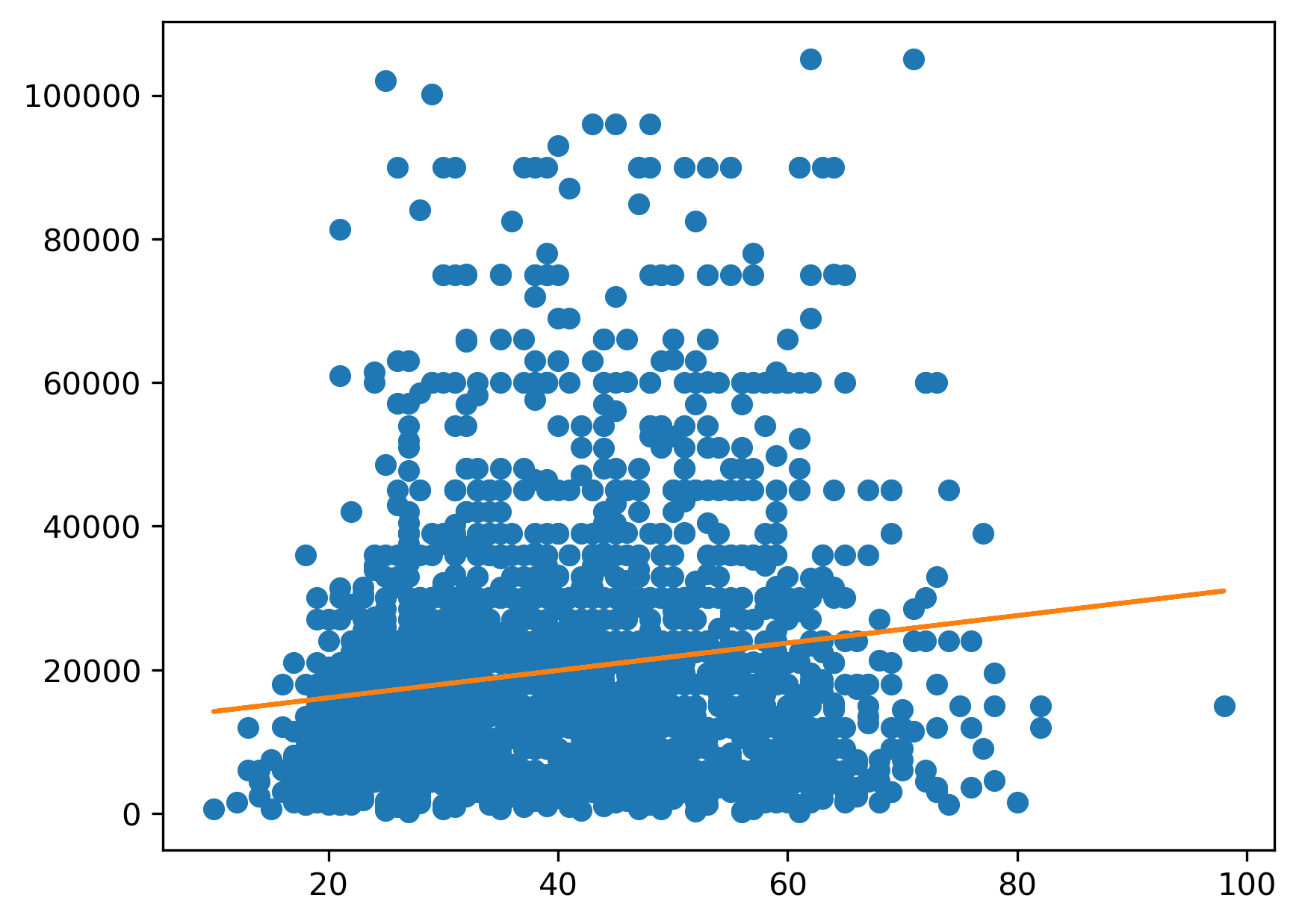
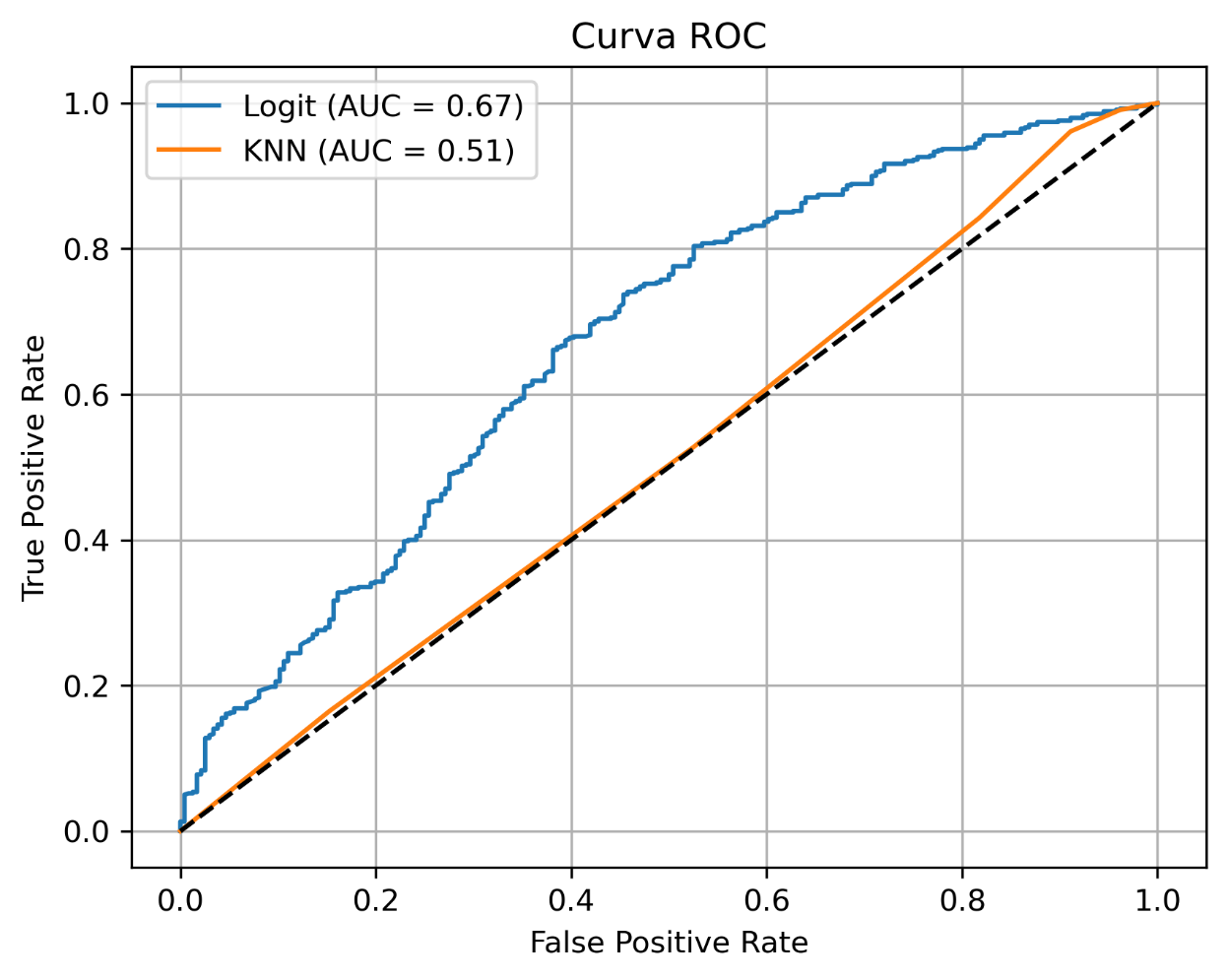
TP4 – Regresión y Clasificación

