Trabajo Práctico Numero 4: Regresión y predicción.

Costanzo Matías Adriel.

Grupo 9.

Big Data UBA- Noelia Romero

Año 2004	Media Train	Media Test	Diferencia
const	1	1	0
Edad2	31	31	0
Edad al cuadrado	1448	1474	25
Educ	10	10	0
Mujer(sexo)	0.5	0.5	0
Horas trab.	15.1	14.9	0.2

Comentario
Las diferencias de medias entre los conjuntos de entrenamiento y prueba son pequeñas, lo que sugiere que la partición aleatoria mantiene representatividad. Esto es importante para evitar sesgos al entrenar el modelo.

Año 2024	Media Train	Media Test	Diferencia
const	1	1	0
Edad2	36	36	0
Edad al cuadrado	1810	1810	0
Educ	10	10	0
Mujer(sexo)	0.5	0.5	0
Horas trab.	36	37	1

Comentario
Las diferencias de medias entre los conjuntos de entrenamiento y prueba son pequeñas, lo que sugiere que la partición aleatoria mantiene representatividad. Esto es importante para evitar sesgos al entrenar el modelo.
Luego de esto, empezamos con nuestra prueba de modelos con variables, agregando 1 a la vez

Var dep: Salario semanal	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Variables					
Edad	0.01	0.0210			
Edad 2	-				
Educacion	-				
Mujer	-				
horas trab	-				
N					

3) Obtenemos los siguientes resultados de 2004 y 2024

```
====== Año 2004 ======
                            OLS Regression Results
==============================================================================
Dep. Variable:        salario_semanal   R-squared:                       0.052
Model:                            OLS   Adj. R-squared:                  0.051
Method:                 Least Squares   F-statistic:                     [illegible]8.4
Date:                Mon, 02 Jun 2025   Prob (F-statistic):           1.64e-98
Time:                        23:32:19   Log-Likelihood:                -887[illegible].
No. Observations:                8679   AIC:                         1.774e+05
Df Residuals:                    8674   BIC:                         1.775e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1533.3089    255.033      6.012      0.000    1033.383    2033.234
edad2         73.6837      5.653     13.034      0.000      62.602      84.765
educ          81.8710      7.452     10.986      0.000      67.263      96.479
mujer      -1876.1729    144.428    -12.990      0.000   -2159.285   -1593.061
horastrab      3.6590      1.205      3.036      0.002       1.297       6.021
==============================================================================
Omnibus:                    22731.356   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       [illegible]78795730.398
Skew:                          30.095   Prob(JB):                         0.00
Kurtosis:                    1807.466   Cond. No.                         277.
==============================================================================
```

```
====== Año 2024 ======
                            OLS Regression Results
==============================================================================
Dep. Variable:        salario_semanal   R-squared:                       0.024
Model:                            OLS   Adj. R-squared:                  0.023
Method:                 Least Squares   F-statistic:                     72.57
Date:                Mon, 02 Jun 2025   Prob (F-statistic):           7.42e-61
Time:                        23:32:21   Log-Likelihood:            -1.2474e+05
No. Observations:               12020   AIC:                         2.495e+05
Df Residuals:                   12015   BIC:                         2.495e+05
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       4129.3330    261.595     15.785      0.000    3616.565    4642.101
edad2         64.1748      5.554     [illegible].554      0.000      53.288      75.062
educ          23.1502      6.771      3.419      0.001       9.878      36.423
mujer      -1678.6890    143.053     -[illegible].735      0.000   -1959.096   -1398.282
horastrab      3.2193      1.506      2.137      0.033       0.266       6.172
==============================================================================
Omnibus:                    17679.006   Durbin-Watson:                   2.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         18258446.006
Skew:                           8.542   Prob(JB):                         0.00
Kurtosis:                     193.169   Cond. No.                         251.
==============================================================================
```

1. R^2 y calidad del modelo

- En ambos años, el **R^2 es muy bajo** (5.2% en 2004 y 2.4% en 2024), indicando que estas variables explican una pequeña parte de la variación en el salario semanal. Esto sugiere que hay muchos otros factores importantes no incluidos en el modelo.

