Big Data and Machine Learning: Trabajo Práctico 4.

Integrantes:

- Gabriela Yael Armoa (898589).
- Kevin Luque Benegas (915007).
- Carlos Ezequiel Martinez (911822).

Parte A: Enfoque de validación.

1. *a. Tabla de diferencia de medias*

Base 2004 - Región GBA.

	Media Train	Media Test	Diferencia
ano4	2004.00	2004.00	0.00
ch06	34.03	34.55	-0.53
p21	255.46	282.25	-26.78
pp03d	0.07	0.07	0.00
p21_ajustado	263614.30	291253.90	-27639.60
salario_semanal	6590.36	7281.35	-690.99
edad2	1657.88	1705.35	-47.46
educ	10.96	10.98	-0.02
estado_desocupado	0.07	0.06	0.01
estado_ocupado	0.41	0.40	0.01

La comparacion entre la base de entrenamiento y la de prueba muestra que, en general, las medias de las variables explicativas son bastante similares, lo que indica que la participacion aleatoria no genero un sesgo significativo entre los subconjuntos.

Variables como educación presentan diferencias practicamente nulas (-0,02), lo cual refuerza la idea de que las muestras son comparables. En los indicadores laborales como estado_desocupado y estado_ocupado, las diferencias son muy pequeñas (0,01), lo cual es importante dado que estamos modelando la condicion desocupacion.

Las diferencias más grandes se observan en variables como p21_ajustado (-27.639,6), salario_semanal (-690,99) y edad (-47,46), lo cual podría deberse a la variabilidad natural de estas variables continuas. Sin embargo, al analizar los valores en relación con sus escalas (por ejemplo, una

media de ingreso semanal de más de 6.000), estas diferencias no parecen lo suficientemente grandes como para afectar la validez del entrenamiento del modelo.

No se observan diferencias sistematicamente relevantes entre los conjuntos, por lo que se puede avanzar con el analisis confiado en que la muestra de entrenamiento es representativa del total.

Base de 2024 - Región GBA.

	Media Train	Media Test	Diferencia
ano4	2024.00	2024.00	0.00
ch06	37.04	36.90	0.14
p21	140602.39	135519.71	5082.68
cat_ocup	1.13	1.15	-0.01
pp03d	0.24	0.23	0.01
salario_semanal	3515.06	3387.99	127.07
edad2	1934.72	1936.32	-1.60
educ	10.64	10.54	0.10
estado_desocupado	0.05	0.05	-0.00
estado_ocupado	0.37	0.38	-0.01

Al comparar las medias de las variables entre la base de entrenamiento y la de prueba para el año 2024, se observa que las diferencias son en general muy pequeñas, lo que indica que generó dos muestras comparables.

Variables como estado_desocupado y estado_ocupado muestran diferencias nulas o practicamente nulas, lo cual es importante ya que la variable dependiente ‘desocupado’ está bien representada en ambos subconjuntos. Las variables cat_ocup, pp03d y educ presentan diferencias mínimas (entre -0,01 y 0,10), lo que sugiere que no hay un sesgo sistemático en la composición ocupacional o educativa de los grupos.

En variables continuas como p21 (ingreso), salario_semanal y edad2, las diferencias son un poco más notorias pero siguen siendo acotadas. Por ejemplo, la media de ingreso (p21) es levemente mayor en la muestra de entrenamiento (una diferencia de alrededor de $5.000), lo mismo ocurre con salario_semanal, aunque la diferencia también es menor al 4% del valor promedio. Estas variaciones son esperables dado el carácter continuo y disperso de estas variables, y no comprometen la representatividad del conjunto de entrenamiento.

En síntesis, la distribución de las características en ambas muestras es similar, lo cual garantiza que los resultados del modelo entrenado podrán generalizarse adecuadamente al conjunto de testeo.

Parte B: Método Supervisado I – Modelo de Regresión Lineal.

