TP4 – Regresión y Clasificación

Integrantes: David Agudelo, Juan Tomasello y Franco Bustelo

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	207.924***	1141.612***	1142.524***	1178.789***	1047.242***
	(29.92)	(142.49)	(142,50)	(140,585)	(155.17)
edad2		-11.736***	-11,750***	-12.173***	-11.008***
		(1.66)	(1,66)	(1,63)	(1.75)
educ			65,945	56,626	58.049
			(72,39)	(71,38)	(70.87)
Mujer				-5028,269***	-4972.265***
				(698,55)	(709.91)
Estado civil_2					4066.256***
					(973.65)
Estado civil_3					3042.810 **
					(1521.96)
Estado civil_4					2612.815
					(2325.59)
Estado civil_5					1852.749*
					(1072.42)
escribe					-13900***
					(3876.988)
N (observaciones)	2557	2557	2557	2557	2557
R^2	0,026	0,046	0,046	0,074	0,09

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

En esta tabla vemos los resultados de varios modelos que intentan explicar el salario semanal según variables/características de las personas. En general, la variable más fuerte es la edad: la edad tiene incidencia en el salario semanal (no es causalidad aclaramos) pero como también se incluye la edad al cuadrado (edad2), eso nos muestra que llega un punto donde la edad ya no impacta tanto en el salario semanal o incluso puede llegar a bajarlo (porque el coeficiente pasa a ser negativo).

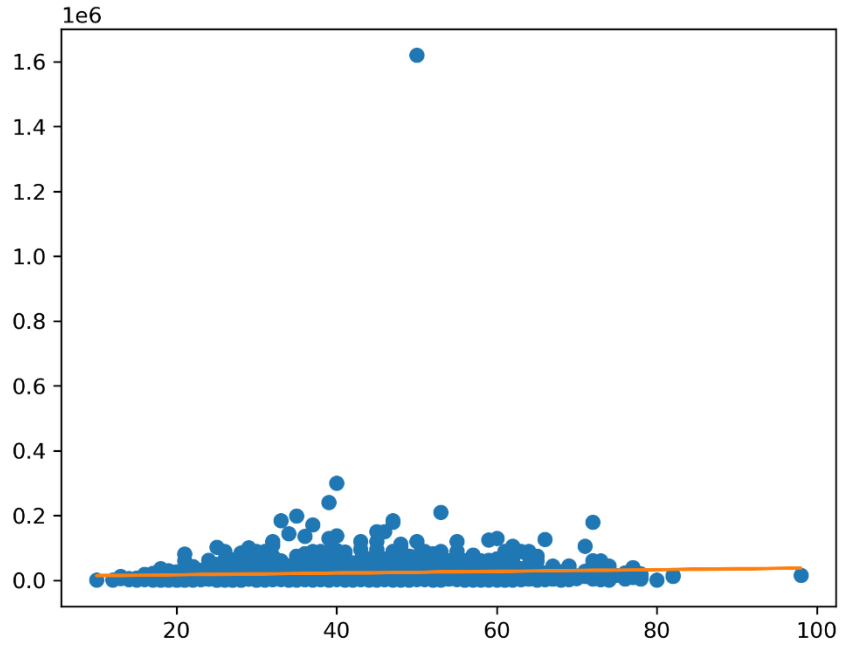
Por otro lado, estar casado (estado civil 2) o separado (estado civil 3) parece estar asociado con salarios más altos que quienes no están en esos estados, especialmente casado que tiene un coeficiente más grande y significativo. En cambio, ser viudo no tiene efecto claro (nada significativo) y ser soltero (estado civil 5) tiene un impacto positivo pero más chico. Nos gustaría aclarar que en este caso al crear las dummies, eliminamos estado civil 1 (unido) para evitar multicolinealidad, por lo que tomamos como referencia estado civil 1, de forma que todas las dummies están comparadas a ella.

