# Big Data & Machine Learning - Trabajo Práctico N°4

**Grupo:** 1
**Integrantes:**

-Benjamin Rodolfo Ayay Quispe (N° Registro 912662)

-Jesús David Ochochoque Mendoza (N° Registro 915690)

## Parte A: Enfoque de validación

### 1. Tabla de diferencia de medias

**Tabla 1** - Región seleccionada: Gran Buenos Aires

| Tabla de diferencia de medias para el año 2004 | | | |
|---|---|---|---|
| | Diferencia de medias | Media Train | Media Test |
| const | 0.00 | 1.00 | 1.00 |
| CH06 | -0.86 | 26.28 | 27.14 |
| edad2 | -70.09 | 1114.74 | 1184.83 |
| educ | -0.04 | 2.76 | 2.80 |
| salario_semanal | -859.75 | 11053.53 | 11913.28 |
| horastrab | -0.51 | 11.76 | 12.27 |
| Tabla de diferencia de medias para el año 2024 | | | |
| | Diferencia de medias | Media Train | Media Test |
| const | 0.00 | 1.00 | 1.00 |
| CH06 | -0.01 | 38.83 | 38.84 |
| edad2 | -34.61 | 1994.74 | 2029.35 |
| educ | 0.14 | 10.61 | 10.47 |
| salario_semanal | 173.58 | 3223.12 | 3049.54 |
| horastrab | 0.67 | 20.18 | 19.51 |

**En la tabla 1** se puede ver cómo se construyeron las variables clave a partir de la base respondieron, utilizando tanto las observaciones del año 2004 como las del año 2024. Se recrearon las variables edad2, educ, salario_semanal, horastrab y mujer, de forma coherente con lo realizado en el TP3, a partir de las variables originales de la EPH (CH06, CH12, CH13, CH14, P21, PP3E_TOT, PP3F_TOT, etc.). A continuación, se separaron las bases por año y se aplicó un muestreo aleatorio estratificado (train/test split) del 70% y 30% respectivamente, utilizando la semilla 444 para garantizar reproducibilidad. Este procedimiento se llevó a cabo por separado para 2004 y 2024, con el objetivo de evitar que sesgos en la muestra entrenada se trasladen al modelo. Es así que se construyeron tablas de diferencias de medias entre los conjuntos de entrenamiento y prueba para verificar la homogeneidad entre ambos. Los resultados muestran diferencias muy pequeñas para todas las variables, en ambos años, lo cual valida la calidad del muestreo aleatorio aplicado. Por consiguiente, en 2004, las diferencias entre train y test son inferiores a 1 año en edad y educación, menos de una hora en horastrab, y una diferencia moderada en salario_semanal, esperable por la dispersión de ingresos. En 2024, las diferencias son incluso menores, con una pequeña variación positiva en educ y horastrab, y una leve diferencia en salario_semanal, consistente con la mayor heterogeneidad del mercado laboral en ese año. Por último, es importante destacar que los patrones observados en la Parte A son coherentes con los hallazgos del TP3, donde se había identificado una relación

entre educación, salario e inserción laboral. La calidad del muestreo logrado en esta parte sienta una base sólida para continuar con los modelos supervisados en las siguientes secciones.

# Parte B: Método Supervisado 1: Modelo de Regresión Lineal

## 2. Estimación por regresión lineal de salarios usando la base de entrenamiento

**Tabla 2** - Región seleccionada: Gran Buenos Aires

| Var. Dep: *salario_semanal* | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 |
|---|---|---|---|---|---|
| Variables | (1) | (2) | (3) | (4) | (5) |
| *edad* | 24.391(16.33) | 436.624(90.43) | 407.916(88.51) | 442.606(87.91) | 412.282(76.96) |
| *edad2* | | -4.669(1.01) | -4.197(0.99) | -4.581(0.98) | -4.420(0.86) |
| *educ* | | | 645.813(64.60) | 704.944(64.73) | 427.267(57.67) |
| *Mujer* | | | | -2821.665(449.44) | -2442.524(393.69) |
| *Variable 1* | | | | | -1.781(2.91) |
| *Variable 2* | | | | | 0.010(0.00) |
| N (observaciones) | 2199 | 2199 | 2199 | 2199 | 2199 |
| $R^2$ | 0.001 | 0.011 | 0.054 | 0.070 | 0.288 |