2. *b. Tabla de Estimación por regresión lineal de salarios usando la base de entrenamiento*

Base del 2004 - Región GBA.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
edad	228.743** (89.71)	2189.080*** (529.01)	1839.308*** (518.19)	1988.343*** (515.84)	2068.735*** (516.32)
edad2		-22.644*** (6.02)	-18.098** (5.91)	-19.718*** (5.88)	-20.260*** (5.88)
educ			2257.449*** (277.01)	2385.458*** (276.71)	2449.872*** (278.34)
Mujer				-10111.051*** (2321.48)	-10385.397*** (2327.41)
Variable 1					-3297.703* (1825.49)
Variable 2					5010.905* (2697.25)
N (obs)	1327	1327	1327	1327	1327
R^2	0.005	0.015	0.062	0.076	0.080

Sobre la base de entrenamiento de *ocupados en 2004*, utilizando como variable dependiente el *salario semanal*, a medida que se incorporan variables explicativas, el R^2 aumenta de forma progresiva (de 0.005 en el modelo 1 a 0.080 en el modelo 5), lo que indica una mejora en la capacidad predictiva del modelo.

En el *modelo 1*, se incluye solo la variable *edad*, que tiene un coeficiente positivo y significativo: cada año adicional de edad se asocia con un aumento promedio de $228,743 en el salario semanal. Sin embargo, el modelo explica muy poco de la variabilidad del ingreso ($R^2 = 0.005$).

El *modelo 2* incorpora *edad2*, lo que permite capturar una relación no lineal entre edad e ingresos. El coeficiente de edad aumenta y el de edad2 es negativo y significativo, lo que sugiere rendimientos crecientes a edades tempranas y decrecientes en edades más avanzadas. El R^2 mejora levemente a 0.015.

En el *modelo 3* se agrega *educ*, una de las variables más relevantes: cada año adicional de educación se asocia con un aumento de $2.257,449 en el salario, con alta significancia estadística. Esto refleja el retorno económico del capital humano. El R^2 sube a 0.062.

El *modelo 4* incorpora la variable mujer, que resulta negativa y significativa: en promedio, las mujeres ganan $10.111,05 menos que los varones, controlando por edad y educación. Este resultado evidencia la existencia de una brecha salarial de género. El R^2 del modelo es 0.076.

En el *modelo 5*, se suman dos variables adicionales: *'pp03d'*, que indica la cantidad de ocupaciones que tiene la persona, y '*cat_ocup_obrero_empleado'*, una dummy que vale 1 si el individuo es obrero o empleado (es decir, está en una posición común de relación de dependencia). La primera tiene un coeficiente negativo y significativo al 10%: tener más de una ocupación se asocia con un menor

salario semanal en la ocupación principal, lo cual podría estar reflejando que quienes tienen múltiples empleos acceden a trabajos peor remunerados individualmente. La segunda variable tiene un coeficiente positivo y también significativo al 10%, lo cual sugiere que estar en una relación de dependencia se asocia con mayores ingresos que otros tipos de inserción laboral (como cuentapropismo o empleos informales).

El modelo 5 tiene el mejor poder explicativo de los cinco ($R^2 = 0.080$), aunque sigue siendo moderado, lo que indica que todavía hay factores no observados que influyen en los salarios.

Base del 2024 - Región GBA.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
edad	53.185** (20.50)	646.248*** (110.86)	617.487*** (106.22)	652.079*** (104.24)	659.493*** (104.49)
edad2		-6.709*** (1.23)	-6.168*** (1.18)	-6.554*** (1.16)	-6.633*** (1.16)
educ			863.312*** (72.45)	938.651*** (71.66)	942.847*** (71.78)
Mujer				-4306.048*** (538.91)	-4262.548*** (540.58)
Variable 1					-371.109 (363.07)
Variable 2					0.000 (0.00)
N (obs)	1572	1572	1572	1572	1572
R^2	0.004	0.023	0.104	0.139	0.140

Para ocupados en *el año 2024*, usando como variable dependiente el *salario semanal*, a medida que se incorporan más variables explicativas, el poder predictivo del modelo mejora, reflejado en un aumento del R^2 de 0.004 en el modelo 1 a 0.140 en el modelo 5.