- El R^2 bajó a la mitad en 2024 respecto a 2004, lo que puede indicar que la relación entre estas variables y el salario se ha debilitado o que la heterogeneidad salarial se explica menos con estos factores básicos en 2024.

2. Intercepto

- El intercepto aumentó mucho en 2024 (4,129) comparado con 2004 (1,533). Esto puede reflejar un aumento general en el nivel de salarios o inflación que no está capturada por las otras variables, dado que el intercepto representa el salario base estimado.

3. Edad al cuadrado (edad2)

- El coeficiente es positivo y significativo en ambos años, indicando que a medida que la edad aumenta, el salario semanal aumenta de manera no lineal (con aceleración positiva).

- El efecto es algo menor en 2024 (64.17 vs 73.68), pero sigue siendo relevante.

4. Educación (educ)

- En 2004, el coeficiente de educación es bastante alto (81.87), lo que indica que cada año adicional de educación aumenta el salario semanal en aproximadamente $81.87.

- En 2024, este efecto se reduce mucho a $23.15 por año de educación. Esto podría interpretarse como una disminución del "retorno" salarial a la educación, o que otros factores complementarios están ganando más peso.

- En ambos casos, el coeficiente es significativo.

5. Género (mujer)

- En ambos años, ser mujer está asociado a una reducción significativa del salario semanal, con coeficientes negativos grandes (-1,876 en 2004 y -1,679 en 2024).

- Esto refleja la persistencia de una brecha salarial de género considerable, aunque la magnitud parece haberse reducido un poco en 2024.

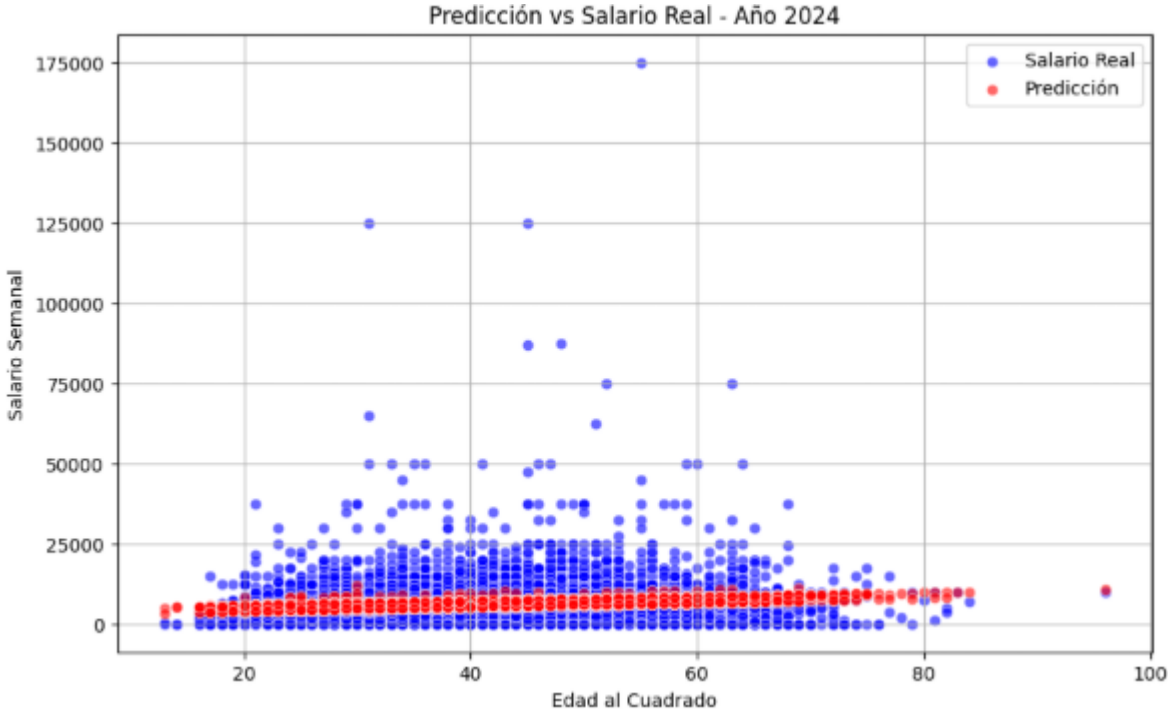
- El efecto es altamente significativo ($p<0.001$).