**En la tabla 2,** se puede ver que se estimaron cinco modelos de regresión lineal para explicar el salario semanal de las personas ocupadas en el Gran Buenos Aires en el año 2024. Se trabajó exclusivamente sobre la base de entrenamiento, obtenida previamente mediante un muestreo aleatorio del 70% con semilla 444, y se usaron como variables explicativas aquellas construidas y validadas en los trabajos anteriores (TP2 y TP3). El Modelo 1, que utiliza solamente la variable edad, muestra un poder explicativo prácticamente nulo ($R^2 \approx 0.001$), lo cual indica que, en forma aislada, la edad no es un buen predictor del salario. El Modelo 2 incorpora edad² y permite capturar una relación no lineal entre edad y salario. Aquí se observa una mejora leve del ajuste ($R^2 \approx 0.011$), y los coeficientes sugieren una curva salarial con rendimientos crecientes hasta cierto punto, seguida de rendimientos decrecientes, algo esperable en mercados laborales reales. El Modelo 3 incorpora la variable educ, que representa los años de educación según el máximo nivel alcanzado y los años cursados. Esta variable resulta altamente significativa y mejora notablemente el poder explicativo del modelo ($R^2 \approx 0.054$), en línea con lo observado en el TP3, donde los niveles de educación mostraban clara correlación con el ingreso y la situación laboral. En el Modelo 4, se incorpora la variable mujer, una dummy que toma el valor 1 si la persona es mujer. El coeficiente estimado para mujer es negativo y altamente significativo, lo cual indica la existencia de una brecha salarial de género dentro de la población ocupada, también consistente con los análisis previos del TP3. Finalmente, el Modelo 5 incorpora dos variables adicionales: horastrab (horas trabajadas semanales) e IPCF (ingreso per cápita familiar). Si bien horastrab no resulta significativa, posiblemente debido a errores de medición o poca variabilidad en los datos, IPCF muestra un efecto positivo y altamente significativo. Este último resultado puede interpretarse como un efecto indirecto de la inserción familiar en el mercado laboral sobre el salario individual, o una señal de mejores condiciones socioeconómicas generales. Finalmente, el Modelo 5 incorpora dos variables adicionales: horastrab (horas trabajadas semanales) e IPCF (ingreso per cápita familiar). Si bien horastrab no resulta significativa, posiblemente debido a errores de medición o poca variabilidad en los datos, IPCF muestra un efecto positivo y altamente significativo. Este último resultado puede interpretarse como un efecto indirecto de la inserción familiar en el mercado laboral sobre el salario individual, o una señal de mejores condiciones socioeconómicas generales. Este último modelo logra un $R^2$ de 0.288, lo cual representa una mejora sustancial respecto a los modelos anteriores y sugiere que la inclusión de variables contextuales (como IPCF) aporta información valiosa al análisis salarial.

