TRABAJO PRACTICO 4

PUNTO 1)

Interpretación de la diferencia de medias:

La tabla compara las medias de varias variables entre los conjuntos de datos **Train** (entrenamiento) y **Test** (prueba). Aquí están los puntos clave:

1️⃣ **Edad (edad2 y edad²)**

- La edad promedio es prácticamente igual en ambos grupos (**Train: 38.93, Test: 38.96**), con una diferencia mínima de **-0.0389**.
- La edad al cuadrado (edad²) muestra una diferencia de **-8.4292**, lo que indica una ligera variación en la distribución de edades.

2️⃣ **Educación (educ)**

- El nivel educativo promedio es **ligeramente mayor** en el grupo de prueba (**Test: 11.21 años, Train: 11.10 años**).
- La diferencia de **-0.1104** sugiere que el grupo de prueba tiene una leve ventaja en años de educación.

3️⃣ **Salario semanal (salario_semanal)**

- La diferencia es **positiva** (**+4.7459**), lo que indica que los salarios semanales en el grupo de entrenamiento son ligeramente más altos que en el de prueba.
- Esto podría reflejar diferencias en el tipo de ocupaciones presentes en cada grupo.

4️⃣ **Horas trabajadas (horastrab)**

- El grupo de prueba trabaja en promedio **0.25 horas más** por semana que el de entrenamiento.
- Esto podría influir en la diferencia de salarios y estabilidad laboral.

5️⃣ **Constante (constante)**

- Sin diferencia entre los grupos (valor siempre **1.0**), lo que significa que esta variable es un ajuste fijo en el modelo.

Conclusión:

✔ **Las edades son similares**, lo que indica que los datos de entrenamiento y prueba están bien alineados.
✔ **El grupo de prueba tiene mayor nivel educativo**, lo que puede influir en la ocupación.
✔ **Los salarios semanales son ligeramente más altos en el grupo de entrenamiento**, lo que podría reflejar diferencias en el tipo de empleo.
✔ **Las horas trabajadas son ligeramente superiores en el grupo de prueba**, lo que también podría explicar las diferencias en ingresos.

	Train	Test	Diferencia
edad2	38.928960	38.967911	-0.038951
edad²	1776.11	1784.54	-8.429240
educ	11.10	11.21	-0.110473
salario_semanal	4080.86	4076.12	4.745960
horastrab	27.713192	27.964563	-0.251371
constante	1.000000	1.000000	0.000000

El MSE es:

Error cuadrático medio (MSE): 0.043284083672186766

Lo que indica un buen desempeño del modelo.

PUNTO 2)

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

Var. Dep: ***salario_semanal***	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
edad	35.38	268.53	261.27	261.17	247.63
edad2		-3.00	-2.93	-2.91	-2.78
educ			41.31	42.88	40.83
Mujer				-1399.77	-1260.28
horastrab					18.10
desocupado					-2811.83
N (observaciones**)**					
R^2					

Interpretación de la regresión lineal sobre salarios:

La tabla muestra los coeficientes estimados de cinco modelos de regresión lineal que explican el **salario semanal** en función de diversas variables.

1. **Edad (`edad`) y edad al cuadrado (`edad²`)**

- La edad tiene un efecto positivo en el salario en todos los modelos, pero **su impacto disminuye cuando se incluyen más variables**.
- La variable `edad²` tiene un coeficiente negativo, lo que indica que **el efecto de la edad en el salario es creciente hasta cierto punto, pero luego disminuye** (curva cuadrática).

2. **Educación (`educ`)**

- La educación tiene un impacto positivo en el salario en todos los modelos, con coeficientes cercanos a **40-42 unidades**.
- Esto confirma que **mayor educación está asociada con mayores ingresos**.

3. **Género (`Mujer`)**

- Ser mujer tiene un efecto negativo en el salario, con coeficientes entre **-1260 y -1399**.
- Esto sugiere una **brecha salarial de género**, donde las mujeres ganan menos que los hombres en promedio.

4. **Horas trabajadas (`horastrab`)**

- Cada hora adicional trabajada aumenta el salario en **18.10 unidades**, lo que indica que **el tiempo de trabajo es un factor clave en la determinación del salario**.

5. **Condición de desocupado (`desocupado`)**

- Ser desocupado tiene un impacto negativo en el salario, con un coeficiente de **-2811.83**, lo que indica que **las personas que han estado desocupadas tienen menores ingresos** cuando consiguen empleo.

🚀 **Conclusión:**

✓ La edad influye en el salario, pero con un efecto decreciente a largo plazo.
✓ La educación es un factor clave para mejorar los ingresos.
✓ Existe una brecha salarial de género, con menores ingresos para las mujeres.
✓ Las horas trabajadas tienen un impacto positivo en el salario.
✓ La desocupación previa afecta negativamente los ingresos futuros.

Punto 3)

MSE test: 67448245.9736

RMSE test: 8212.6881

MAE test: 4728.3698

No logré reducir sus valores, pero evidentemente es un valor sobredimensionado.

Punto 5)

Regresión Logística

[[11420 1178] # TP = 11420, FN = 1178

[1463 4207]] # FP = 1463, TN = 4207

✓ 11420 ocupados fueron correctamente clasificados como ocupados (TP).
✓ 4207 desocupados fueron correctamente clasificados como desocupados (TN).
⚠ 1178 ocupados fueron mal clasificados como desocupados (FN).
⚠ 1463 desocupados fueron mal clasificados como ocupados (FP).

