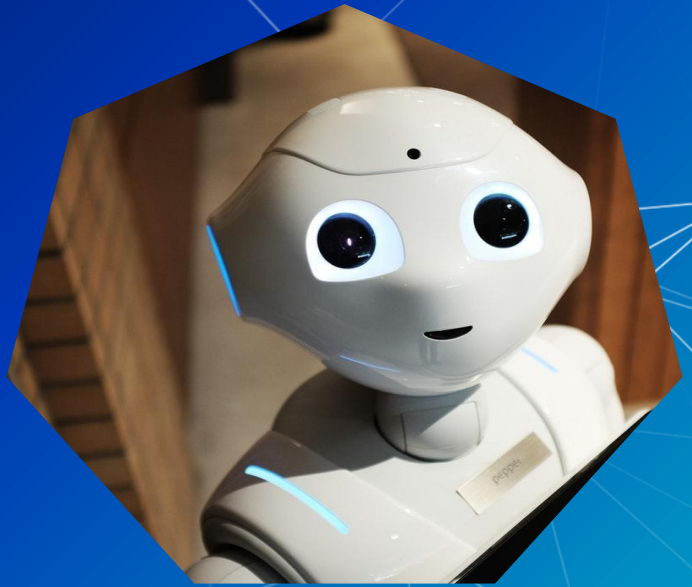
MACHINE LEARNING



TRABAJO FINAL

INTEGRANTES

- Fátima Gómez
- Beatriz Meléndez
- Alfonso Merayo
- Lisbeth Carrillo
- David Mayo

Objetivo

Buscamos comparar la eficacia de los algoritmos de regresión, árboles de decisión y Clustering en la predicción de los salarios del sector tech (India).

Variables: Años de experiencia, la titulación, el tipo de trabajo y brecha salarial entre géneros.

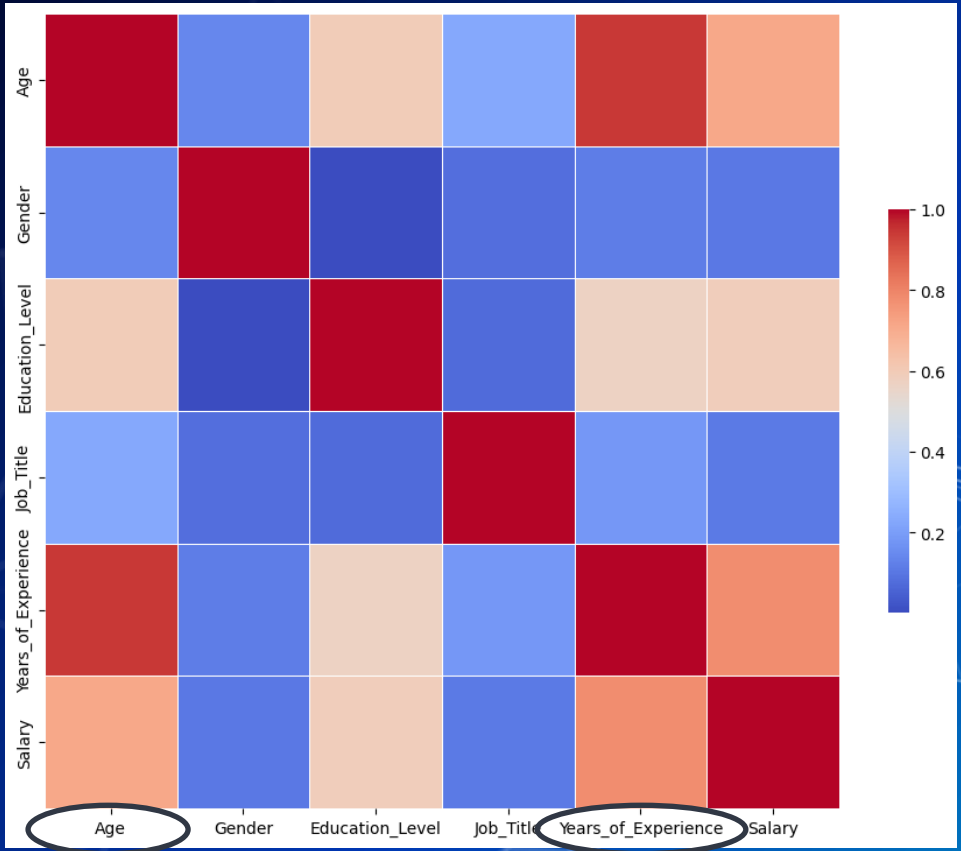
1. **Regresión: nos ayudaría a predecir como aumenta el salario en función de los años de experiencia de la persona.**
2. **Árboles de decisión: podríamos utilizar un árbol de decisión para predecir el salario en función de varias variables.**
3. **Clustering: podríamos utilizar el clustering para identificar los diferentes tipos de trabajadoes**