## 3. Enfoque de validación

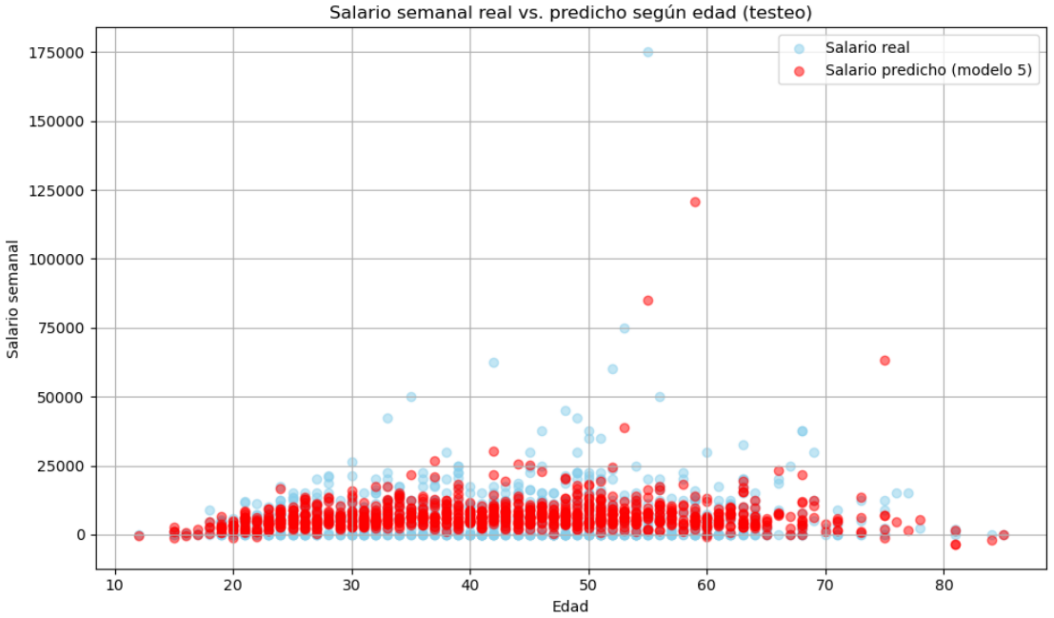
**Tabla 3** - Región seleccionada: Gran Buenos Aires

| Var. Dep: *salario_semanal* | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| *MSE test* | 95267261.12 | 94238747.22 | 90084619.62 | 90224631.24 | 59750630.59 |
| *RMSE test* | 9760.49 | 9707.66 | 9491.29 | 9498.66 | 7729.85 |
| *MAE test* | 5915.94 | 5860.87 | 5663.53 | 5615.58 | 4586.04 |

**En la tabla 3** vemos que se validó el desempeño predictivo de los modelos estimados en la Parte B sobre el conjunto de testeo (30% de los ocupados de 2024), utilizando como variable dependiente salario_semanal. Para cada modelo se calcularon tres métricas de error fuera de muestra: MSE (error cuadrático medio), RMSE (raíz del error cuadrático medio) y MAE (error absoluto medio). Los resultados se presentan en la Tabla 3 y muestran una mejora progresiva en el ajuste predictivo a medida que se incorporan más variables al modelo. El Modelo 1, que solo incluye edad, muestra un error promedio elevado (RMSE ≈ 9760.49), indicando que la edad, por sí sola, no es suficiente para explicar el salario. El Modelo 2 mejora levemente al incorporar edad², lo que permite capturar una relación no lineal con el ingreso, reduciendo marginalmente los errores. El Modelo 3 representa un avance importante: al incluir educ, el RMSE baja a 9491.29 y el MAE cae más de 250 pesos. Esto valida la fuerte influencia de la educación en el salario, en línea con los hallazgos del TP3. El Modelo 4, al incorporar la variable mujer, logra una leve mejora en MAE, aunque no reduce mucho el MSE. Esto indica que la variable de género permite ajustar diferencias sistemáticas de ingreso, pero no mejora sustancialmente la capacidad de predicción individual. El Modelo 5 muestra una mejora considerable en todas las métricas (RMSE cae a 7729.85, MAE baja a 4586.04). La inclusión de horastrab e IPCF potencia la capacidad predictiva del modelo. Esto sugiere que el ingreso semanal no solo está determinado por factores individuales (edad, educación, género), sino también por el contexto familiar (IPCF) y la intensidad laboral (horas trabajadas), aunque esta última no fue estadísticamente significativa en la Parte B. En conclusión, los resultados validan la progresiva mejora de los modelos al incluir más variables relevantes, y muestran que el Modelo 5 es el que presenta mejor capacidad de predicción fuera de muestra. Esta validación refuerza la coherencia interna del trabajo y la consistencia con los patrones detectados en el TP3.