El nivel educativo (educ) tiene un efecto positivo, pero tiene un p valor alto, por lo que no es estadísticamente significativo, en otras palabras, no guarda buena relación con el salario semanal. Ser mujer parece tener un impacto negativo en el salario (coeficiente negativo y significativo), lo cual muestra una posible brecha salarial, también tiene un p valor bastante bajo por lo que es estadísticamente significativo. Podemos decir que el genero si tiene que ver en el salario semanal.

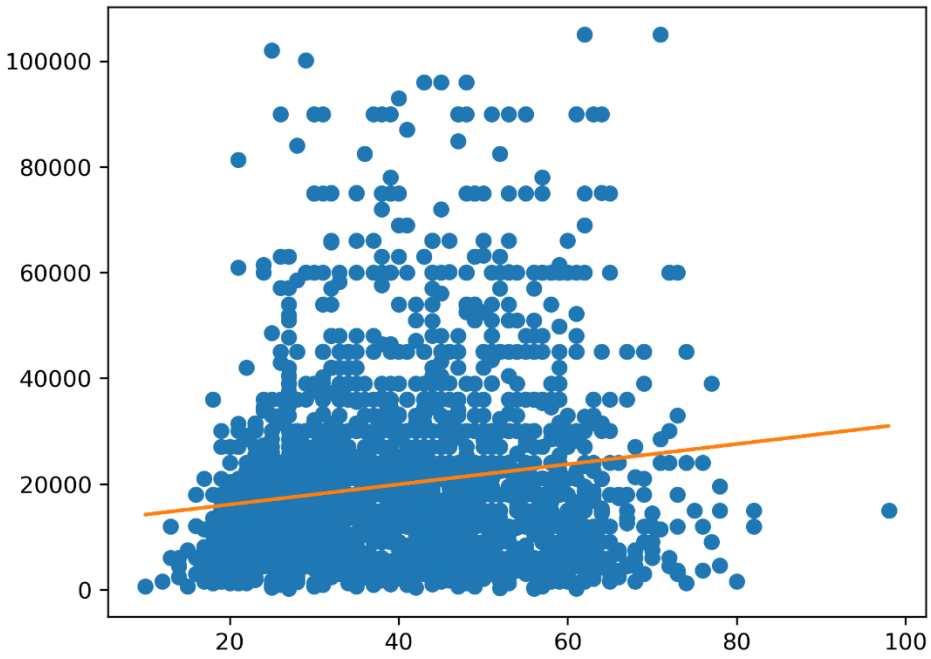
La variable escribe es una dummy que toma el valor 1 cuando no sabe escribir y el valor 0 cuando la persona si sabe escribir, en este caso eran dos columnas y eliminamos la primera para evitar la multicolinealidad. El p valor es bastante bajo así que podemos decir que es estadísticamente significativo, por lo que guarda relación si la persona sabe escribir o no con el salario semanal.

Por último, los R^2 son bastante bajos, por lo que no explican tanto de la varianza del salario semanal, el valor más alto lo adopta en la última columna llegando a un 0,09.

Punto 2 i)



Este grafico es el primero que hicimos pero había outliers o los valores extremos. Para solucionar esto, tal como lo aclaramos en los códigos mediante el uso del #, consideramos el percentil 99 y así poder tener un histograma mucho más legible y agradable a la vista. Con este grafico queríamos mostrarte un poco el paso a paso para que se entienda luego el otro histograma de la página de abajo.



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.

Tabla 3. Performance por regresión lineal de la predicción de salarios usando la base de testeo

