Informe

Trabajo Práctico N°4

Big Data y Machine Learning

Grupo 24: Agustín Cané, Joaquín García Lucchesi,Joaquín Tesler

Facultad de Ciencias Económicas, Universidad de Buenos Aires

Junio 2025

1)Las siguientes tablas presentan, junto con información complementaria, para cada año los resultados del test-t aplicado a la comparación entre la muestra de testeo (Test) y de entrenamiento (Train), formadas a partir de segmentar la base de datos elaborada en el TP 3. Por motivos de claridad expositiva, se incluyen los resultados correspondientes a 15 variables.

Tabla 1.a: Diferencia de medias entre muestra de testeo y entrenamiento,Año 2004

Variables	Nº obs, Train	Media, Train	Std, Train	Nºobs, Test	Media, Test	Std, Test	Test-t	P-valor del test-t
Desocupado (Variable dependiente)	5352	0.067	0.251	2295	0.071	0.258	-0.643	0.52
Ingreso P.C Familiar	5352	320299.530	742581.036	2295	318582.237	382252.559	0.098	0.92
Edad	5352	33.763	22.698	2295	33.131	22.650	1.117	0.26
DECCFR	5352	4.944	2.903	2295	5.030	2.847	-1.197	0.23
RDECCFR	5352	4.650	2.867	2295	4726.	2.811	-1.066	0.28
GDEFCCR	5352	4.756	2.907	2295	4.840	2.854	-1.158	0.24
ADECCFR	5352	4.737	2.886	2295	4.805	2.845	-0.945	0.34
PP3F_TOT	5352	1.161	27.437	2295	0.463	2.981	1..217	0.22
PP3E_TOT	5352	17.057	50.135	2295	27.931	58.975	-0.661	0.50
Ingreso Ocupación Principal	5352	265.865	915.889	2295	256.839	546.958	0.439	0.66
Años de educación	5352	4.126	5.320	2295	4.019	5.312	0.802	0.42
Salario Semanal	5352	5863.791	20200.407	2295	5664.726	12063.434	0.439	0.66
Horas trabajadas p/semana	5352	15.364	23.449	2295	15.346	23.674	0.030	0.97
Es mujer	5352	0.529	0.499	2295	0.527	0.499	0.198	0.84
Sabe leer y escribir	5352	1.038	0.327	2295	1.048	0,335	-1-228	0.21
Tiene vacaciones pagas	5352	0.433	0.722	2295	0.414	0.695	1.627	0.1
Edad ^{2}	5352	1655.115	1839.209	2295	1610.501	1820.816	0.975	0.32

Tabla 1.b: Diferencia de medias entre muestra de testeo y entrenamiento,Año 2024

Variables	Nº obs, Train	Media, Train	Std, Train	Nºobs, Test	Media, Test	Std, Test	Test-t	P-valor del test-t
Desocupado (Variable dependiente)	4907	0.042	0.201	2103	0.049	0.216	-1.354	0.17
Ingreso P.C Familiar	4907	160624.514	370872.681	2103	161901.385	418887.030	-0.126	0.89
Edad	4907	38.031	22.958	2103	37.561	22.666	0.788	0.43
DECCFR	4907	7.960	4.088	2103	8.084	4.110	-1.158	0.24
RDECCFR	4907	7.882	4.111	2103	8.003	4.135	1.126	0.26
GDEFCCR	4907	7.927	4.101	2103	8.047	4.129	-1.114	0.26
ADECCFR	4907	7.884	4.109	2103	8.008	4.132	-1.148	0.35
PP3F_TOT	4907	1.565	32.025	2103	0.570	3.236	1.420	0.15
PP3E_TOT	4907	18.777	66.073	2103	18.916	60.584	-0.08	0.934
Ingreso Ocupación Principal	4907	118203.810	275008.828	2103	127931.526	38649.165	-1.20	0.22
Años de educación	4907	8.959	5.225	2103	8.954	5.157	0.03	0.97
Salario Semanal	4907	2955.095	6875.220	2103	3198.288	9516.229	-1.20	0.22
Horas trabajadas p/semana	4907	15.225	21.529	2103	16.098	22.038	-1.545	0.12
Es mujer	4907	0.524	0.499	2103	0.520	0.499	0.333	0.73
Sabe leer y escribir	4907	1.028	0.237	2103	1.037	0.259	-1.346	0.18
Tiene vacaciones pagas	4907	0.435	0.668	2103	0.434	0.658	0.027	0.97
Edad ^{2}	4907	1963.370	1962.632	2103	1924.412	1897.738	0.966	0.33