Data

```
df_salary = pd.read_csv('Salary_Data.csv')
```

	Age	Gender	Education_Level	Job_Title	Years_of_Experience	Salary
Age	1.000000	0.140344	0.595318	0.230264	0.944147	0.714632
Gender	0.140344	1.000000	0.000420	0.080926	0.115929	0.102250
Education_Level	0.595318	0.000420	1.000000	0.072884	0.566691	0.588323
Job_Title	0.230264	0.080926	0.072884	1.000000	0.182468	0.107738
Years_of_Experience	0.944147	0.115929	0.566691	0.182468	1.000000	0.777988
Salary	0.714632	0.102250	0.588323	0.107738	0.777988	1.000000

1.
MODELO DE REGRESIÓN



En este mapa de calor podemos ver como las variables se correlacionan con nuestro objetivo.

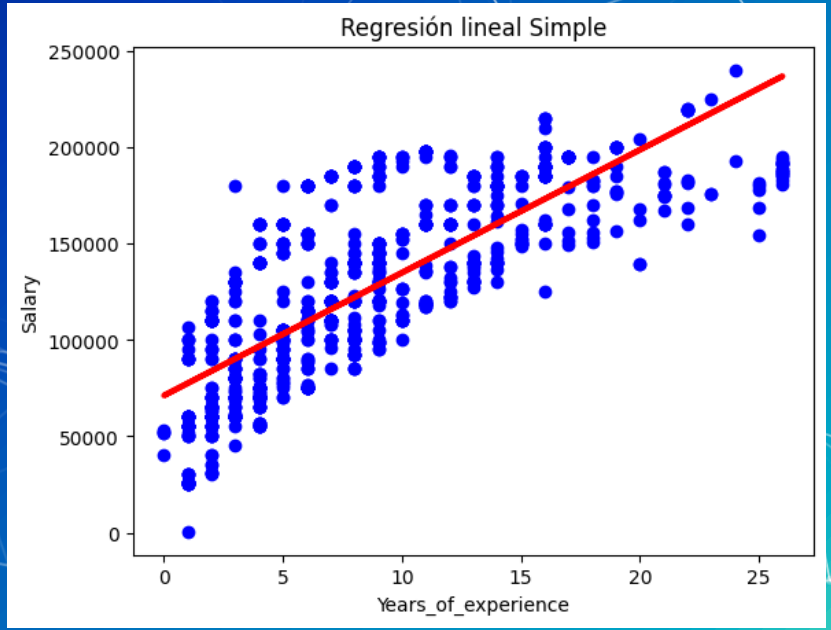
Las que mayor se correlacionan son:

-“Years_of_experience”
- “Age”

Regresión lineal simple

- progresión del salario = 6371.807028123424 Years_of_experience + 71131.80882175866

- Raíz error cuadrático medio: 30609.13
  Coeficiente de determinación: 0.60



CONCLUSIÓN:

Un error cuadrático medio de $ 30609.13 es elevado, lo que sugiere que hay una cantidad considerable de error en las predicciones. El coeficiente de determinación explica el 60% de la variación en los datos de los salarios, todavía hay un 40% de la variación en los salarios que no puede ser explicada.

Es probable que las variables "Age", "Education_Level", "Gender", y "Job_Title" también influyan en el salario de una persona.