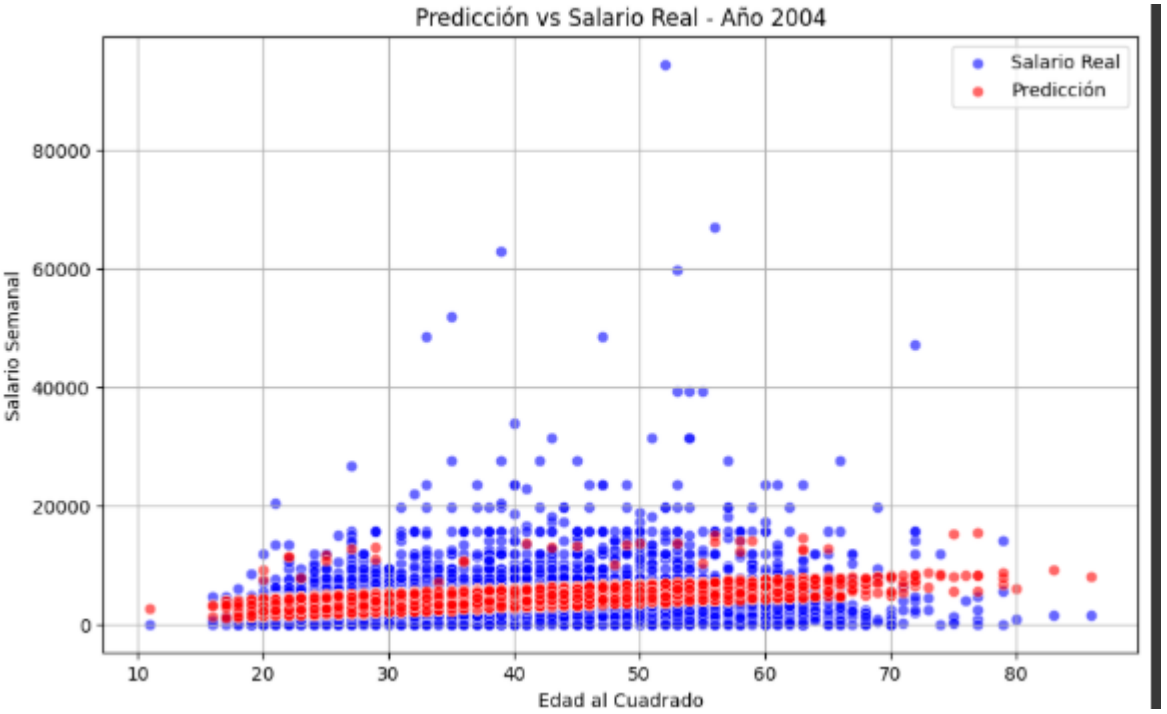
6. Horas trabajadas (horastrab)

- En ambos años, más horas trabajadas se asocian con un aumento pequeño pero significativo del salario semanal ($3.66 en 2004 y $3.22 en 2024 por hora extra trabajada).