## 4. Opcional

**Gráfico 1** - Región seleccionada: Gran Buenos Aires

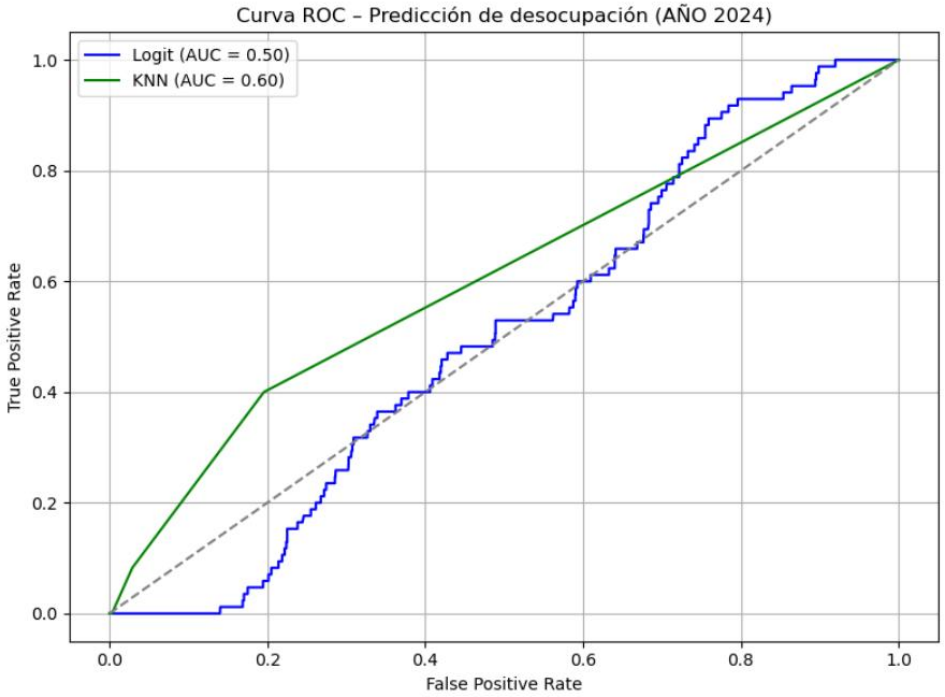



**En el gráfico 1,** se muestra la relación entre edad y salario semanal tanto para los valores reales como para las predicciones del Modelo 5, utilizando los datos de testeo. Se observa que el modelo logra capturar la tendencia general de crecimiento del salario con la edad hasta aproximadamente los 45 años, seguido de una leve caída, reflejando rendimientos decrecientes. Sin embargo, como es característico de los modelos lineales, las predicciones son más conservadoras y no logran captar completamente la gran dispersión de los ingresos reales, especialmente en edades productivas donde existen ingresos atípicos. A pesar de ello, el modelo muestra un patrón coherente con lo esperado teóricamente y valida el uso de edad y edad² en la predicción.