Las variables DECCFR,RDECCFR, GDECFR y ADECCFR refieren respectivamente al decil de ingreso per cápita familiar del total de la EPH, de la región ,de los aglomerados de más de 500 mil habitantes y del aglomerado particular del relevamiento. PP3E_TOT indica el total de horas trabajadas semanalmente en la ocupación principal y PP3F_TOT al total de horas trabajadas en la ocupación secundaria.

Con respecto a los datos presentados, el valor del test-t muestra la diferencia entre las medias de las dos muestras comparadas. El p-valor de este resultado marca la posibilidad de que se obtengan los resultados que muestra la tabla en caso de que la hipótesis nula sea verdadera.

La hipótesis nula para el test-t involucra que no exista una diferencia con significancia estadística entre los dos grupos de datos, lo que marca la aleatoriedad de la asignación de la información en la división entre muestras. Es por medio de evaluar el p-valor que llegamos a conocer este grado de significancia.

Los valores lejanos a 0,05 de los p-valores del test-t marcan que no hay evidencia que nos permita negar la hipótesis nula y, por lo tanto, podemos afirmar que la división entre datos de entrenamiento y de testeo es efectivamente aleatoria.

2)La Tabla 2 presenta los coeficientes obtenidos utilizando StatsModels para cada variable presente en 5 modelos de regresión lineal distintos.

Se utilizó para conformar los datos de la tabla la muestra de entrenamiento mencionada en el punto anterior.Debajo de los coeficientes se indica entre paréntesis el desvío estándar correspondiente a cada β_n estimado. Marcadas con *, ** y *** se encuentran las estimaciones que cuentan con un p-valor menor a 0.1 , 0.05 y 0.001 respectivamente.

Tabla 2: Estimación por regresión lineal utilizando la base de entrenamiento

Variable dependiente: salario_semanal	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Variables	1	2	3	4	5
edad	67,190 *** (14,88)	720,040 *** (80,13)	691,550 *** (80,91)	715,703 *** (80,28)	697,950*** (80,36)
edad^2		-7,572 *** (0,91)	-7,260 *** (0,92)	-7,5605 *** (0.91)	-7,492*** (0,91)
educ			92,540 (37,47)	121,750 ** (37,31)	100,232 ** (37,63)
Mujer				-3609,921 *** (414.22)	-3459,094 (415,19)
Tiene vacaciones Pagas					-1115,515 *** (292,19)
Sabe leer y escribir					-5896,814 (3137,94)
N(observaciones)	4412	4412	4412	4412	4412
R^2	0,5%	2%	2,1%	3,8%	4,2%

Los coeficientes en cada modelo nos indican cuál es la sensibilidad de nuestra variable dependiente a la variable independiente a la que se encuentran asignados.

Se identifica en la tabla una amplia presencia de coeficientes estimados con un p-valor menor a 0,05. Esta medida refiere a la posibilidad de obtener estos resultados si la hipótesis nula fuese cierta . Así, aunque no existiera relación entre salario_semanal y uno de los predictores, el valor del coeficiente asignado no sería igual a 0.

Sin embargo, con valores por debajo de 0.05 puede decirse que la relación entre variables es significativa y que , por lo tanto, puede rechazarse la hipótesis nula. Con esto presente podemos afirmar que la asignación de los coeficientes en cada modelo responde ,en gran medida, a relaciones efectivas entre las variables involucradas.

Puede apreciarse también un aumento en el valor del R^2 ,el porcentaje de la variación en los datos explicada por la regresión, entre los distintos modelos. Esto implica una caída en el valor de la RSS y un aumento de la capacidad predictiva de la regresión lineal dentro de la muestra a medida que se incluyen nuevas variables en el modelo.

De esta forma , podemos concluir a partir de la información presentada en la tabla que , a cuantas más variables se incorporen, más se aproximarán los resultados que se obtengan utilizando los modelos a la información de la muestra de entrenamiento, aunque esto no será una garantía de que nuestro modelo sea capaz de efectuar predicciones con el mismo grado de certeza fuera de este conjunto de datos.