En el *modelo 1*, se observa una relación positiva entre *edad* y el salario semanal: cada año adicional se asocia con un aumento de $53,185 en promedio, con significancia al 5%. Sin embargo, el R^2 es muy bajo, lo que indica una escasa capacidad explicativa.

En el *modelo 2*, se incluye *edad2* para capturar una relación no lineal. El coeficiente de edad sube y el de edad2 es negativo y altamente significativo, lo cual sugiere que el salario crece con la edad, pero a un ritmo decreciente, como suele esperarse en mercados laborales. El R^2 mejora a 0.023.

El *modelo 3* incorpora la variable *educ*, cuyo coeficiente es positivo y altamente significativo: cada año adicional de educación se asocia con un aumento de $863,312 en el salario semanal. Esta variable aporta notablemente al modelo, elevando el R^2 a 0.104, lo que evidencia la relevancia del capital humano en la determinación del ingreso.

En el *modelo 4* se suma la variable *mujer*, una dummy que toma valor 1 si la persona es mujer. El coeficiente es negativo y significativo al 0.1%: en promedio, las mujeres ganan alrededor de $4.300 menos que los varones, manteniendo constantes edad y educación. La inclusión de esta variable eleva el R^2 a 0.139, sugiriendo que parte de la variabilidad del salario está asociada al género.

En el *modelo 5*, se incorporan dos variables adicionales *'pp03d'* (cantidad de ocupaciones que tiene la persona) y *'cat_ocup_obrero_empleado'* (dummy que indica si es obrero o empleado en relación de dependencia). A diferencia de 2004, en este caso pp03d no resulta estadísticamente significativa, aunque su signo negativo se mantiene, lo que podría estar reflejando una relación débil o heterogénea

entre la multiocupación y los ingresos. En cuanto a *cat_ocup_obrero_empleado*, su coeficiente es exactamente cero, lo que podría indicar un problema en la codificación o que la variable no aporta información adicional al modelo en esta muestra. El R^2 se mantiene prácticamente igual al del modelo anterior (0.140).

Los *resultados de 2024* confirman el impacto positivo de la edad (aunque decreciente) y la educación sobre el salario, así como la persistencia de una brecha de género. A diferencia de *2004*, las variables adicionales no mejoran sustancialmente la capacidad explicativa del modelo, lo que podría deberse a cambios estructurales en el mercado laboral o a menor relevancia de estas variables en la muestra de ese año.

3.

En este punto aplicamos un enfoque de validación para evaluar los cinco modelos, se utilizaron las observaciones del conjunto de testeo. Después se predijo el salario semanal sobre ese conjunto de test y se calcularon las tres métricas de testeo: MSE, RMSE y MAE.

Lo que nos muestran los resultados es que a medida que agregamos variables relevantes, el error tiende a bajar, lo que era esperable. Además, al comparar los dos años, se nota que en 2024 los errores son algo más bajos.

4.

En este punto graficamos la predicción del salario semanal en un gráfico de dispersión. En el gráfico mismo se observa la relación entre edad y salario semanal.

El gráfico muestra que hay dispersión, pero el modelo puede llegar a lograr capturar la tendencia general. Al comparar ambos años, se puede ver que el error de predicción es más bajo que en 2024

Gráfico 2004 - Región GBA.

