst	US	100	US	L
2	2020	MI	FT	Data Analyst	US	100	US	L
3	2020	MI	FT	Machine Learning Engineer	CN	0	CN	M
4	2020	MI	FT	Product Data Analyst	IN	100	IN	L

```
#Límites del criterio turkey para identificar outliers
```

```
Q1 = df['salary_in_usd'].quantile(0.25)
```

```
Q3 = df['salary_in_usd'].quantile(0.75)
```

```
IQR = Q3 - Q1
```

```
outliers = df.groupby('job_title')['salary_in_usd'].mean()[df.groupby('job_title')['salary_in_usd'].mea
```

```
outliers
```

```
job_title
```

```
Data Analytics Lead      405000.00
```

```
Financial Data Analyst  450000.00
```

```
Name: salary_in_usd, dtype: float64
```

```
Q11: 64594.5
```

```
Q33: 150000.0
```

```
IQRR: 85405.5
```

```
Upper Limit for Outliers: 278108.25
```

```
Lower Limit for Outliers: 21891.75
```

# OBJETIVOS DE LA PRESENTACIÓN



ELEGIR MODELO



ENTRENAR



TESTEAR

Model	Adjusted R-Squared
Lars	698168442509481163906183114995979575195860992.00
RANSACRegressor	55597888842277934003424067584.00
TransformedTargetRegressor	3175467926990929956649828352.00
LinearRegression	3175467926990929956649828352.00
MLPRegressor	37.64
LinearSVR	37.56
KernelRidge	35.31
GaussianProcessRegressor	14.75
SVR	10.15
QuantileRegressor	10.14
DummyRegressor	10.04
NuSVR	10.01
ElasticNetCV	9.45
AdaBoostRegressor	8.56
KNeighborsRegressor	8.32
Lasso	7.12