Regresión múltiple con dos variables ('Age' y 'Years_of_Experience')

- progresión del salario = 7680.247480905887 Years_of_Experience + -1096.9003613167804 Age + 97253.86736727759

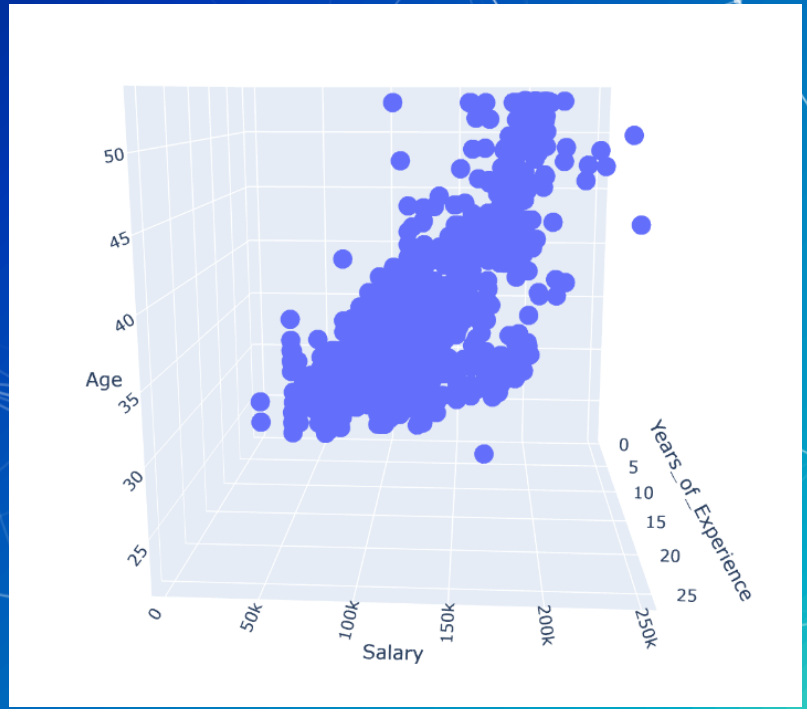
- Raíz error cuadrático medio: 30458.77
  Coeficiente de determinación: 0.60

CONCLUSIÓN:

A la vista del error cuadrático medio, observamos una ligera mejora en comparación con el RMSE del modelo anterior de $30,609.13, pero sigue siendo una cantidad considerable de error en las predicciones.

El modelo aún explica el 60% de la variación en los datos de los salarios. Esto no ha cambiado con respecto al modelo anterior, lo que indica que añadir "Age" como variable independiente no ha aumentado significativamente la cantidad de variación que el modelo puede explicar. Sugiere puede no estar fuertemente correlacionada con "Salary", o que cualquier correlación que exista ya está en gran medida capturada por la variable "Years_of_Experience".

En este punto, consideramos agregar otras variables independientes al modelo que podrían tener una correlación más fuerte con "Salary", como "Education_Level"



Regresión múltiple con dos variables ('Years_of_Experience','Education_Level')

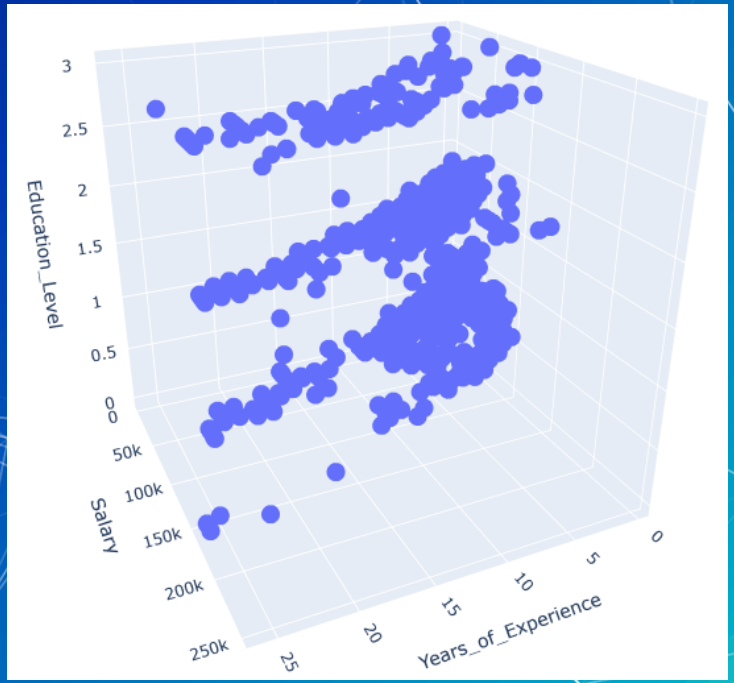
- progresión del salario = 5385.629286008472 Years_of_Experience + 12972.938620489704 Education_Level + 58194.966718827156