3)La Tabla 3 muestra las métricas del error cuadrático medio (MSE test), Mean Absolute Error (MAE test) y el Root-Mean Squared Error (RMSE test) para cada modelo de regresión lineal visto en el punto anterior.

Para realizar la tabla, se realizó un entrenamiento previo del modelo utilizando las librerías de SciKit-Learn; se realizó una iteración de cada modelo, del 1 al 5, ordenados del modelo con menos variables presentadas (edad+intercepto) hasta el modelo más complejo.

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

Variable dependiente *salario semanal*	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Variables	1	2	3	4	5
MSE test	1,132,976,695.4	1,128,820,310.9	1,124,998,171.9	1,114,844,642.3	1,115,022,751
RMSE test	33,659.7	33,597.9	33,540.9	33,389.3	33,391.9
MAE test	8,878.4	8,867.1	8,808.2	8,520.2	8,404.6

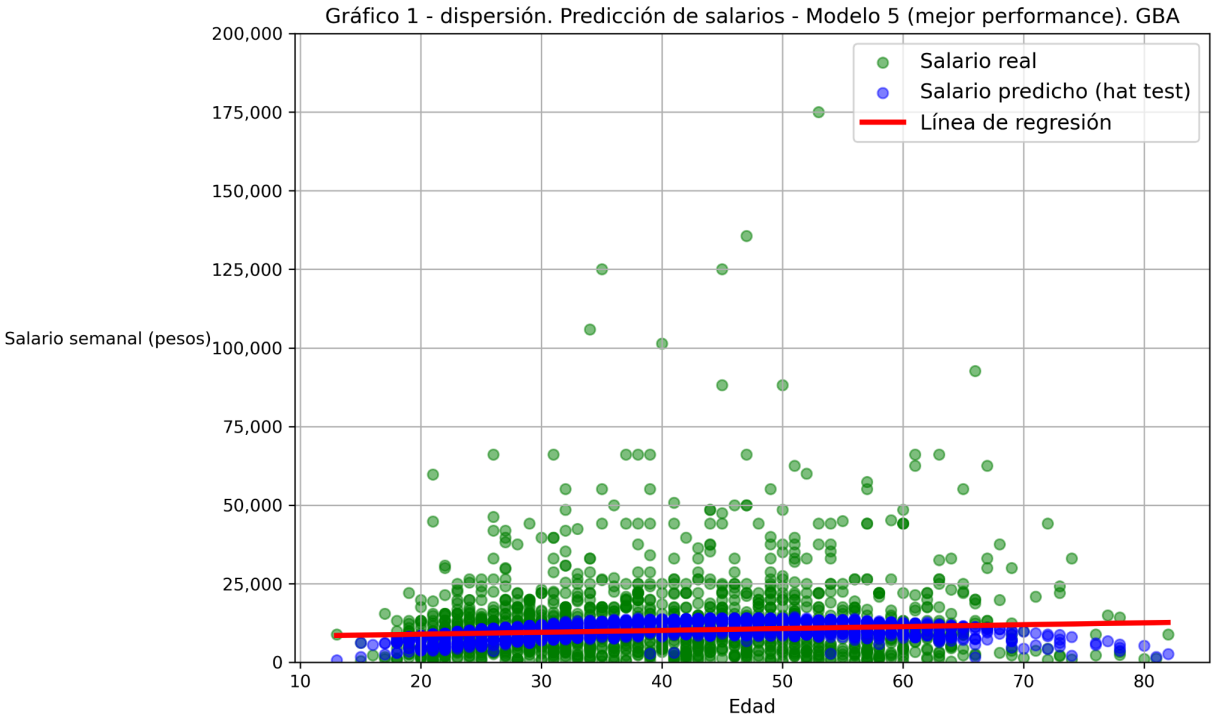
Con respecto a los resultados de la tabla, en primer lugar, vemos valores de MSE de testeo y de RMSE de testeo altos; esto implica por definición un error cuadrático medio alto, lo cual puede ser preocupante a la hora de analizar la viabilidad de nuestro modelo de predicción. Sin embargo, podemos estar en presencia de un conjunto de datos que presentan fuerte variabilidad y desviación entre ellos. Es decir, si el salario promedio semanal es alto o la desviación estándar es alta, coincide con un error cuadrático medio alto. Esto lo vamos a poder visualizar mejor en el gráfico del siguiente punto.