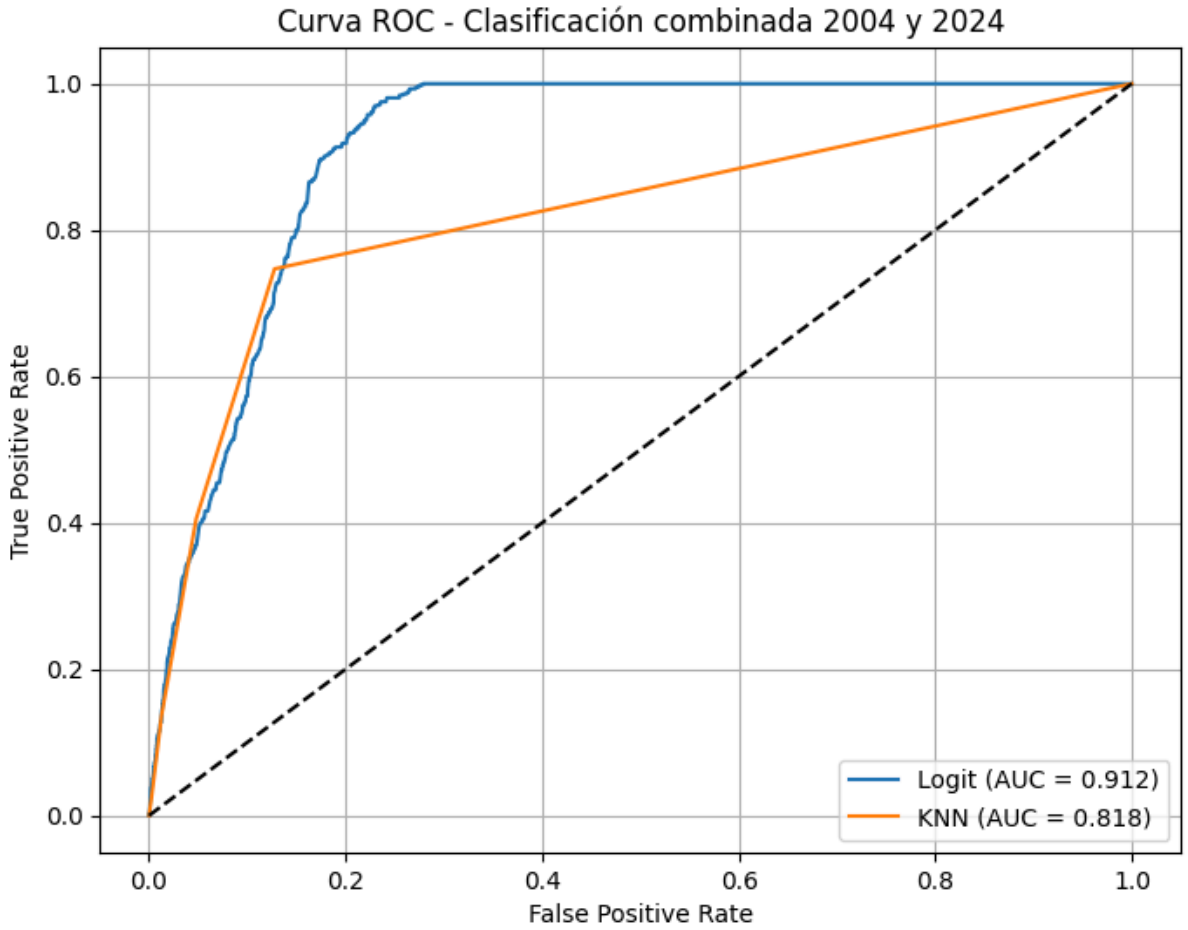
- El efecto se mantiene similar, ligeramente menor en 2024.

Podemos así, ver ambos graficos de dispersión:





