**BIG DATA & MACHINE LEARNING**

**Trabajo Práctico N° 4**

**Métodos Supervisados: Regresión & Clasificación usando la EHP**

*Parte A:*

1)

**2004**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Media Train | Media Test | Diferencia |
| CH06 | 33.451076 | 34.050611 | -0.599535 |
| NIVEL\_ED | 2.691113 | 2.661867 | 0.029246 |
| IPCF | 358.227007 | 373.852669 | -15.625662 |
| CH04 | 1.532647 | 1.520942 | 0.011705 |
| CH07 | 3.512067 | 3.456370 | 0.055697 |

**2024**

|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Media Train | Media Test | Diferencia |
| CH06 | 38.205544 | 38.105314 | 0.100231 |
| NIVEL\_ED | 3.832892 | 3.849263 | -0.016371 |
| IPCF | 165539.752107 | 150432.498954 | 15107.253153 |
| CH04 | 1.518647 | 1.533999 | -0.015352 |
| CH07 | 3.431424 | 3.436519 | -0.005095 |

Las variables están bien balanceadas entre train y test, con diferencias mínimas en las medias. Solo el IPCF muestra una brecha mayor, pero no compromete la comparabilidad general de los conjuntos.

*Parte B:*

2)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var Dep: Salario Semanal | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 |
| Variables | 1 | 2 | 3 | 4 | 5 |
| Edad | 1023.953\*\*\* (196.93) | 8286.717\*\*\* (1082.10) | 8534.898\*\*\* (1045.06) | 9276.996\*\*\* (1023.47) | 9487.966\*\*\* (972.59) |
| Edad2 |  | -83.879 \*\*\* (12.29) | -81.887 \*\*\* (11.87) | -89.977 \*\*\* (11.62) | -87.881 \*\*\* (11.00) |
| Educ |  |  | 10470.562 \*\*\* (657.91) | 11381.397 \*\*\* (647.25) | 14622.772 \*\*\* (631.32) |
| Mujer |  |  |  | -65231.176 \*\*\* (5134.86) | -61495.312 \*\*\* (4837.16) |
| CH11 |  |  |  |  | 1996.553 (1617.27) |
| CH09 |  |  |  |  | -104097.032 \*\*\* (5083.11) |
| N(Observaciones) | 3488 | 3488 | 3488 | 3488 | 3488 |
| R2 | 0.008 | 0.021 | 0.087 | 0.128 | 0.228 |