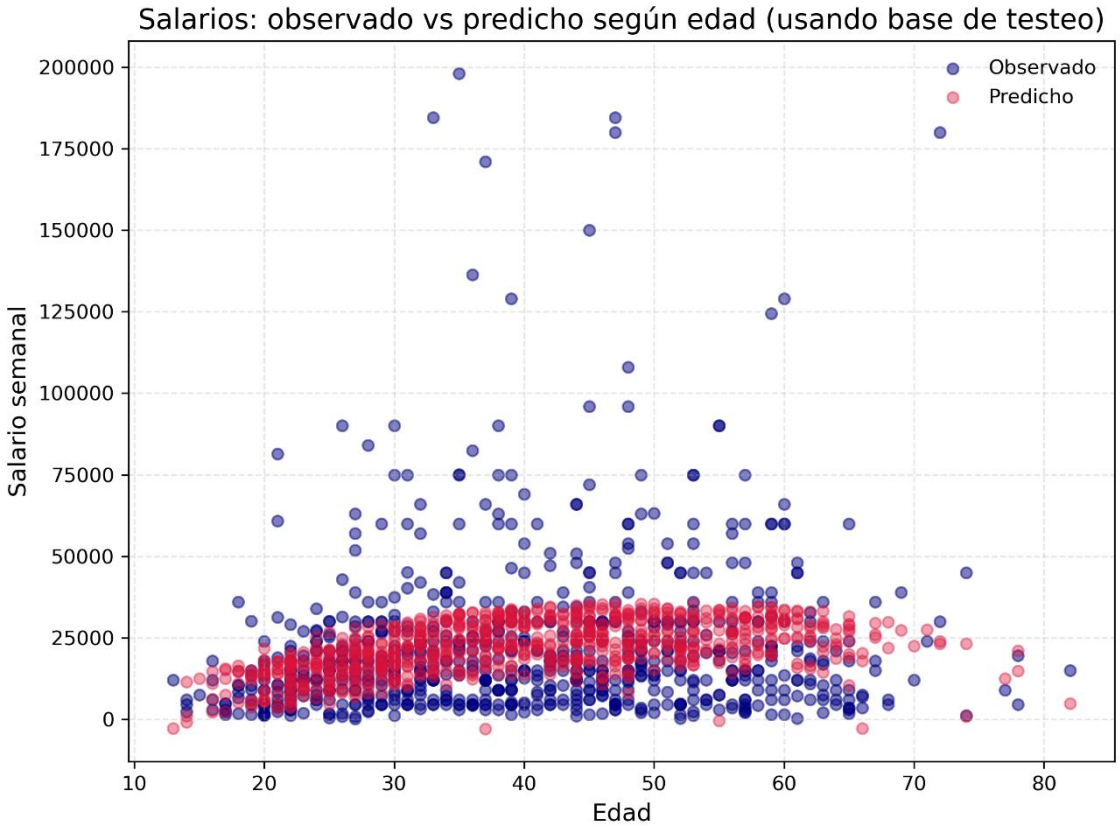
Var. Dep: ***salario_semanal***	**Modelo 1** **(1)**	**Modelo 2** **(2)**	**Modelo 3** **(3)**	**Modelo 4** **(4)**	**Modelo 5** **(5)**
MSE test	591364718	581252431	586252962	574227462	561771502
RMSE test	24317	24109	24212	23963	23701
MAE test	14857	14631	14800	14654	14656

En esta tabla se ve cómo rindieron los distintos modelos a la hora de predecir el salario usando los datos de testeo. Los tres indicadores (MSE, RMSE y MAE) van bajando a medida que se agregan más variables, lo que muestra que el modelo va mejorando un poco.

El MSE del modelo 3 sube con respecto al 2 porque variable educ tiene un p valor alto. Esto significaría que no es significativo, es decir, que educ y salario semanal no guardan una relación relevante como si lo hace con edad.