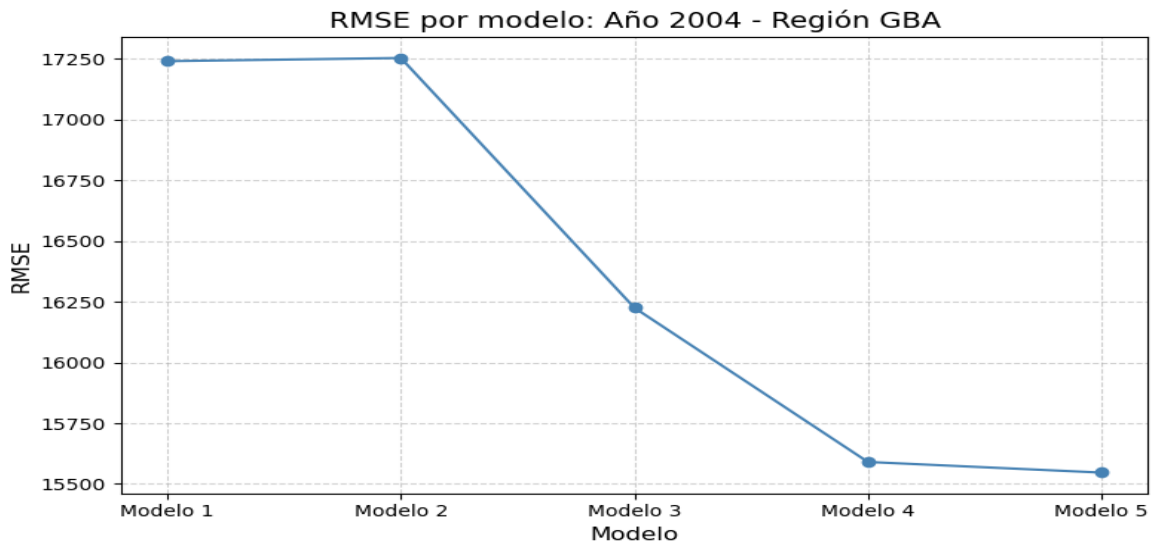
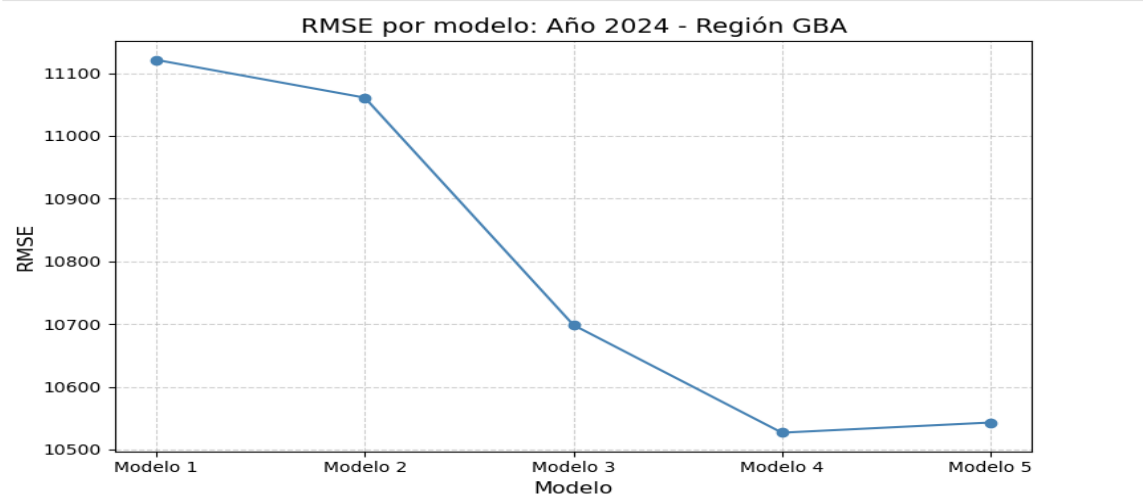


Gráfico 2024 - Región GBA.