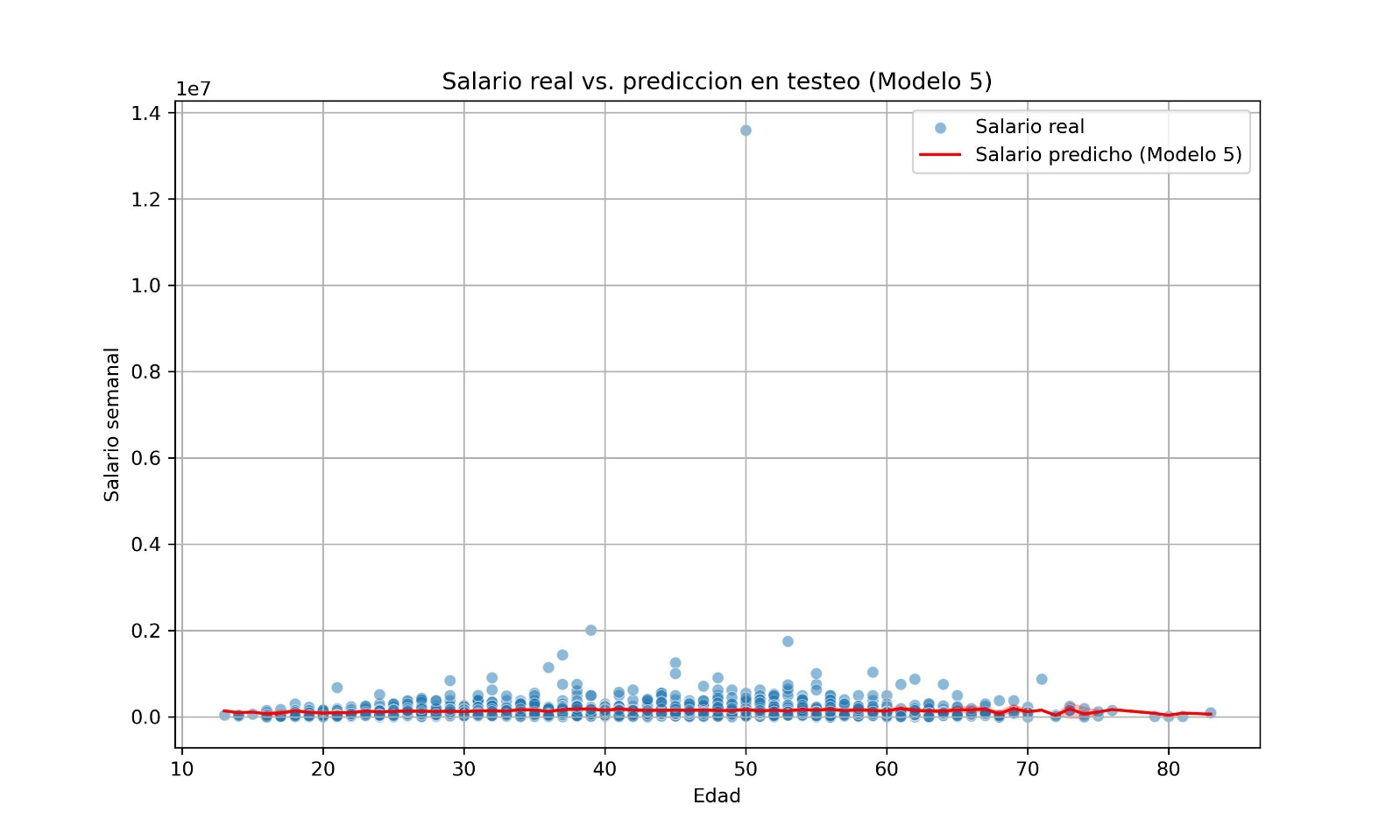
Nota: destaque con \*, \*\*, y \*\*\* cuando el p-valor de los coeficientes reportados sean menor que 0.1, 0.05 y 0.001 respectivamente.

Los modelos muestran que la educación y el género explican gran parte del salario. Al agregar más variables, mejora el R², destacando el efecto positivo de la educación y la penalización salarial a las mujeres.

3)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var Dep: Salario Semanal | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 |
| Variables | 1 | 2 | 3 | 4 | 5 |
| MSE | 450157e+11 | 443898e+11 | 418411e+11 | 406124e+11 | 369515e+11 |
| RMSE | 380809.271280 | 379986.629640 | 376617.979429 | 374983.135157 | 370069.648701 |
| MAE | 104537.118639 | 103508.467708 | 100291.055788 | 98721.840874 | 90692.702781 |