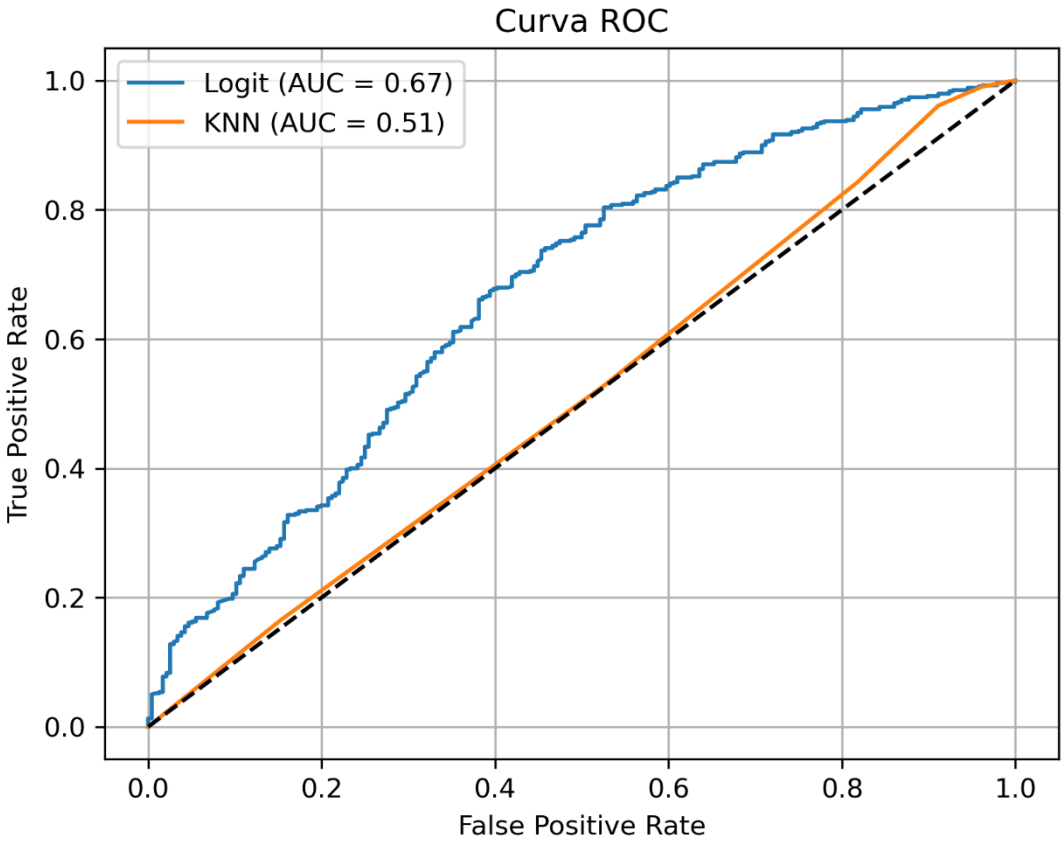
El Modelo 5 es el que tiene el mejor desempeño, con el menor error (MSE, RMSE y MAE más bajos). Igual, los errores siguen siendo bastante altos (por ejemplo, el RMSE tiene 23.700), así que aunque mejora, todavía no predice con mucha precisión. O sea, los modelos hacen un esfuerzo pero todavía les falta para acercarse más a los valores reales del salario.

Punto 4



En este gráfico se comparan los salarios reales observados con los que predijo el modelo, en función de la edad. Se ve que el modelo (los que están en rojo) sigue más o menos la tendencia general de los datos reales (los que están en azul), pero no llega a capturar bien los valores extremos (una especie de outliers). En general, el modelo predice mejor los salarios semanales más comunes, pero le cuesta con los casos atípicos. También se puede ver que los salarios no aumentan con la edad como uno esperaría, lo cual puede indicar que está relacionado con otros factores además de la edad.