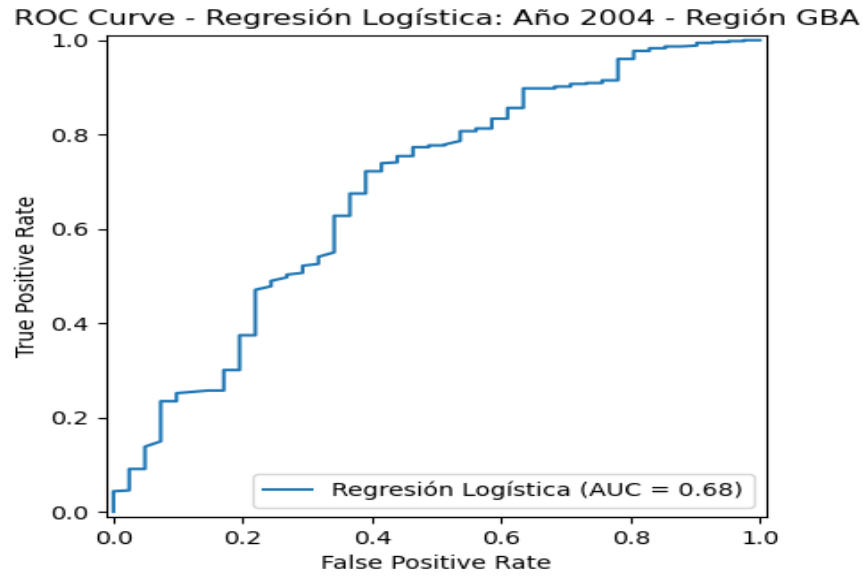


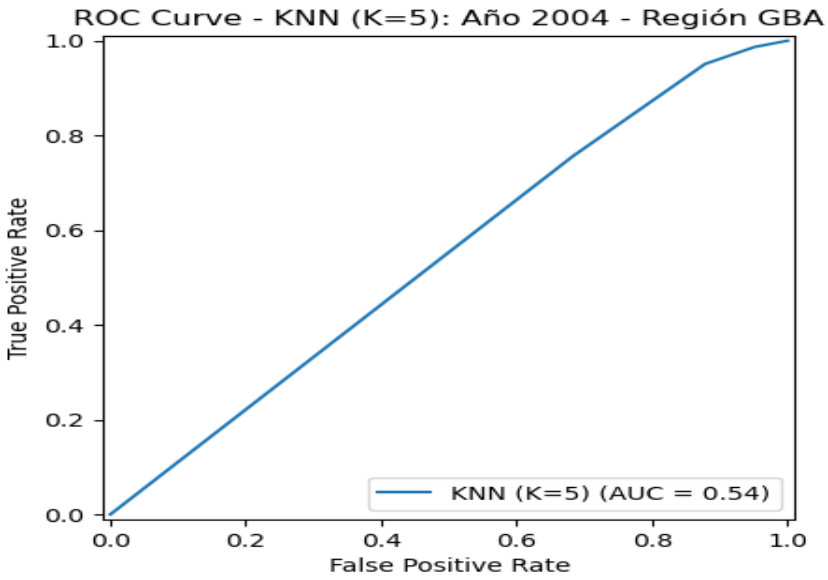
Parte C: Métodos de Clasificación y Performance.

5.