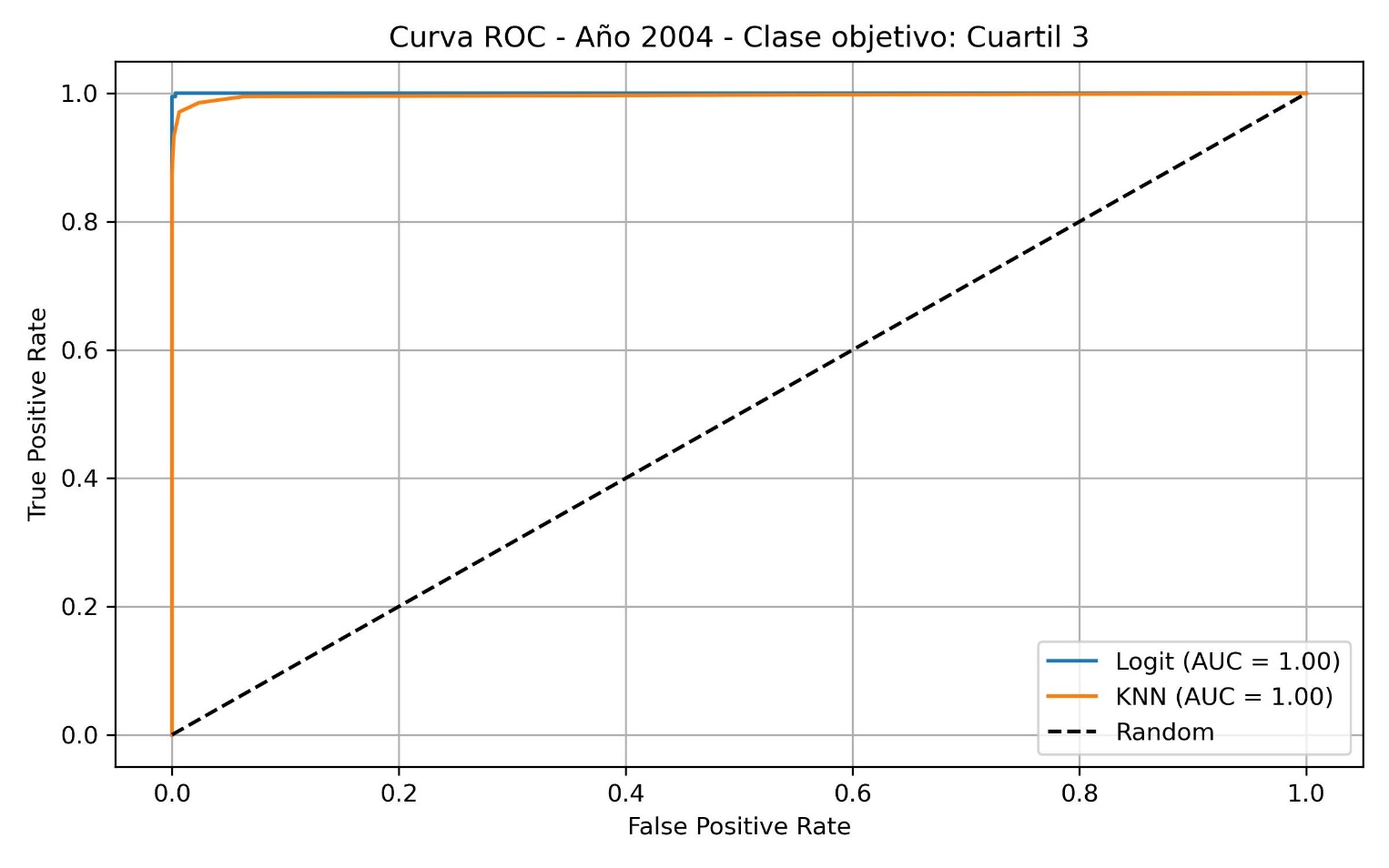
A medida que se agregan variables, el MSE, RMSE y MAE disminuyen, lo que indica que los modelos predicen mejor el salario semanal con más información relevante.

