Trabajo practico N°4

Iván Martínez, Joaquín Reyes y Santiago Reinoso Sokol

1) Tanto para la base 2004 como para 2024 vemos que la diferencia entre la media para la base de entrenamiento y la base de test es relativamente baja en la mayoría de las variables. Es decir, si bien hay diferencias en los valores del promedio estas son bastante similares, por lo que podríamos decir que la base de entrenamiento es representativa de la base de testeo.
2) Tabla:

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Edad	-1117.443	-5087.510	-7310.923	-6724.342	-3362
Edad2		426.321	653.929	610.846	610
Educ			1525.947	1646.343	1646
mujer				-2218.556	-3703
Analfabetismo					-3362
Nivel Ed					823
observaciones	3224	3224	3224	3224	3224
R^2	0	0	0.04	0.052	0.0516

Podemos ver que tanto los modelos 1 y 2 tienen un R^2 de 0. A partir de la variable Educación (entendido como máximo nivel educativo alcanzado, CH12), la cual tiene una relación positiva con el salario semanal aumenta R^2. Un mayor se ve al añadirse la variable de género (en este casó mujer), teniendo esta una relación negativa con respecto al salario semanal. Lo mismo ocurre con una de las variables escogidas (analfabetismo), esto significa que al aumentar este, el salario semanal disminuye.

Y por último el modelo 5, el más completo incluye la variable Nivel educativo (entendido esta vez como máximo nivel educativo completado), esto se distingue de educación (como se veía antes) ya que esta toma en cuenta simplemente los niveles educativos terminados, aunque esta diferencia parezca menor, podemos ver como al agregar esta variable y la consideración del analfabetismo, el modelo 5 toma mucha más significancia.

Cabe aclarar que en todos los modelos el número de observaciones fue el mismo y se puede ver una evolución progresiva, modelo a modelo, en el R^2.

3) Tabla 2: Año 2004

Var. Dep: *salario_semanal*	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
	(1)	(2)	(3)	(4)	(5)
MSE test	1,15	1,11	9,81	9,00	9,00
RMSE test	10.738,80	10.543,20	9.909,22	9490,61	9490,61
MAE test	6.910,78	6.582,75	6.534,51	6.100,01	6.100,01

La Tabla muestra el desempeño de cinco modelos de regresión lineal para predecir el salario semanal. A medida que se incorporan más variables explicativas, se observa una mejora progresiva en las métricas de error:
El MSE aumenta de 1,15 en el Modelo 1 a 9,00 en los Modelos 4 y 5.
Pero el RMSE disminuye sustancialmente, de 10.738,80 a 9.490,61.
El MAE baja desde 6.910,78 a 6.100,01.
Sin embargo, el Modelo 5 no mejora respecto al Modelo 4, lo que indica que la última variable agregada no aporta poder predictivo adicional.
Los Modelos 4 y 5 parecen ser los más eficientes, logrando un buen equilibrio entre complejidad del modelo y capacidad predictiva.