- Raiz error cuadrático medio: 29512.85
- Coeficiente de determinación: 0.62

CONCLUSIÓN:

En este nuevo modelo podemos ver que el resultado ha variado un mínimo, En este caso, el modelo continúa con precisión moderada pero aumentó al 62% y bajo su medida de dispersión a 29512.85
Quiere decir que las variables si se correlacionan positivamente.

Este es el modelo que menor error cuadrático medio presenta, pero el coeficiente de determinación sólo ha mejorado un 2%. Vamos a probar con un modelo de regresión múltiple



Regresión múltiple con tres variables

- progresión del salario = 7617.136773435293 Years_of_Experience + -1973.738034304606 Age + 14589.278906156467 Education_Level + 103586.57308314556

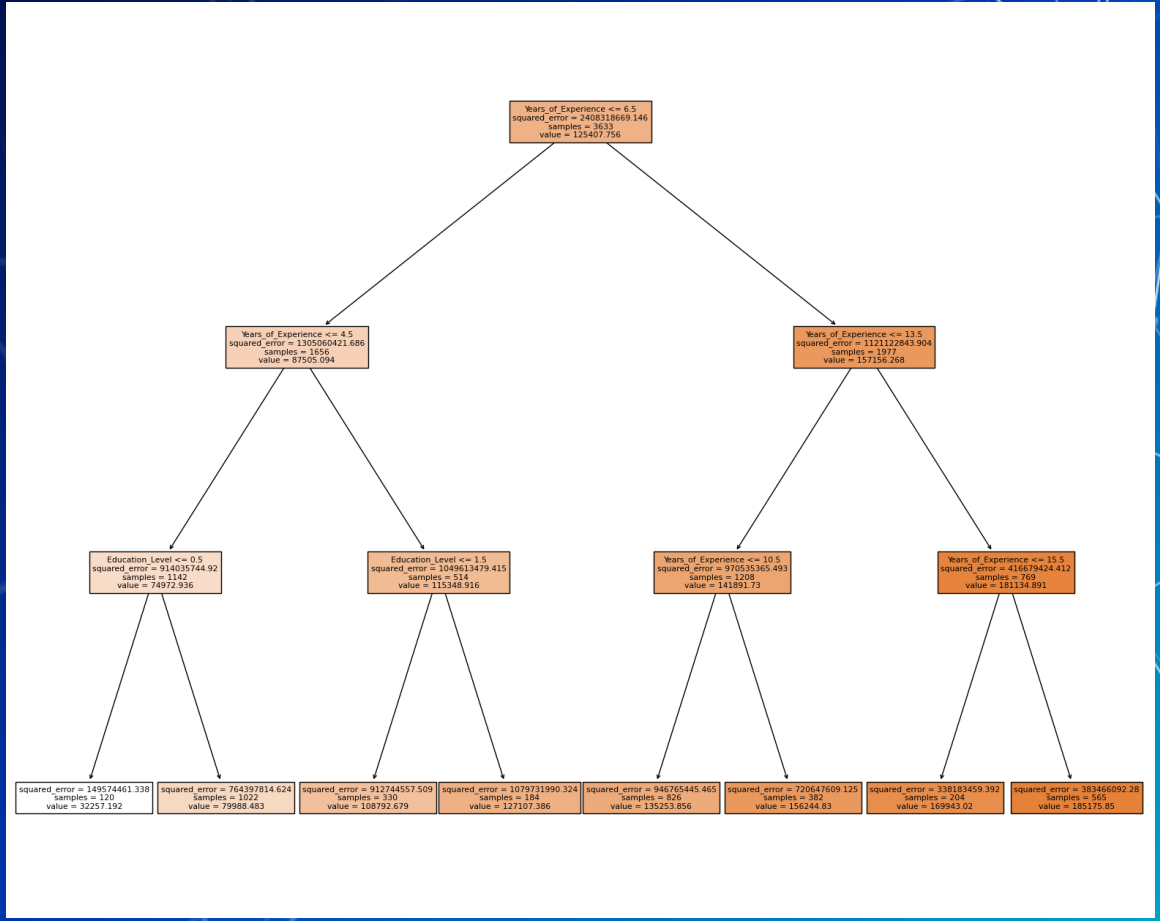
- Raiz error cuadrático medio: 29105.77
- Coeficiente de determinación: 0.64

