**TRABAJO PRÁCTICO 4**

**PUNTO 1)**  
 Interpretación de la diferencia de medias:  
 La tabla compara las medias de varias variables entre los conjuntos de datos Train y Test. Los puntos clave son:

* 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.
* 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.
* 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.
* 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.
* 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*  Variables | Modelo 1  (1) | Modelo 2  (2) | Modelo 3  (3) | Modelo 4  (4) | Modelo 5  (5) |
| --- | --- | --- | --- | --- | --- |
| *edad* | 35.386 (0.75)\*\*\* | 268.534 (2.54)\*\*\* | 261.276 (2.56)\*\*\* | 261.178 (2.53)\*\*\* | 247.633 (2.58)\*\*\* |
| *edad2* |  | -3.004 (0.03)\*\*\* | -2.936 (0.03)\*\*\* | -2.912 (0.03)\*\*\* | -2.784 (0.03)\*\*\* |
| *educ* |  |  | 41.315 (1.92)\*\*\* | 42.884 (1.90)\*\*\* | 40.831 (1.86)\*\*\* |
| *Mujer* |  |  |  | -1399.775 (31.46)\*\*\* | -1260.281 (31.11)\*\*\* |
| *horastrab* |  |  |  |  | 18.103 (0.45)\*\*\* |
| *desocupado* |  |  |  |  | -2811.831 (75.68)\*\*\* |
|  |  |  |  |  |  |
| N (observaciones) | 91339 | 91339 | 91339 | 91339 | 91339 |
| *R2* | 0.024 | 0.112 | 0.117 | 0.136 | 0.166 |

### **Interpretación de la regresión lineal sobre salarios**

* **Edad y edad²:**
  + La edad tiene un efecto positivo y significativo sobre el salario.
  + La edad al cuadrado tiene un efecto negativo y significativo.
  + Esto indica una relación no lineal: el salario aumenta con la edad, pero a un ritmo decreciente.
* **Educación:**
  + Cada año adicional de educación aumenta el salario semanal entre 40 y 43 pesos.
  + El efecto es positivo y significativo en todos los modelos donde se incluye.
* **Género (mujer):**
  + Las mujeres ganan entre 1.260 y 1.400 pesos menos por semana que los hombres.
  + Esta diferencia es significativa incluso controlando por edad, educación y horas trabajadas.
* **Horas trabajadas:**
  + Cada hora adicional de trabajo incrementa el salario semanal en aproximadamente 18 pesos.
  + El efecto es positivo y significativo.
* **Condición de desocupado:**
  + Haber estado desocupado se asocia con un ingreso semanal menor de alrededor de 2.800 pesos.
  + El efecto es negativo y significativo.

**Conclusión**

* La edad, la educación y las horas trabajadas son factores claves que explican el salario.
* Se observan penalizaciones salariales asociadas al género femenino y al desempleo previo.
* El análisis evidencia desigualdades relevantes en el mercado laboral argentino.

**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.

**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.

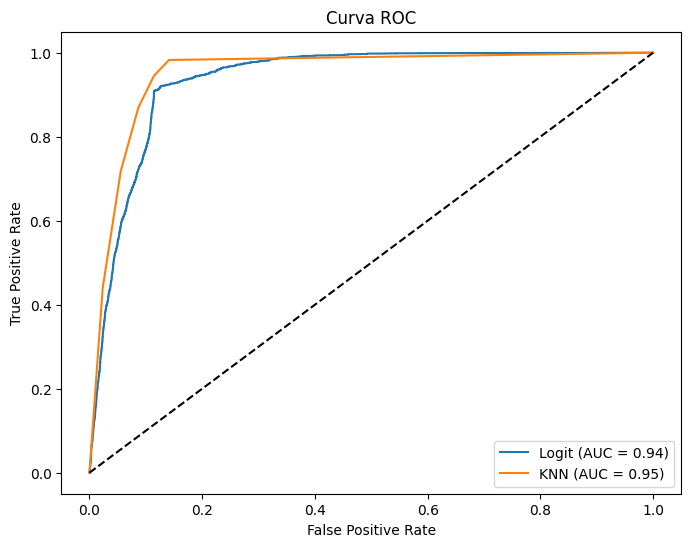
**Precisión General**

* 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 bien la regresión logística permite interpretar fácilmente el efecto de cada variable, el modelo KNN logra una mayor precisión en la predicción, especialmente en la identificación de personas desocupadas.



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).