5) Podemos ver el gráfico



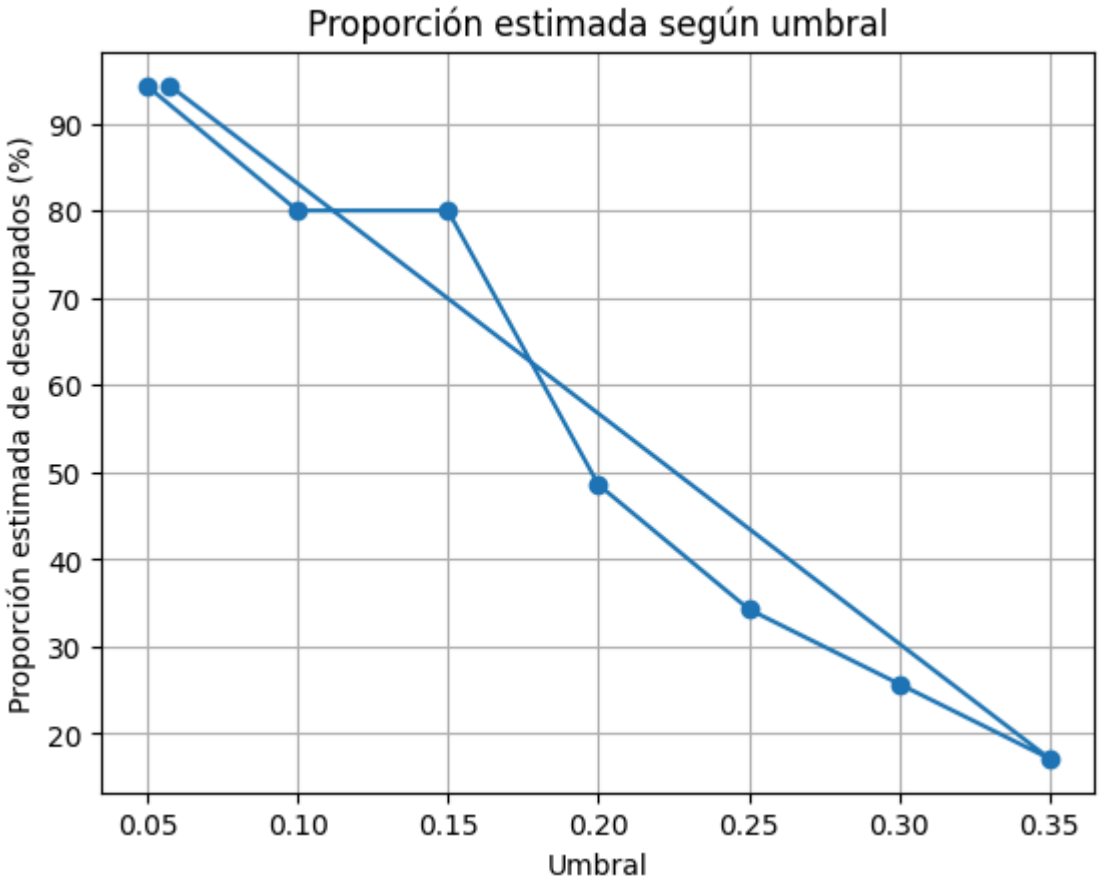
Conclusión:
La regresión logística tiene mejor AUC y accuracy global, pero no identifica ningún desocupado en el test, lo cual es un problema serio para un modelo de clasificación binaria con clases desbalanceadas.

El KNN tiene menor AUC y accuracy, pero sí logra identificar algunos desocupados, lo que la hace más útil para detectar esa clase minoritaria, aunque con errores.

- SI el objetivo es maximizar la capacidad de detección de desocupados, KNN es mejor porque predice esta clase.

- Si lo que importa es una alta exactitud general y buen ordenamiento según la probabilidad, la regresión logística tiene mejor performance, pero hay que ajustar el umbral o balancear clases para que también detecte desocupados.

Por ultimo, lo que pretendemos entonces, es tomar el KNN, ya que logra identificar desocupados, y preferimos maximizar la capacidad de detección, para poder tomar en cuenta factores como desarrollo. Y lo aplicamos a nuestro modelo



Lo que muestran tus datos
El umbral óptimo según Youden es 0.058 — muy bajo, lejos del clásico 0.5.

Eso indica que, para maximizar la capacidad de detectar desocupados (sensibilidad), el modelo tiene que clasificar como desocupado a casi todos con una probabilidad mayor a 0.058.

Es decir, es muy “generoso” para asignar la clase desocupado.

La tabla

Umbral	% estimado de desocupados (proporción predicha)
0.05	94.29%
0.10	80.00%
0.15	80.00%
0.20	48.57%
0.25	34.29%
0.30	25.71%
0.35	17.14%

0.058 94.29%

Al bajar el umbral a 0.05 (o cercano 0.058), muchísimas personas son clasificadas como desocupadas (94.29%).

A medida que sube el umbral, el porcentaje de clasificados como desocupados baja.

El umbral que maximiza Youden es justamente el que marca el punto con alto porcentaje de detección (sensibilidad alta), aunque a costa de clasificar erróneamente muchos casos como desocupados (falsos positivos).