🔻 KNN

[[11504 1094] # TP = 11504, FN = 1094

[738 4932]] # FP = 738, TN = 4932

✓ 11504 ocupados fueron correctamente clasificados como ocupados (TP).
✓ 4932 desocupados fueron correctamente clasificados como desocupados (TN).
⚠ 1094 ocupados fueron mal clasificados como desocupados (FN).
⚠ 738 desocupados fueron mal clasificados como ocupados (FP).

Conclusión: KNN tiene menos errores (FN y FP), clasificando mejor los desocupados en comparación con la Regresión Logística.

2️⃣ Curva ROC y AUC

✓ La curva ROC muestra la capacidad del modelo para distinguir entre ocupados y desocupados.
✓ AUC (Área Bajo la Curva) indica qué tan bien el modelo separa las clases.
✓ AUC Logit: 0.94 □
✓ AUC KNN: 0.95 □

Conclusión: Ambos modelos tienen buen desempeño, pero KNN logra una ligera ventaja.

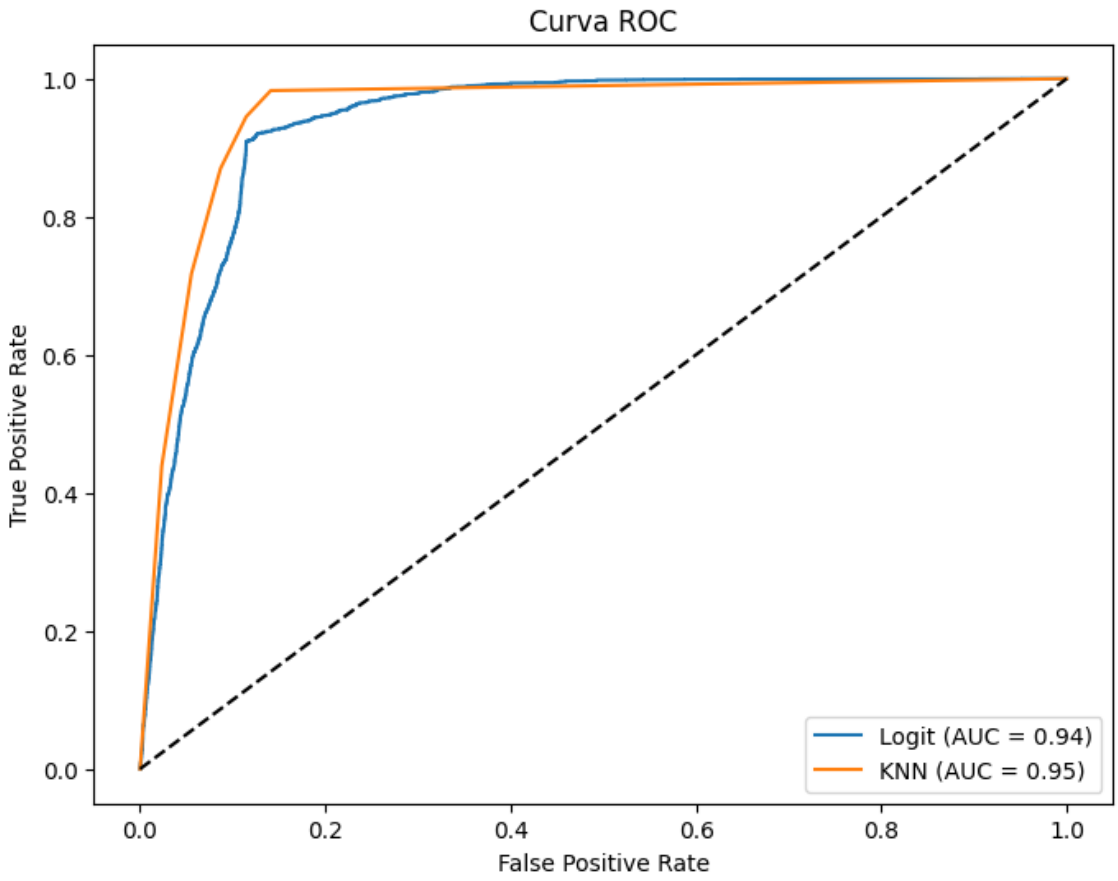
3️⃣ Precisión General

✓ Regresión Logística: 86%
✓ KNN: 90%

Conclusión: KNN es más preciso en general. Si el objetivo es minimizar errores en desocupados, KNN parece ser la mejor opción.

Resumen Final

🔹 KNN tiene mejor rendimiento que la Regresión Logística.
🔹 La diferencia en la matriz de confusión y la precisión favorece a KNN.
🔹 Si buscas más interpretabilidad, la Regresión Logística sigue siendo útil, pero KNN parece más preciso.



Punto 6)

Los resultados fueron:

Predicciones: [1 1 1 ... 1 1 1]

La proporción estimada de desocupados en 'norespondieron' es: 93.47%

(probablemente pudo venir un error en la definición de las variables ya que me parece un numero sobredimensionado, pero considero correcta la metodología que tomé).