UNIVERSIDAD DE BUENOS AIRES - 2025

BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N° 4

MÉTODOS SUPERVISADOS: REGRESIÓN & CLASIFICACIÓN USANDO LA EHP

A. Enfoque de validación - punto 1

	N Train	Mean Train	SD Train	N Test	Mean Test	SD Test	T-Test	P-Value
Constante	31702	1	0	13587	1	0		
Desocupado	31702	0	0	13587	0	0		

Año 2024: el promedio estimado del resto de las variables en el entrenamiento son 1, y el promedio estimado del resto de las variables en el testeo son 1.
El promedio estimado de la variable 'desocupados' en el entrenamiento es 0,029523, y el promedio estimado de desempleo en el testeo es 0,029875.

	N Train	Mean Train	SD Train	N Test	Mean Test	SD Test	T-Test	P-Value
Constante	31702	1	0	13587	1	0		
Desocupado	31702	0	0	13587	0	0		

Año 2004: el promedio estimado del resto de las variables en el entrenamiento son 1, y el promedio estimado del resto de las variables en el testeo son 1.
El promedio estimado de la variable 'desocupados' en el entrenamiento es 0. Y el promedio estimado de desempleo en el testeo es 0.

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	2180.320 (75.49)***	11997.859 (267.48)***	12102.218 (267.58)***	12096.915 (265.12)***	10369.479 (280.03)***
edad2		-129.499 (3.40)***	-130.177 (3.39)***	-129.651 (3.36)***	-107.082 (3.48)***
educ			-640.494 (96.39)***	-691.245 (95.55)***	-448.534 (91.97)***
Mujer				-50433.716 (2921.38)***	-36139.598 (2803.77)***
Variable1 (CHO9)					-12977.419 (6184.40)**
Variable2 (PP02I)					-58711.361 (1485.02)***
N (observaciones)	15914	15914	15914	15914	15914
R^2	0.050	0.129	0.132	0.147	0.227

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

En el modelo 1 la variable edad tiene un coeficiente positivo. Con un desvió estándar de 75.49.

En el modelo dos se le agrega la variable de 'edad²'. El coeficiente de edad aumenta a 11997.859, con un desvió estándar de 267.48, y en cuanto a la variable edad² tiene un coeficiente negativo de -129.499 con un desvió estándar de 3.40.

En el modelo tres se agrega la variable 'educ', cuyo coeficiente es negativo (-640.494) y posee un desvió estándar de 96.39. También podemos observar que los valores de los coeficientes de edad y edad² disminuyeron.
En el modelo 4 se incorpora la variable 'mujer' con un coeficiente negativo y significativo (-50433.716) y un desvió estándar de 2921.38.

Por último, el modelo 5 incluye dos nuevas variables 'CH09' (¿Sabe leer y escribir?) Y 'PP02I' ('En los últimos 12 meses ¿trabajó en algún momento?'). Ambas variables poseen coeficientes y desvíos con efectos negativos y significativos, lo que podría indicar que, el hecho de que la persona sepa o no leer, y haya o no trabajado en os últimos 12 meses, podría impactar negativamente en el salario semanal.

Tabla 3 – Enfoque de validación – Predicción de salarios

	(1)	(2)	(3)	(4)	(5)
MSE test	6.798489e+10	6.788802e+10	6.670322e+10	6.638681e+10	6.638115e+10
RMSE test	2.607391e+05	2.605533e+05	2.582697e+05	2.576564e+05	2.576454e+05
MAE test	1.004325e+05	9.964191e+04	9.730917e+04	9.670545e+04	9.671565e+04

Teniendo en cuenta que:

- MSE: Error cuadrático medio
- RMSE: Raíz del MSE
- MAE: Error absoluto medio

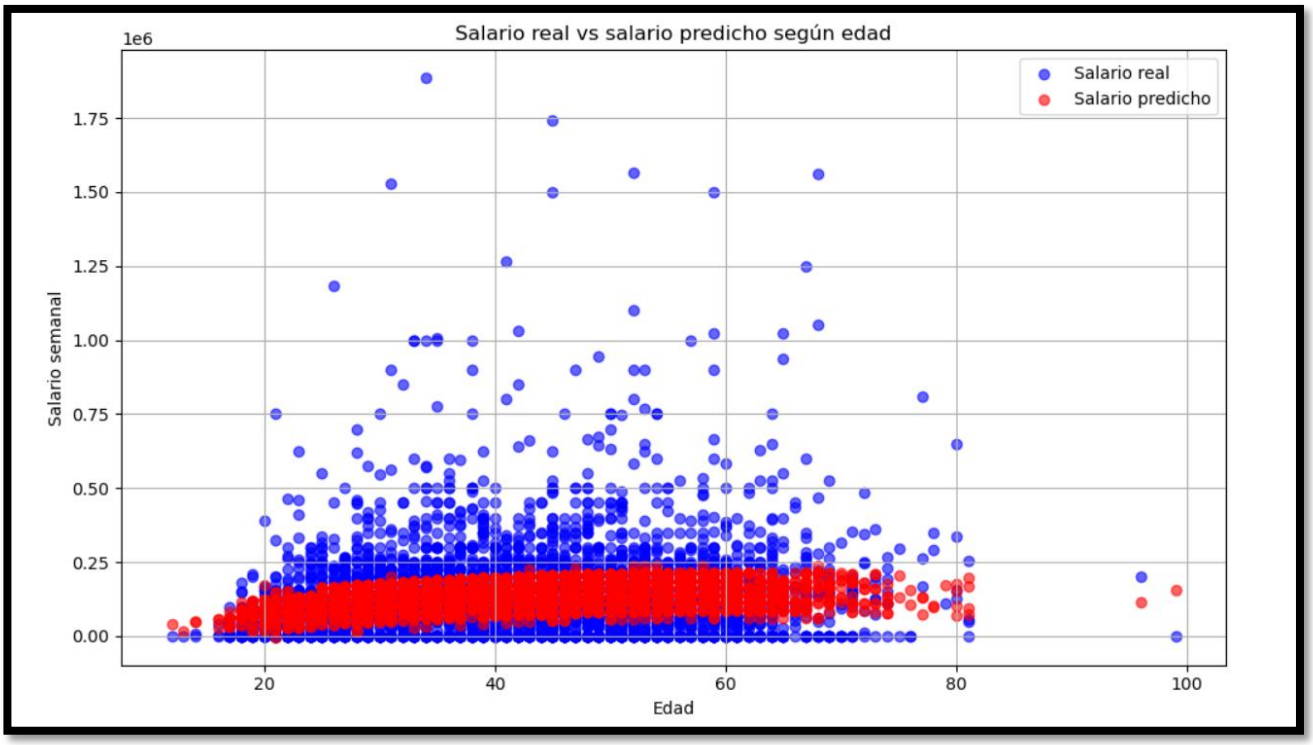
En el modelo 1, sólo se incluye la variable "edad", que justamente presenta demasiados errores en las tres métricas. el error promedio es mayor a $260 mil pesos y la diferencia absoluta promedio con el salario real supera los $100 mil. Esto indica que la edad, por sí sola, no alcanza para explicar las diferencias salariales.

En el modelo 2, se incorpora "edad2" donde observamos una leve mejora bajando el MAE en menos de $ 1.000 pesos.

En el modelo 3, se incorpora la variable "educ" donde vemos que mejora notablemente los valores de MSE, RMSE y MAE. Esto confirma que el nivel educativo es un factor importante en la determinación del salario semanal, como se espera en la literatura laboral.

En el Modelo 4, al sumar la variable "mujer" sigue mejorando ligeramente el desempeño del modelo. Si bien la mejora es pequeña, esto sugiere que existen diferencias sistemáticas de ingreso por género, que el modelo logra capturar.

Gráfico 4 - de dispersión – Predicción de Salarios



El gráfico ilustra cómo se relaciona la edad con el salario semanal, comparando lo que realmente ganan las personas (en azul) con lo que estima el modelo (en rojo).

Se puede ver que las predicciones se ubican en un rango bastante acotado, con una ligera tendencia creciente entre los 20 y los 60 años. Sin embargo, los datos reales muestran mucha más variabilidad, sobre todo a partir de los 30, donde aparecen varios casos con ingresos muy altos que el modelo no logra anticipar.

Esta diferencia se explica, en parte, porque el modelo usa pocas variables y no tiene en cuenta factores que también influyen en el salario, como el tipo de trabajo, las horas que se trabajan, la experiencia o incluso el sector económico.

Gráfico del punto 5

Se decide implementer lo que son modulo de clasificación para predecir la desocupación: regresión logística (Logit) y vecinos más cercanos (KNN) con K=5K = 5K=5. Se entrenaron ambos modelos sobre el 70% de los datos con respuesta, y se evaluaron sobre el 30% restante.

Modelo	Accuracy	AUC
Logit	0.928	0.955
KNN (k=5)	0.910	0.947

Lo que es la proporción de aciertos y AUC (área bajo la curva ROC) fueron ligeramente superiores en el modelo de Logit.
La matriz de confusión del modelo Logit mostró menos falsos positivos y menos falsos negativos que KNN.
En ambos modelos, la curva ROC presentó buen poder discriminante, pero la de Logit fue levemente mejor.

Logit) mostró un mejor rendimiento general en términos de precisión y discriminación. Por lo tanto, se selecciona Logit como el mejor modelo para predecir la desocupación en este contexto.

Logit: AUC 0.955 — Accuracy 92.8%
KNN: AUC 0.947 — Accuracy 91.0%