CONCLUSIÓN:

En este nuevo modelo podemos ver que el resultado ha mejorado en cuanto a dispersión. En este caso, el modelo continúa con precisión moderada al 64% y bajo un poco mas su medida de dispersion hasta 29105.77. Es decir, este modelo tiene un menor número de variabilidad de los errores.

En resumen, este nuevo modelo con "Age", "Education_level" y "Years_of_Experience" como variables independientes es un modelo mejor que los anteriores basados en las métricas de raíz del error cuadrático medio y coeficiente de determinación.
Sin embargo, como el coeficiente de determinación es todavía 0.64, hay un 36% de la variación en los salarios que aún no se explica con este modelo.
Vamos a ver que ocurre con modelos de árboles de decisión, porque puede que haya relaciones no lineales

2.
ÁRBOLES DE DECISIÓN

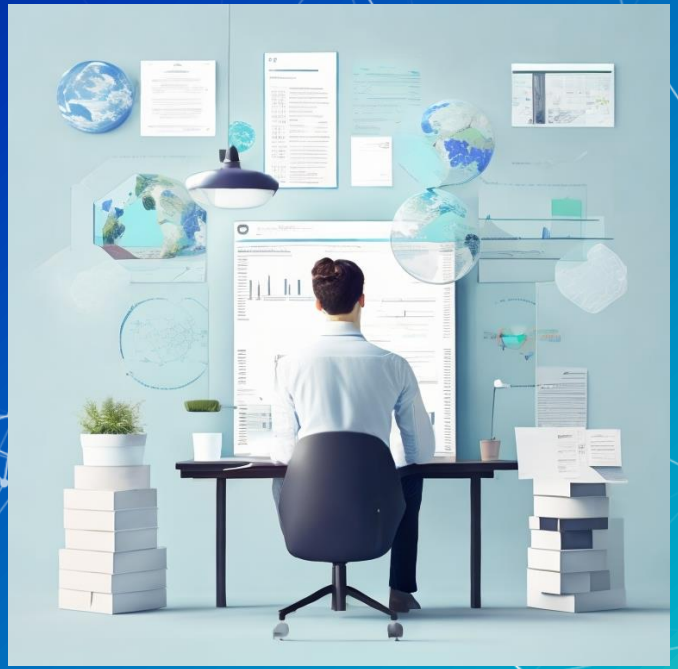

Years_of_Experience <= 6.5
squared_error = 2408318669.146
samples = 3633
value = 125407.756
Years_of_Experience <= 4.5
squared_error = 1305060421.686
samples = 1656
value = 87505.094
Years_of_Experience <= 13.5
squared_error = 1121122843.904
samples = 1977
value = 157156.268
Education_Level <= 0.5
squared_error = 914035744.92
samples = 1142
value = 74972.936
Education_Level <= 1.5
squared_error = 1049613479.415
samples = 514
value = 115348.916
Years_of_Experience <= 10.5
squared_error = 970535365.493
samples = 1208
value = 141891.73
Years_of_Experience <= 15.5
squared_error = 416679424.412
samples = 769
value = 181134.891
squared_error = 149574461.338
samples = 120
value = 32257.192
squared_error = 764397814.624
samples = 1022
value = 79988.483
squared_error = 912744557.509
samples = 330
value = 108792.679
squared_error = 1079731990.324
samples = 184
value = 127107.386
squared_error = 946765445.465
samples = 826
value = 135253.856
squared_error = 720647609.125
samples = 382
value = 156244.83
squared_error = 338183459.392
samples = 204
value = 169943.02
squared_error = 383466092.28
samples = 565
value = 185175.85

Raíz error cuadrático medio: 27747.84
Coeficiente de determinación: 0.69

CONCLUSIÓN:

Podemos ver que este modelo es mucho más acertado que la regresión múltiple , ya que aumenta la precisión a un 69% y la dispersión es mucho menor.