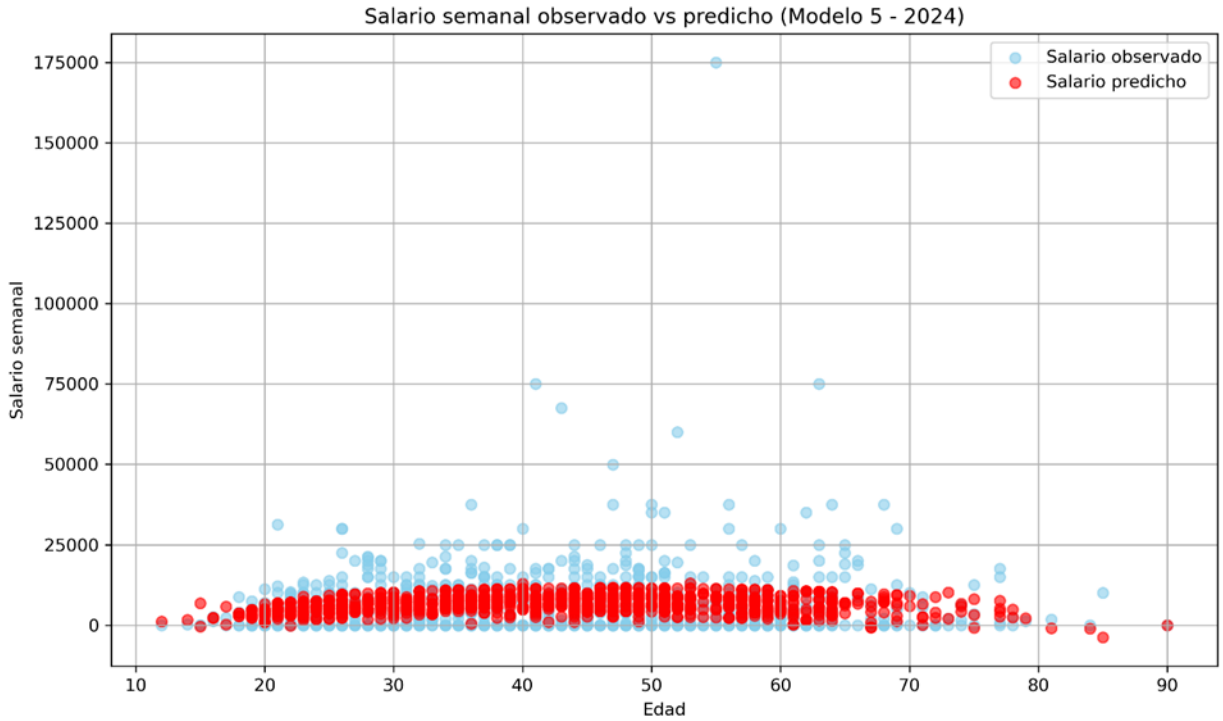
Tabla3: Año 2024

Var. Dep: *salario_semanal*	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
	(1)	(2)	(3)	(4)	(5)
MSE test	97.337.918	96.728.951	92.614.507	91.613.815	91.613.815
RMSE test	9.866,00	9.853,09	9.623,64	9.571,51	9.571,51
MAE test	6.014,12	5.947,48	5.727,75	5.657,40	5.657,40

La tabla muestra el desempeño de cinco modelos de regresión lineal para predecir el salario semanal en 2024, evaluados con métricas de error sobre la base de testeo.
El MSE disminuye significativamente del Modelo 1 (97.337.918) al Modelo 4 (91.613.815), lo que indica que añadir variables mejora la precisión del modelo.
El RMSE pasa de 9.866,00 a 9.571,51 en el mismo recorrido.
El MAE también mejora: de 6.014,12 a 5.657,40.

Sin embargo, del Modelo 4 al Modelo 5 no hay mejora alguna: los valores de MSE, RMSE y MAE se mantienen exactamente iguales. Esto sugiere que la última variable añadida no aporta valor predictivo adicional.
Los Modelos 4 y 5 representan el mejor balance entre complejidad y rendimiento predictivo para el año 2024.

4) Gráfico de la predicción de salarios del año 2024



El gráfico muestra la relación entre la edad y el salario semanal, comparando los valores observados (en celeste) y predichos por el Modelo 5 (en rojo). Los salarios observados tienen una dispersión mucho mayor, con varios valores extremos (outliers) que superan ampliamente los $50.000 y en algunos casos incluso los $150.000. En contraste, los salarios predichos se encuentran agrupados en una banda mucho más estrecha y conservadora, sin capturar los extremos. Sin embargo, el modelo subestima sistemáticamente los valores altos del salario. El Modelo 5 para 2024 captura de forma general la tendencia entre edad y salario, pero tiene limitaciones claras en su capacidad para predecir salarios altos.

5)

Results: Logit

Model:	Logit	Method:	MLE
Dependent Variable:	desocupado	Pseudo R-squared:	0.037
Date:	2025-06-03 11:07	AIC:	2672.2862
No. Observations:	5344	BIC:	2698.6211
Df Model:	3	Log-Likelihood:	-1332.1
Df Residuals:	5340	LL-Null:	-1383.5
Converged:	1.0000	LLR p-value:	3.9615e-22
No. Iterations:	10.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
ch09	-3.3481	0.7056	-4.7453	0.0000	-4.7310	-1.9652
nivel_ed	0.1521	0.0322	4.7300	0.0000	0.0891	0.2152
ch04	0.1966	0.1078	1.8237	0.0682	-0.0147	0.4079
intercepto	0.1123	0.7251	0.1549	0.8769	-1.3089	1.5335