Por otro lado, vemos que a medida que le agregamos variables al modelo, en general las métricas bajan su valor. En particular vemos una fuerte disminución del MSE y el MAE cuando pasamos del modelo 3 al 4, esto es, cuando pasamos del modelo 3 que incluye ‘edad’, ‘edad2’, ‘intercepto’ y ‘educ’ al modelo 4 que le añade a esas cuatro la variable ‘mujer’. Esto implica también que la variable ‘mujer’ es muy significativa en el modelo y tiene mucho peso como variable explicativa, ya que su inclusión mejora positivamente las métricas.

Finalmente, vemos que la inclusión de las dummies ‘vp_dummy’ (dummy de vacaciones pagas) y ‘lee_dummy (dummy de sabe leer) mejoran levemente las métricas, pero no al nivel de la variable ‘mujer’. De esta forma, podemos concluir que el modelo 5 es el que cuenta con una mejor capacidad predictiva.

4) El Gráfico 1 muestra un gráfico de dispersión que representa la predicción del salario semanal , comparado con los datos efectivos de salario semanal. Por motivos de claridad expositiva , se muestra como límite superior del eje de ordenadas el monto de 200.000 pesos.

La predicción de salario semanal se basa en el modelo con mejor performance en las métricas de MSE, MAE y RMSE de testeo. El performance mejora cuando las métricas bajan en número, es decir, los errores disminuyen y hay una mejoría en la predicción del modelo. En el desarrollo de la anterior consigna, terminamos concluyendo que ese modelo era el 5.



En el eje de abscisas vemos la edad de los individuos que componen las observaciones. En el eje de ordenadas, vemos el valor del salario semanal (en pesos). Se presentan dos tipos de puntos de dispersión; en color azul, tenemos el salario predicho ($\hat{y}$), en color verde, los datos de salario real. Finalmente, en color rojo tenemos la línea de regresión lineal sobre los salarios predichos.

Como dijimos anteriormente, visualizar el gráfico nos permitiría entender las métricas de los errores con mayor claridad. Cuando lo contraponemos con los datos de salario real, las métricas cobran más sentido. En primer lugar, vemos fuertes outliers (puntos que se salen de la condensación general) en los datos reales de la variable dependiente (puntos de dispersión de color verde).

En segundo lugar, vemos una mayor variabilidad entre los datos de salario efectivos y los predichos, que se incrementa incluso a medida que la edad de las observaciones es mayor. Estos dos factores en conjunto (outliers y mayor dispersión) pueden ser la razón de que nuestro mean-squared error tenga un valor tan alto, además del propio tamaño de las cifras que representan los salarios..

Esto , sin embargo, no lo vemos reflejado en los puntos correspondientes a la predicción del modelo de regresión lineal; hay mucha menor variabilidad en los datos y no hay outliers. Podemos inferir que las variables explicativas del modelo de regresión lineal no son buenas para predecir estos valores extremos que se salen de la media. Esto se visualiza con la línea de regresión; la inclinación es muy poca, se ve una ligera mejoría cuando incrementa la edad, pero después esta cae.

Se aprecia en el gráfico una leve inclinación por ejemplo en los extremos de los datos vemos la misma dispersión, tanto de individuos que recién se incorporan a la población económicamente activa (18 o más) como de aquellos más longevos que trabajan muy pasada la edad de jubilación (individuos con 80 años que perciben un salario). Esto difiere del res los datos reales; la mayoría se condensa entre los 20 y los 60 años, y vemos mayores ingresos entre los 30 y 50 años.

5) Para determinar qué método predice mejor, analizamos las métricas de performance de los modelos de clasificación aplicados (regresión logística y KNN) sobre la base de testeo:

1. Regresión logística (Logit)
Matriz de confusión:

126	215
80	681

Marcado en verde ,de izquierda a derecha, aparece el número de verdaderos positivos y verdaderos negativos obtenido para este modelo. Los falsos positivos y falsos negativos se ubican en los recuadros rojos, también en la misma orientación..

Las medidas de performance para este modelo son :

- Accuracy: 0.7323049001814882
- AUC: 0.7333016057741588
- Curva ROC: El área bajo la curva (AUC = 0,733) indica una buena capacidad de discriminación entre las dos clases.