## Parte C: Métodos de Clasificación y Performance

## 5. Implementación de métodos

**Gráfico 2** - Región seleccionada: Gran Buenos Aires



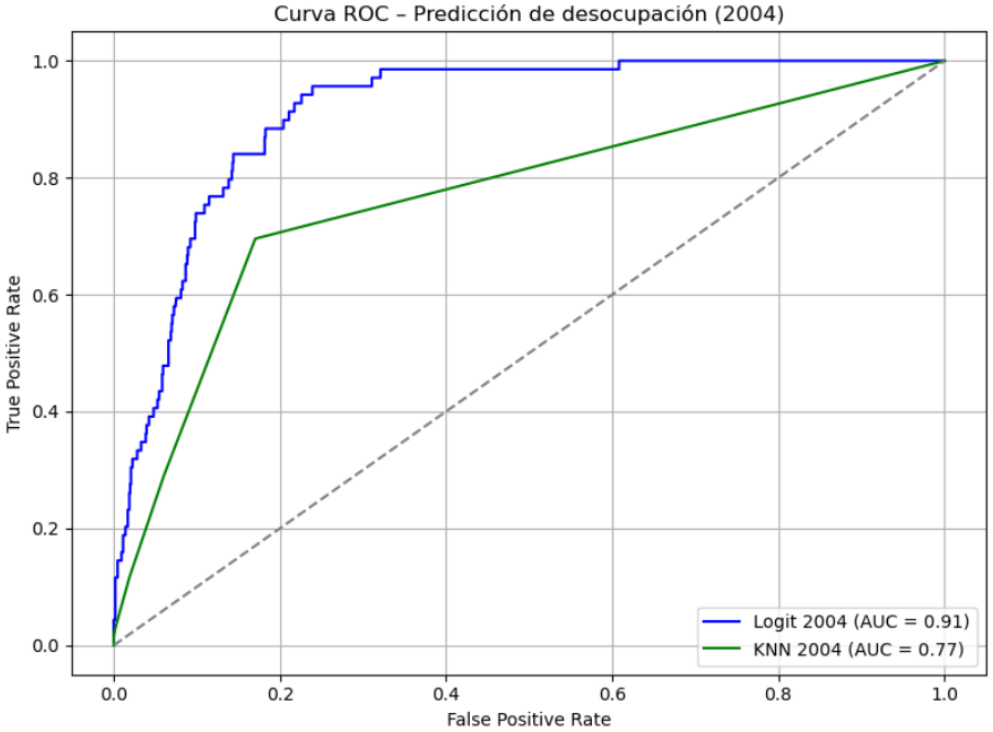

```
=== Logit ===
Matriz de confusión:
 [[1901    0]
 [  85    0]]
Accuracy: 0.957
AUC: 0.501

=== KNN ===
Matriz de confusión:
 [[1895    6]
 [  85    0]]
Accuracy: 0.954
AUC: 0.604
```

**En el gráfico 2,** se evaluaron dos modelos supervisados para predecir la condición de desocupación con base del año 2024: regresión logística (logit) y K-vecinos más cercanos (KNN, con K=5). Se entrenaron sobre la base de entrenamiento de 2024 y se evaluaron sobre la base de testeo mediante la matriz de confusión, el AUC de la curva ROC y la precisión (accuracy). Los resultados muestran que ambos modelos logran un accuracy muy alto (superior al 95%), pero esto se debe a que la clase dominante ("ocupado") representa más del 95% de los casos. En este tipo de problemas desbalanceados, la métrica de AUC resulta más útil para evaluar la capacidad real del modelo para distinguir entre clases. La regresión logística tuvo un AUC de 0.501, indicando que no predice mejor que el azar y que simplemente clasifica a todos los individuos como

ocupados. En contraste, el modelo KNN alcanzó un AUC de 0.604 y logró identificar algunos casos de desocupación, aunque de forma limitada. Por lo tanto, aunque ambos modelos presentan dificultades para predecir correctamente los casos de desocupación, el modelo KNN mostró un desempeño superior al del modelo logístico en esta base, al menos en términos de capacidad discriminante.

**Gráfico 3** - Región seleccionada: Gran Buenos Aires




```
=== LOGIT 2004 ===
Matriz de confusión:
 [[1057    5]
 [  61    8]]
Accuracy: 0.942
AUC: 0.913

=== KNN 2004 ===
Matriz de confusión:
 [[1042   20]
 [  61    8]]
Accuracy: 0.928
AUC: 0.767
```