Resultados de la regresión logit en el año 2004

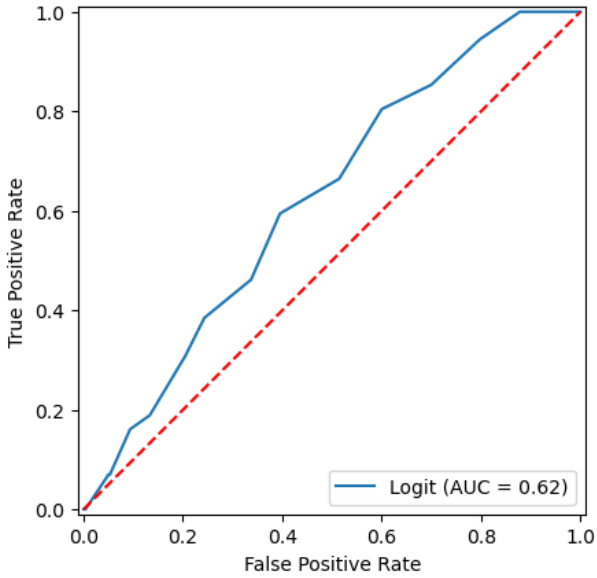
Results: Logit

Model:	Logit	Method:	MLE
Dependent Variable:	desocupado	Pseudo R-squared:	0.020
Date:	2025-06-03 11:08	AIC:	1706.9356
No. Observations:	4907	BIC:	1732.9292
Df Model:	3	Log-Likelihood:	-849.47
Df Residuals:	4903	LL-Null:	-867.22
Converged:	0.0000	LLR p-value:	9.5672e-08
No. Iterations:	35.0000	Scale:	1.0000

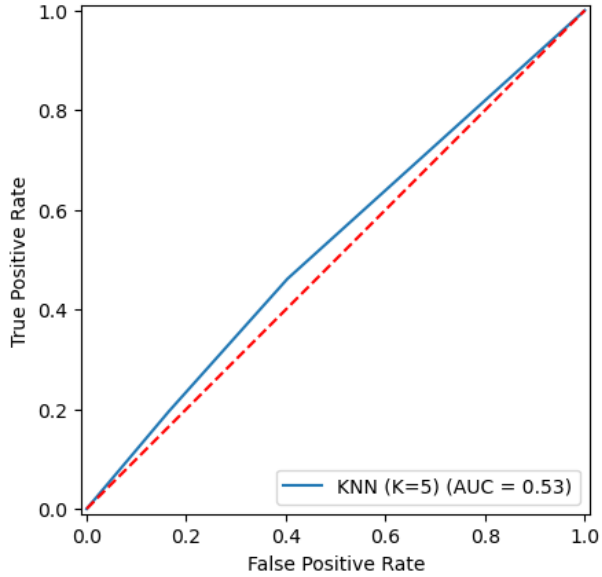
	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
CH09	-13.8479	268.1216	-0.0516	0.9588	-539.3566	511.6607
CH04	0.0234	0.1417	0.1651	0.8689	-0.2543	0.3011
NIVEL_ED	0.1220	0.0451	2.7033	0.0069	0.0335	0.2104
intercepto	10.3059	268.1217	0.0384	0.9693	-515.2029	535.8147

Resultados de la regresión logit en el año 2024

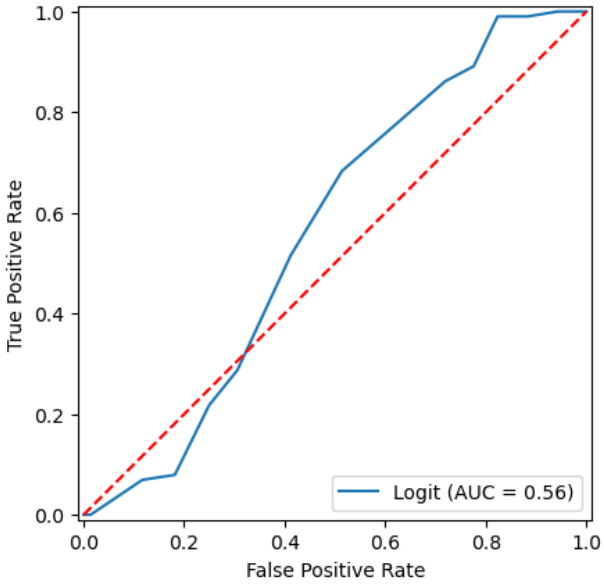
	Accuracy	AUC score	Matriz de confusión
Logit (2004)	0.937582	0,62	$\begin{pmatrix} 2148 & 0 \\ 143 & 0 \end{pmatrix}$
KNN (K=5) (2004)	0.938	0.5305	$\begin{pmatrix} 0 & 2148 \\ 1 & 143 \end{pmatrix}$
Logit (2024)	0.951973	0.56	$\begin{pmatrix} 2002 & 0 \\ 101 & 0 \end{pmatrix}$
KNN (K=5) (2024)	0.952	0.5	$\begin{pmatrix} 0 & 2002 \\ 1 & 101 \end{pmatrix}$



