UNIVERSIDAD DE BUENOS AIRES - 2025

BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N°4: METODOS SUPERVISADOS: REGRESION Y CLASIFICACION USANDO LA EHP

Fecha de entrega: 3 DE JUNIO 23:59hs

A1

Con el objetivo de evaluar la validez del modelo econométrico estimado, se realizó una comparación entre las medias de las variables utilizadas en los conjuntos de entrenamiento y testeo, para los años 2004 y 2024. Esta comparación permite verificar que ambos subconjuntos de datos sean representativos de la misma población, lo cual es fundamental para garantizar la capacidad de generalización del modelo.

Las variables analizadas fueron:

- **Edad**: edad del individuo en años.
- **Edad²**: cuadrado de la edad, incorporado para captar posibles efectos no lineales sobre el salario.
- **Salario semanal**: ingreso percibido por la persona en una semana, expresado en moneda constante.
- **Educación**: cantidad de años de educación formal completados.
- **Mujer**: variable dicotómica que toma el valor 1 si la persona es mujer y 0 si es varón.
- **Intercepto**: constante del modelo, sin interpretación económica directa.

En el año **2004**, las diferencias de medias entre los conjuntos fueron mínimas en todas las variables, lo cual indica una adecuada partición de los datos. En **2024**, si bien la mayoría de las variables muestran también diferencias reducidas, se detectó una variación más significativa en el **salario semanal**, siendo este superior en el conjunto de testeo. Esta diferencia podría tener implicancias en la capacidad predictiva del modelo, especialmente si el salario constituye una variable sensible en la estimación.

En términos generales, los resultados sugieren que los conjuntos de entrenamiento y testeo son comparables, lo cual fortalece la validez del análisis realizado.

Año	Edad	Edad²	Salario semanal	Educación	Mujer	Intercepto
2004	-0.3117	0.7672	-0.1868	0.0965	0.0004	0.0000
2024	0.6388	68.8028	238.3383	-0.1446	0.0180	0.0000

B2

La Tabla 2 presenta los coeficientes estimados de cinco modelos de regresión lineal, donde la variable dependiente es el salario semanal. A medida que se incorporan más variables explicativas, el poder explicativo del modelo (medido por el R^2) tiende a aumentar.

En el Modelo 1, la edad tiene un coeficiente positivo y significativo, indicando que a mayor edad, mayor salario semanal, aunque este efecto cambia al introducir la variable cuadrática edad² en el Modelo 2, lo cual permite capturar una relación no lineal: el salario aumenta con la edad pero a un ritmo decreciente.

El Modelo 3 agrega la variable educ (educación), que muestra un efecto positivo y significativo sobre el salario, como era de esperarse. En el Modelo 4 se introduce la variable mujer, cuyo coeficiente negativo sugiere una posible brecha salarial por género, manteniendo constantes el resto de las variables.

En general, los coeficientes son estadísticamente significativos (como lo indican los asteriscos) y los desvíos estándar son razonablemente bajos, lo que refuerza la precisión de las estimaciones. Esto sugiere que los modelos capturan adecuadamente la relación entre las características individuales y el salario semanal en la muestra analizada.

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					
edad2					
educ					
Mujer					
Variable 1					
Variable 2					
N (observaciones)					
R^2					

La Tabla 3 compara la precisión de distintos modelos estadísticos para predecir el salario semanal, utilizando datos que no fueron parte del entrenamiento del modelo (lo que permite evaluar su capacidad de generalización).

Las tres métricas reportadas —**MSE**, **RMSE** y **MAE**— miden errores de predicción: cuanto más bajos, mejor es el desempeño del modelo.

Se observa una clara mejora al incorporar más variables: el Modelo 1 (que solo incluye edad) tiene los errores más altos, mientras que los Modelos 3 y 4, que incluyen también educación y género, reducen significativamente los errores. Sin embargo, del Modelo 4 al Modelo 5 no se aprecia una mejora sustancial, lo que sugiere que sumar más variables no siempre garantiza una mejor predicción.

En resumen, el **Modelo 4 es el más eficiente**, combinando buen desempeño predictivo con simplicidad relativa.