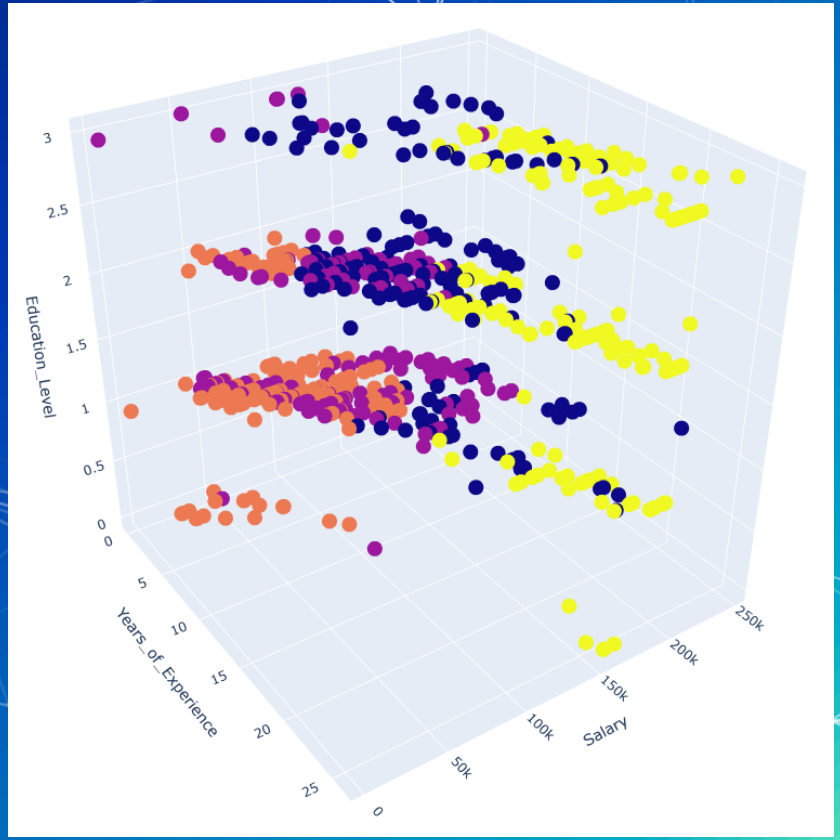
Comparado con el error cuadrático medio anterior de $29105.77, las predicciones han mejorado.
Según el coeficiente de determinación. modelo ahora explica el 69% de la variación en los salarios. Esto es una mejora con respecto al valor anterior de 0.64.
Estos resultados sugieren que el modelo de árbol de decisión es una mejor elección para los datos que el modelo de regresión lineal.



3.
CLUSTERING

Observamos que hay cuatro grupos:

- Naranjas: Menor nivel de educación, suelen tener menos años de experiencia y estudios no superiores a Bachelor o grado universitario
- Morados: Mayor nivel de educación, han estudiado en su mayoría Bachelor o grado universitario
- Azules: Mayor nivel de educación, han estudiado hasta un master y tienen más salario
- Amarillos: Tienen un salario más alto y han estudiado Bachelor, algunos también un master y otros incluso doctorado. Aquí el nivel de educación es más indiferente y pesan más los años de experiencia. No obstante, eso sí, el mayor número de doctorados se encuentra en este grupo.



Conclusiones

Después de aplicar el método de árboles de decisión pudimos ver que teníamos una predicción mucho mas efectiva que una regresión y esto sería porque hay una interacción mas compleja entre variables.

Al aplicar Clustering, nos hemos dado cuenta que es útil cuando se busca explorar, resumir y agrupar datos sin tener una variable objetivo específica y nos ayudar a descubrir estructuras y patrones ocultos en los datos sin la necesidad de una variable objetivo conocida

A través de este estudio, nos damos cuenta que la variable más importante para determinar el salario son los años de experiencia.


THANKS!