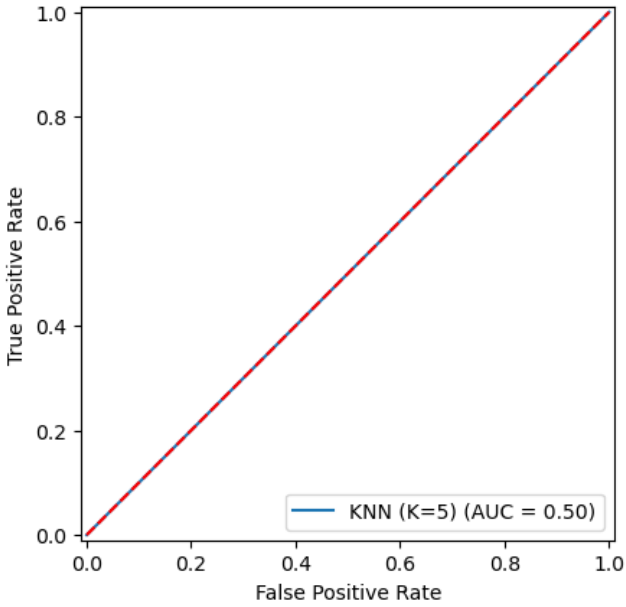
Curva ROC Logit del año 2004



Curva ROC KNN del año 2004



Curva ROC Logit del año 2024



Curva ROC KNN del año 2024

Dependiendo el año podemos ver diferencias entre cada uno de los modelos, así que vamos a examinarlos por año:

Para el año 2004 vemos que tanto el modelo Logit como el modelo de vecinos cercanos tienen una accuracy (precisión o exactitud) muy cercana, por no decir igual, lo que nos dice que clasifican correctamente la misma proporción de observaciones, ósea que su proporción de la suma de true positives y true negatives sobre los positivos y negativos da muy parecido ((TP+TN) /(P+N)).
Por lo que pasaremos a la siguiente medida de precisión, la AUC, o Area under curve, que nos muestra el área bajo la curva de los scores predichos, en este caso tenemos un AUC score de 0,62 para logit y de 0,5305 para KNN. Esto nos dice que nuestro modelo logit suele

discriminar más entre ocupados y desocupados, los que nos sirve, por que significa que el modelo no es totalmente aleatorio (como seria si es 0,5 o muy cercano) y que tenemos mejor capacidad de predicción, todo esto, puede notarse en las curvas ROC del año 2004, que representan justamente esta capacidad gráficamente, por lo que, en conclusión, el Modelo logit es mejor para predecir en el año 2004.

Por otro lado, para el año 2024, vemos que tenemos un caso similar, pero más reñido, donde vemos que hay una accuracy de 0,9519 para el modelo logit y una accuracy de 0,952 para KNN, por lo que podemos decir que su accuracy es la misma. En cuanto a la AUC, vemos que los modelos de logit y KNN tienen un score de 0,56 y 0,5 respectivamente, por lo que vemos que el modelo logit tiene una pequeña capacidad de discriminar entre ocupados y desocupados, mientras que en KNN vemos que esta capacidad no está, es totalmente aleatorio y, por lo tanto, no es viable usarla para predecir, como se puede ver en la curva ROC donde las líneas se superponen. Por lo tanto, el modelo de regresión logit es el mas ideal para predecir en el año 2024.

6) con el modelo logit, en el año 2004, pudimos predecir que 0,69 personas están desocupadas, por lo que en proporción a las 10 observaciones que tiene la base, sería un 6,9% de personas desocupadas.

En el año 2024, pudimos predecir que 1,46 personas están desocupadas, por lo que en proporción a las 41 observaciones que tiene la base, sería un 3,6% de personas desocupadas.