## Integrantes: David Agudelo, Juan Tomasello y Franco Bustelo



En el gráfico de arriba se muestra la relación entre la edad (eje X) y el salario semanal (eje Y). Cada punto azul representa a una persona, y la línea naranja es lo que aparece al aplicar una regresión lineal.

Para este histograma se usó el 70% de los datos para entrenar el modelo y el 30% para probarlo. Aunque la línea naranja muestra que a mayor edad tiende a subir el salario, la dispersión de los puntos indica que la edad no explica muy bien el salario semanal por sí sola. Esto sirve para ver si el modelo puede mejorar agregando más predictores.



En este grafico se ve que el modelo KNN tuvo un desempeño más bajo, con una precisión del 64% y un Auc de 0.51 (parecido al de tirar una moneda), lo cual vendría a indicar que no fue muy útil que digamos. Por el otro lado se ve que la regresión logística clasificó correctamente el 77% de los casos y el área bajo la curva ROC fue de 0.67, lo cual indica un desempeño, dentro de lo que cabe, **aceptable**.

También se ve claramente que la curva de **Logit** está más alejada de la línea diagonal (que representa un modelo aleatorio), mientras que la curva de **KNN** está muy cerca de esa línea. Esto confirma que la regresión logística funciona mejor para este caso.

Con estas variables elegidas, la **regresión logística funciona mucho mejor que KNN** para predecir quién gana más de $10.000 semanales. La información como tener más educación o saber escribir está relacionada con salarios más altos, y la regresión logística fue mejor para ver eso.

**Tabla 2. Estimación por regresión lineal de salarios usando la base de entrenamiento**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var. Dep: *salario\_semanal*  Variables | Modelo 1  (1) | Modelo 2  (2) | Modelo 3  (3) | Modelo 4  (4) | Modelo 5  (5) |
| *edad* |  |  |  |  | 1.239 |
| *edad2* |  |  |  |  | -1.223 |
| *educ* |  |  |  |  | 3.367 |
| *Mujer* |  |  |  |  | -6.109 |
| *Variable 1* |  |  |  |  | -1.568 |
| *Variable 2* |  |  |  |  | 2.114 |
|  |  |  |  |  |  |
| N (observaciones) | 2557 | 2557 | 2557 | 2557 | 2557 |
| *R2* |  |  |  |  |  |

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