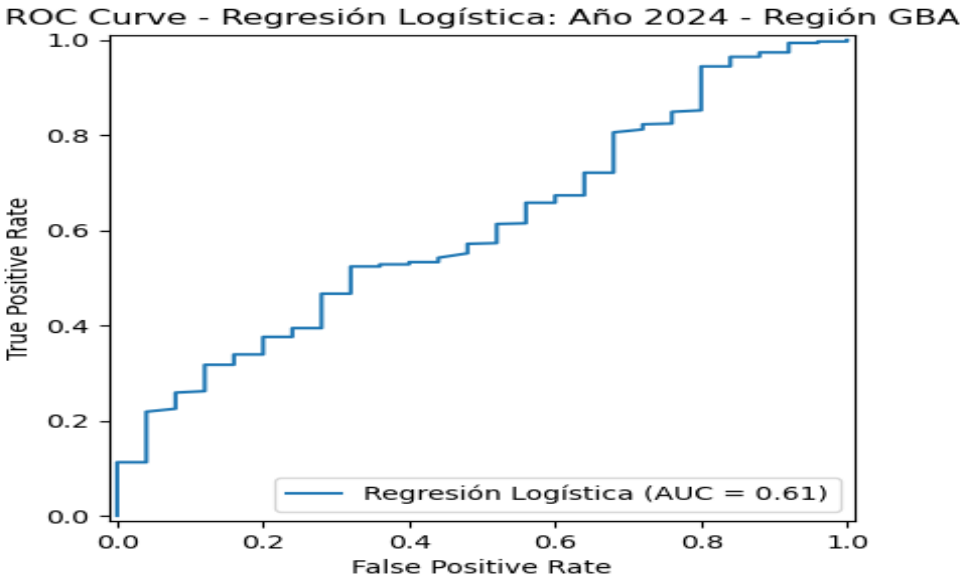
Podremos observar e identificar a través de los gráficos que el mejor método es sin ninguna duda la "Regresión logística" ya que nos da números más altos en "Precisión" y "AUC" en comparación con el método de "KNN" para ambos años (2004 y 2024), esto se debe a que regresión logística es un modelo paramétrico que puede calcular los coeficientes para cada una de las variables predictoras, mientras que KNN es lo contrario.

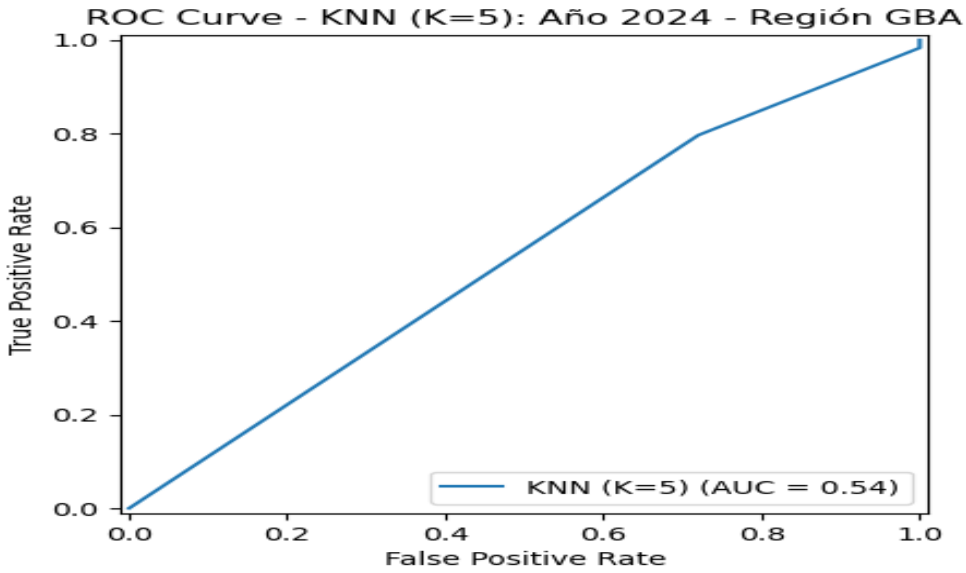
Gráficos 2004 - Región GBA.





Gráficos 2024 - Región GBA.





6.

En este caso observaremos algo muy interesante, en 2024 tenemos los datos y podemos observar los números y el porcentaje de personas son desocupadas dentro de la base norespondieron, los cuales nos indican que de 101 personas que no respondieron 41 son desocupadas, es decir el 40.59%. Pero, por otro lado, podemos observar algo interesante en 2004, simplemente no hay datos, o por lo menos no hay suficientes datos completos que nos indiquen lo que buscamos, esto se debe a la manera en que la base de datos de 2004 está construida originalmente, y que se cambió en retrospectiva con el 2024 para si poder obtener estos datos.