2. K-Vecinos más cercanos (KNN, k=5)
Matriz de confusión:

145	196
143	618

Las medidas para este modelo son:

- Accuracy: 0.6923774954627949
- AUC: 0.669220927857696
- Curva ROC: El área bajo la curva (AUC = 0,669) es menor que la del modelo logit, lo que indica menor capacidad predictiva.

En base a esta información, podemos afirmar que en este caso la regresión logística (Logit) supera al KNN en todas las métricas de performance:

Se reconoce comparando ambas matrices una mayor presencia de verdaderos positivos en los resultados de la regresión logística,además de un menor número de falsos negativos.

Tiene mayor accuracy, un mayor porcentaje de aciertos sobre el total de observaciones (73,2% respecto a 69,2% de KNN). Esto significa que el modelo clasifica correctamente una mayor proporción de casos fuera de la muestra de entrenamiento.

Su AUC es más alto (0,733 vs 0,669), mostrando mejor capacidad para distinguir entre las clases.

Además, la curva ROC del modelo logit se encuentra por encima de la del KNN, mostrando mejor desempeño general.

Se aprecia en los siguientes gráficos las curvas ROC correspondientes a cada modelo:

Gráfico A) : Regresión logística

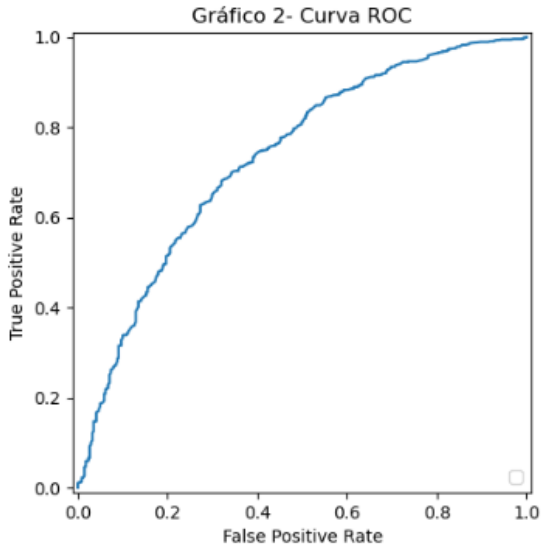
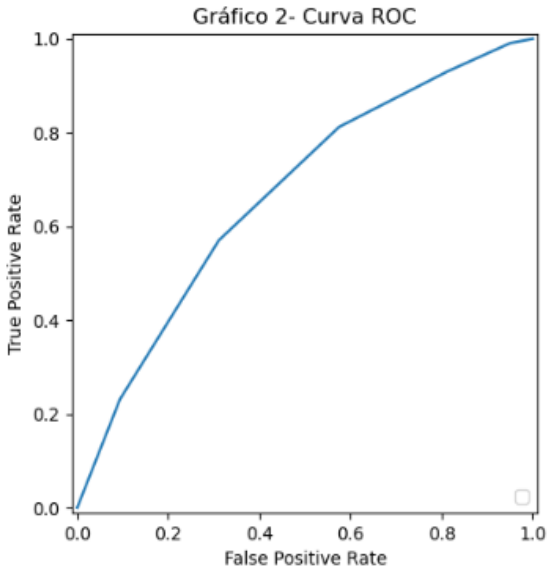


Gráfico B): KNN



6) Siguiendo la misma lógica y utilizando las variables auxiliares del punto 5, se aplicó el modelo logit previamente entrenado para estimar la probabilidad de estar desocupado entre las personas que no respondieron a la encuesta.

Para ello:

-Se analizaron los datos de las(edad, educ, mujer, lee_dummy).

-Se imputaron los valores faltantes de la misma manera que en el set de entrenamiento.

-Se aplicó el modelo logit entrenado, obteniendo la probabilidad de desocupación para cada individuo.

Como resultado, se estimó la proporción de personas identificadas como desocupadas dentro del grupo de no respondientes, obteniéndose un valor del 85,37%.

Este análisis sugiere que, según el modelo estimado, una gran parte de quienes no respondieron la encuesta podrían ser clasificados como desocupados. Esto puede indicar también una menor predisposición de las personas desempleadas a contestar. Es un punto importante a considerar para la interpretación de los resultados y la representatividad de la muestra.