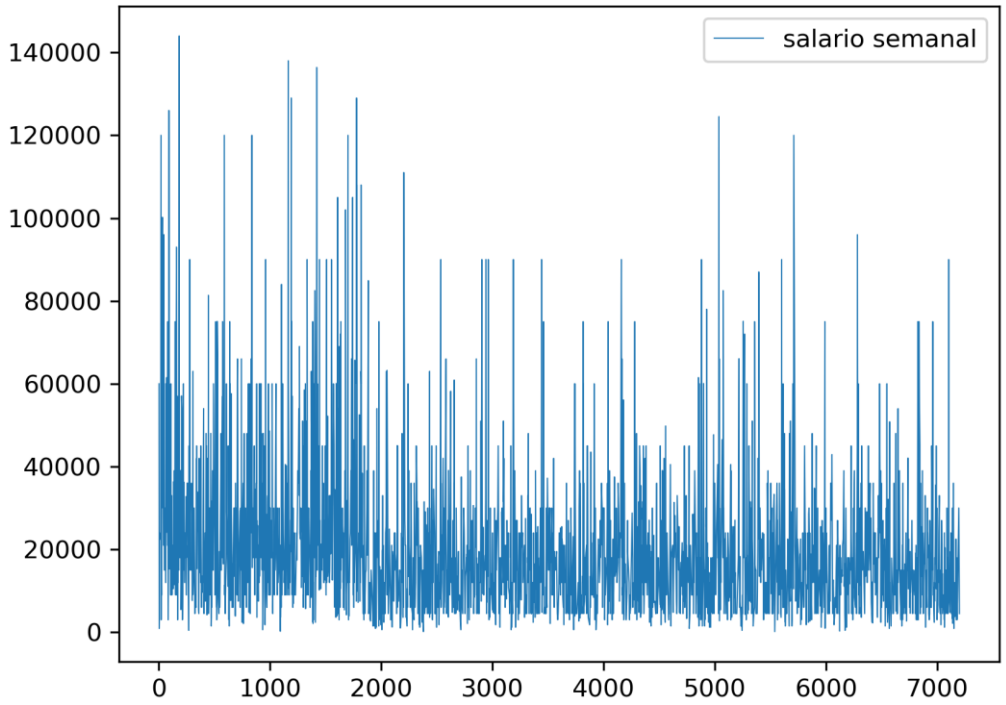
Punto 5



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 0.5), 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.



Un gráfico sencillo que muestra el volumen o densidad de personas con sus respectivos salarios semanales. Acá tuvimos que acortar con <150000 para que se vea mejor porque había un outlier o un valor extremadamente alto y alejado del resto de los salarios semanales.

Punto 6

	año	cond. de actividad	edad	edad2	educ	salario semanal	mujer	estado civil_2	estado civil_3	estado civil_4	estado civil_5	sabe escribir_2	prediccion_desocupado	probabilidad_desocupado
3	2004	0.00	54	2916	10	60000	0	1	0	0	0	0	1	0.84
5	2004	0.00	25	625	16	900	1	0	0	0	1	0	1	0.51
6	2004	0.00	20	400	11	3000	1	0	0	0	1	0	0	0.42
7	2004	0.00	55	3025	9	24000	0	0	0	0	0	0	1	0.75
12	2024	0.00	50	2500	11	22500	0	1	0	0	0	0	1	0.85
18	2024	0.00	62	3844	14	23400	1	0	1	0	0	0	1	0.57
21	2024	0.00	45	2025	5	120000	0	0	0	0	0	0	1	0.76
22	2024	0.00	41	1681	15	3000	1	0	0	0	0	0	1	0.56
23	2024	0.00	21	441	9	3000	1	0	0	0	1	0	0	0.44
25	2024	0.00	31	961	7	30000	0	1	0	0	0	0	1	0.81
26	2024	0.00	19	361	4	12000	1	1	0	0	0	0	0	0.44
27	2024	0.00	22	484	0	9000	1	0	0	0	1	0	0	0.44
28	2024	0.00	54	2916	12	51000	0	1	0	0	0	0	1	0.85
29	2024	0.00	52	2704	12	30000	1	1	0	0	0	0	1	0.69
31	2024	0.00	51	2601	4	90000	1	0	0	0	1	0	1	0.63
34	2024	0.00	29	841	14	100200	0	0	0	0	1	0	1	0.77
43	2024	0.00	47	2209	13	21000	1	1	0	0	0	0	1	0.70
46	2024	0.00	43	1849	14	96000	0	1	0	0	0	0	1	0.86

Nuestro modelo logit, nos predice que 14 de 18 personas de las que no respondieron la encuesta serán desocupados. Aclaramos que no sabemos si son desocupados o no realmente, debido a que no respondieron. Pero nuestro modelo nos predice eso en base a todos los predictores que usamos y a como lo entrenamos previamente. Si la probabilidad de desocupados es mayor a 0,5 entonces lo predice como desocupado, si es menor a 0,5 entonces es ocupado. Porque el umbral es 0,5.