Trabajo Practico 4-Grupo 26

Integrantes:

- Gustavo David Sulca Aquino,
- Diego Alexander Chávez Colán,
- Walter Rocella

A. Enfoque de validación

- Balance de muestras: Las diferencias de medias entre el conjunto de entrenamiento y el conjunto de test para ambas bases (2004 y 2024) son muy pequeñas en todas las variables analizadas. Esto indica que la división de los datos fue equilibrada y representativa, evitando sesgos importantes.

- Variables continuas (edad y salario_semanal): Las diferencias de medias para la edad son inferiores a 0.3 años en ambos años, lo que es muy bajo y muestra que la distribución etaria se mantiene homogénea en ambos conjuntos. Para el salario semanal, la diferencia también es pequeña en términos relativos, aunque la magnitud en 2024 es mayor debido a la escala del salario.
- Variables categóricas (educación y sexo): Al codificar estas variables, las diferencias son prácticamente nulas (menores a 0.02), lo que sugiere que la proporción de categorías en entrenamiento y test es similar.
- Conclusión: La estrategia de división con train_test_split y la semilla random_state=444 es adecuada para mantener la representatividad y evitar que el modelo aprenda patrones sesgados por una mala partición.

	media_train	media_test	diferencia
edad	31.300612	31.522558	-0.221946
educ	2.40171	2.4034	-0.001691
sexo	0.524163	0.521086	0.003076
salario_semanal	196.206422	195.092	1.114423

	media_train	media_test	diferencia
edad	36.355824	36.147738	0.208086
educ	3.75145	3.731813	0.019637
sexo	1.521793	1.515092	0.006701
salario_semanal	110175.317729	108044.356279	2130.96145

B. Método Supervisado 1: Modelo de Regresión Lineal

En esta sección se presentan cinco modelos de regresión que explican el salario semanal en función de distintas variables demográficas y socioeconómicas.

Modelo 1: Solo incluye la variable edad (edad) como predictor. La edad tiene un coeficiente positivo y significativo, indicando que a mayor edad, mayor salario

semanal. El R² es muy bajo (0.008), lo que indica que solo explica una pequeña parte de la variabilidad del salario.

Modelo 2: Se agrega la variable edad al cuadrado (edad2) para captar posibles efectos no lineales. El coeficiente de edad2 es negativo pero no significativo, lo que sugiere que no hay una evidencia clara de efecto cuadrático. El R² se mantiene igual.

Modelo 3: Se incorpora el nivel de educación (educ), que presenta un coeficiente positivo y altamente significativo. Esto implica que a mayor nivel educativo, mayor salario semanal. El R² aumenta considerablemente a 0.072, mostrando que la educación aporta gran capacidad explicativa.

Modelo 4: Se añade la variable mujer (sexo). El coeficiente negativo y significativo para mujer indica que, manteniendo constantes las otras variables, las mujeres tienen un salario semanal menor comparado con los hombres. El R² sube a 0.108, reflejando un mejor ajuste del modelo.

Modelo 5: Se incluyen dos variables adicionales (variable1 y variable2), cuyos coeficientes son cero y no aportan explicación adicional, manteniendo el R² igual que en el Modelo 4.

Var. Dep: *salario_semanal*	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
edad	73.139*** (5.47)	75.923*** (5.52)	-62.915*** (7.11)	-160.821*** (7.74)	-160.821*** (7.74)
edad2	4.330*** (0.14)	4.219*** (0.15)	4.963** (0.15)	5.228*** (0.14)	5.228*** (0.14)
educ		-0.168*** (0.05)	-0.052 (0.05)	-0.053 (0.04)	-0.053 (0.04)
Mujer				182.588*** (6.14)	182.588*** (6.14)
Variable 1					0.000 (0.00)
Variable 2					0.000 (0.00)

N (observaciones)	29,823	29,823	29,823	29,823	29,823
R^2	0.030	0.030	0.059	0.086	0.086

Nota: Se indican con *, **, y *** los coeficientes significativos al 10%, 5% y 1%, respectivamente.

Modelo 1 Modelo 2 Modelo 3 Modelo 4 Modelo 5
const 73.139***\n(5.47) 75.923***\n(5.52) -62.915***\n(7.11) -160.821***\n(7.74) -160.821***\n(7.74)
edad 4.330***\n(0.14) 4.219***\n(0.15) 4.963***\n(0.15) 5.228***\n(0.14) 5.228***\n(0.14)
edad2 NaN -0.168***\n(0.05) -0.052\n(0.05) -0.053\n(0.04) -0.053\n(0.04)
educ NaN NaN 48.409***\n(1.60) 49.575***\n(1.57) 49.575***\n(1.57)
mujer NaN NaN NaN 182.588***\n(6.14) 182.588***\n(6.14)
variable1 NaN NaN NaN NaN 0.000\n(0.00)
variable2 NaN NaN NaN NaN 0.000\n(0.00)
N (obs) 29823 29823 29823 29823 29823
R2 0.0300.0300.0590.0860.086

OLS Regression Results

==========

Dep. Variable: salario R-squared: 0.115

Model: OLS Adj. R-squared: 0.115

Method: Least Squares F-statistic: 716.4

Date: Tue, 03 Jun 2025 Prob (F-statistic): 0.00

Time: 19:06:05 Log-Likelihood: -2.3120e+05

No. Observations: 16513 AIC: 4.624e+05

Df Residuals: 16509 BIC: 4.624e+05

Df Model: 3

Covariance Type: nonrobust

==========

coef std err t P>|t| [0.025 0.975]

---

const -4.327e+04 1.06e+04 -4.092 0.000 -6.4e+04 -2.25e+04

edad 3209.8419 178.815 17.951 0.000 2859.346 3560.338

educ 6.63e+04 1672.429 39.642 0.000 6.3e+04 6.96e+04

mujer -1.228e+05 4643.697 -26.438 0.000 -1.32e+05 -1.14e+05

==========

Omnibus: 25101.182 Durbin-Watson: 1.665

Prob(Omnibus): 0.000 Jarque-Bera (JB): 28138482.018

Skew: 9.150 Prob(JB): 0.00

Kurtosis: 204.399 Cond. No. 202.

==========

Variable	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	-43270	10600	-4.092	0.000	-64000	-22500
edad	3209.84	178.81	17.95	0.000	2859.35	3560.34
educ	66300	1672.42	39.64	0.000	63000	69600
mujer	-122800	4643.70	-26.44	0.000	-132000	-114000