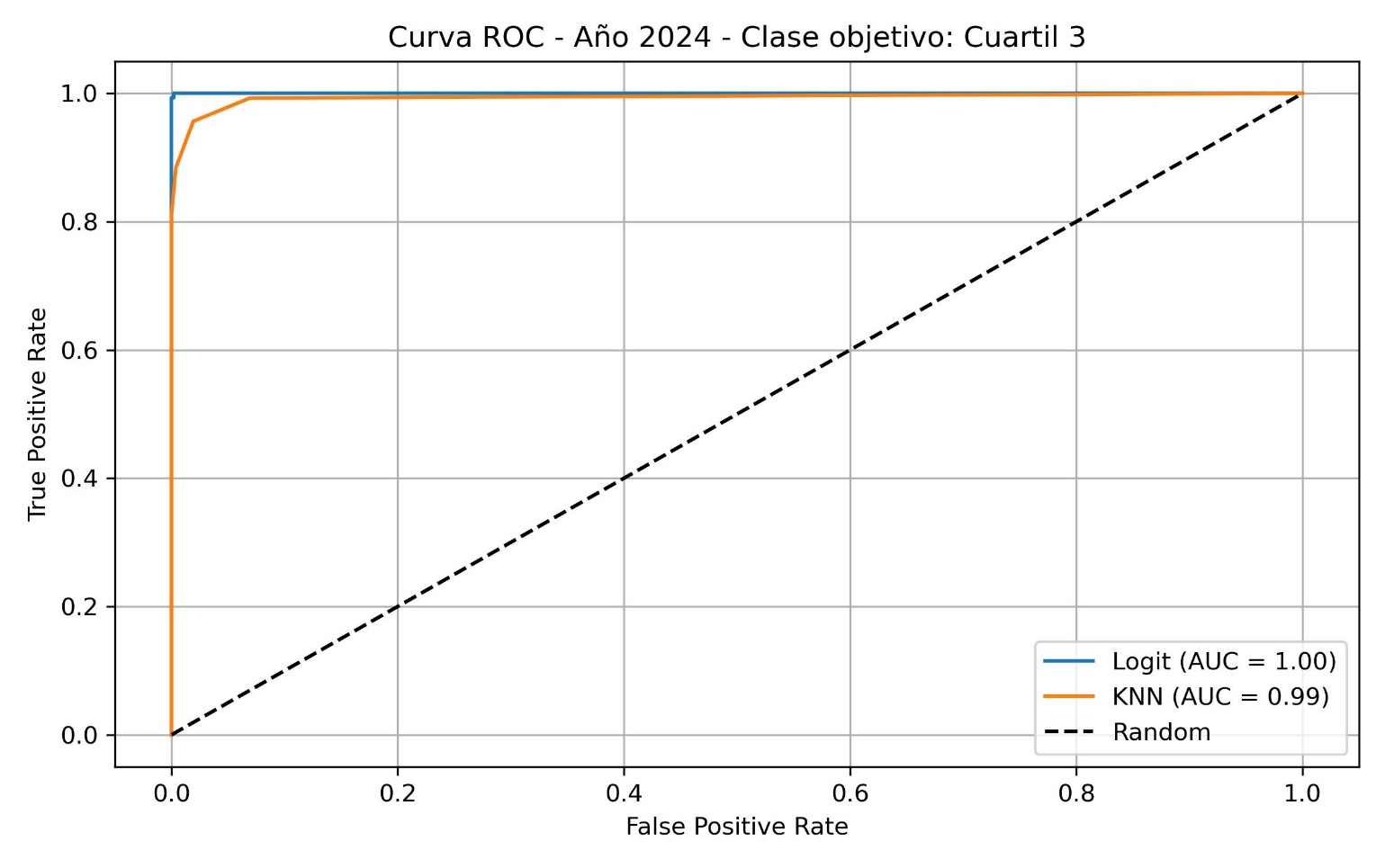
4)

El modelo 5 predice bien el salario promedio por edad, pero no capta los valores extremos. Hay alta dispersión, lo que limita su precisión individual.

*Parte C:*

5)





En ambos años, 2004 y 2024, la regresión logística funciona mejor que KNN para predecir el ingreso en cuartiles. El logit tiene mayor accuracy y un AUC perfecto (1.0), mostrando que discrimina mejor las clases. KNN también funciona bien, pero comete más errores y tiene un AUC un poco menor.

Esto indica que la relación entre las variables y el ingreso se ajusta mejor a un modelo lineal. Por eso, el logit es el método más confiable para esta predicción en ambos períodos.

6)

Al entrenar el modelo con la base de personas que respondieron, compuesta por casi toda gente ocupada, la regresión predecía inicialmente a todos como ocupados, consiguiendo una gran efectividad sobre la totalidad de la base pero con todos falsos negativos en la sección de desocupados. Luego de ajustar el modelo para ponderar los casos minoritarios (en este caso: los desocupados) llegamos a estos resultados que logran una gran mejora prediciendo desocupados en la base de los que no respondieron, a cambio de sacrificar un poco de la efectividad del modelo en términos generales (ocupados y desocupados) y un aumento en los falsos positivos en desocupados. Aproxidamente el 72% de la base de norespondieron, los predice como desocupados (varios de ellos pueden ser falsos positivos, debido a los ajustes explicados previamente, pero si se quiere mayor rigurosidad es posible considerar únicamente como desocupados aquellos con probabilidades muy altas)