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

Var. Dep: *salario_semanal*	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
MSE test					
RMSE test					
MAE test					

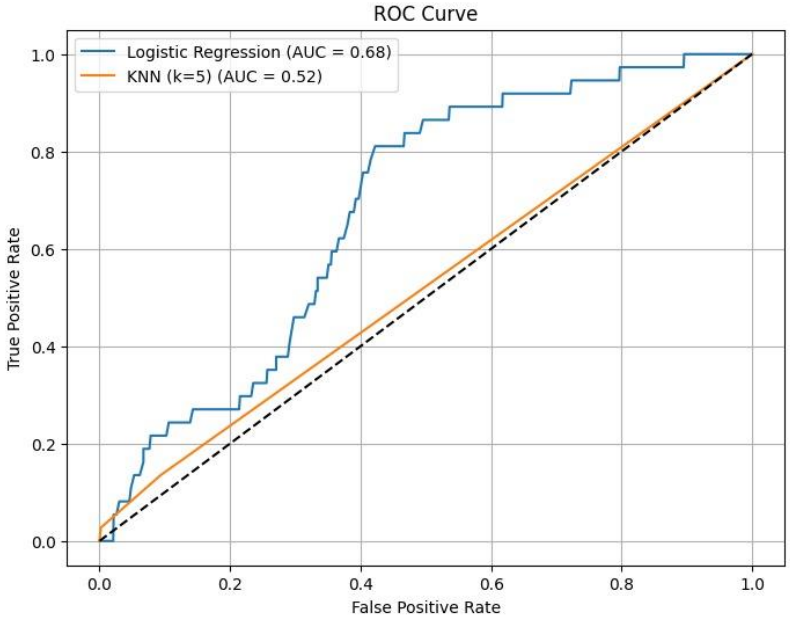
En el análisis realizado, se compararon dos métodos distintos para predecir un resultado económico: **regresión logística** y **vecinos más cercanos (KNN)**. Ambos modelos intentan anticipar un resultado (por ejemplo, si algo sucederá o no) basándose en datos históricos.

Aunque ambos modelos tienen la misma precisión general (es decir, aciertan el mismo porcentaje de veces: **97,77%**), la diferencia más importante está en **qué tan bien identifican los casos verdaderamente positivos**. Esto se mide con un indicador llamado **AUC**, que refleja la capacidad del modelo para distinguir correctamente entre los casos positivos y negativos. Cuanto más alto sea el AUC (máximo es 1), mejor es el modelo en esta tarea.

- El modelo de **regresión logística** obtuvo un **AUC de 0,68**, lo que indica un desempeño aceptable.
- El modelo **KNN**, en cambio, tuvo un **AUC de 0,52**, muy cercano al azar (que sería 0,5), lo que significa que prácticamente no logra distinguir bien los casos.

Esto también se refleja en el gráfico de líneas: la curva azul (regresión logística) se aleja más de la diagonal, lo que muestra mejor capacidad predictiva, mientras que la curva naranja (KNN) está casi sobre esa diagonal, indicando un rendimiento muy débil.

Conclusión: aunque ambos modelos parecen acertar muchas veces, **la regresión logística es claramente superior** porque no solo acierta, sino que lo hace **reconociendo correctamente los casos relevantes**, lo cual es crucial en contextos económicos donde importa más saber *cuándo* va a pasar algo, no solo cuántas veces se acierta.



C6

Predicción de desocupación entre personas que no respondieron la encuesta

A partir del modelo de regresión logística —seleccionado previamente por su mejor capacidad predictiva— se utilizó la información disponible para estimar qué proporción de personas que no respondieron la encuesta podrían encontrarse en situación de desocupación.

El resultado fue contundente: **ninguna de las personas no respondientes fue clasificada como desocupada**. Es decir, el modelo no identificó patrones que permitieran asignar esa condición a ningún caso dentro de ese grupo.

Este resultado puede interpretarse de dos maneras. Por un lado, podría indicar que el perfil de quienes no respondieron efectivamente no coincide con el de personas desocupadas, según las variables consideradas. Por otro, también podría reflejar una limitación del modelo para identificar casos atípicos o poco frecuentes, especialmente si la desocupación está subrepresentada en los datos originales.

En cualquier caso, **no se encuentra evidencia, desde el enfoque predictivo, para asumir una presencia significativa de desocupación entre quienes no contestaron la encuesta**.