**En el gráfico 3,** se opinará comparando con el año 2024, por ende, nuevamente se evaluaron dos modelos de clasificación supervisada, regresión logística (logit) y K vecinos más cercanos (KNN, con K=5) para predecir la probabilidad de estar desocupado. Se entrenaron y evaluaron los modelos por separado en los años 2004 y 2024, utilizando como métrica principal el AUC de la curva ROC, junto con la matriz de confusión y la precisión (accuracy) en la base de testeo. En el año 2024, ambos modelos presentaron un rendimiento muy bajo: la regresión logística obtuvo un AUC de 0.501 (equivalente al azar), y clasificó a todos los individuos como ocupados, fallando por completo en detectar casos positivos. El modelo KNN mostró un rendimiento algo mejor (AUC = 0.604), logrando identificar algunos pocos casos de desocupación. Sin embargo, el desequilibrio de clases en los datos de 2024 (menos del 5% de los casos eran desocupados) dificultó el aprendizaje de ambos modelos. En contraste, en el año 2004, ambos modelos tuvieron un desempeño mucho mejor. El modelo logit alcanzó un AUC de 0.913, con una matriz de confusión que muestra una capacidad clara de detectar desocupados (TPR alto y pocos falsos positivos). KNN también mostró un rendimiento razonable (AUC = 0.767), aunque inferior al logit. Esto sugiere que en 2004, la estructura de los datos y la relación entre variables explicativas y desocupación eran más predecibles y lineales, facilitando el aprendizaje del modelo logístico. Es así que podemos decir que, en 2024, KNN superó levemente al logit, aunque ambos modelos fallaron en la predicción de desocupación por el fuerte desbalance de clases y la menor capacidad predictiva de las variables. En 2004, la regresión logística fue claramente superior, alcanzando un AUC de 0.91 y destacándose como el modelo más efectivo para esta tarea.

## 6. Predicción con método preferente

Proporción de desocupados predichos en “norespondieron”: 17.241%

Se uso el modelo de regresión logística, sobre datos de 2004 a la base de personas que no respondieron (norespondieron.csv). La idea fue predecir la condición de desocupación en una población sobre la cual no se tenía esa información directamente. Luego de aplicar el modelo y predecir en base a las variables disponibles (edad, educación, horas trabajadas, IPCF, etc.), se encontró que el 17.2% de las personas que no

respondieron fueron clasificadas como potencialmente desocupadas. Este valor es considerablemente mayor que la tasa de desocupación observada en la base original de 2004 (~6%) y 2024 (~4–5%), lo que podría indicar que la no respuesta está asociada a condiciones más precarias del mercado laboral, como informalidad, trabajos discontinuos o directamente desempleo. Este hallazgo ilustra el valor práctico de los modelos supervisados: no solo permiten analizar lo observable, sino también inferir información relevante sobre grupos ocultos o no respondientes, permitiendo así obtener una mejor aproximación al fenómeno laboral completo.

# *Conclusión*

Este trabajo práctico aplicó técnicas de aprendizaje supervisado (regresión lineal, clasificación con logit y KNN) sobre bases de la Encuesta Permanente de Hogares (EPH), con el objetivo de analizar los determinantes del salario y la condición laboral (ocupado/desocupado) en el Gran Buenos Aires, durante los años 2004 y 2024.

En la Parte A, se logró dividir adecuadamente la muestra de cada año en bases de entrenamiento y testeo, con diferencias de medias mínimas que validan la representatividad de las muestras.

En la Parte B, se aplicaron modelos de regresión lineal crecientes en complejidad. Se confirmó que la edad, la educación y el género tienen efectos significativos sobre el salario, y que variables como IPCF mejoran considerablemente el poder predictivo del modelo.

En la Parte C, se evaluaron dos modelos de clasificación para predecir desocupación. Se encontró que, en 2024, ambos modelos mostraron baja capacidad predictiva por el fuerte desbalance de clases. En 2024, ambos modelos mostraron baja capacidad predictiva por el fuerte desbalance de clases. En 2004, el modelo logístico alcanzó un AUC de 0.91, mostrando gran poder discriminante. También aplicó el mejor modelo para estimar la condición de desocupación entre personas que no respondieron. El resultado (17.2% de predicción de desocupación) sugiere que los no respondientes podrían tener una inserción laboral más precaria.

En conjunto, el trabajo mostró cómo el uso combinado de regresión y clasificación permite no solo analizar fenómenos sociales relevantes como el empleo y el ingreso, sino también inferir patrones en datos incompletos, mejorando así la capacidad de diagnóstico y diseño de políticas públicas.

# *Referencias*


- *INDEC (2025). Encuesta Permanente de Hogares: Metodología. Disponible en: https://www.indec.gob.ar/*
- *Schwabish, J. A. (2014). An economist's guide to visualizing data. Journal of Economic Perspectives, 28(1), 209-23*