# TRABAJO PRACTICO Nº4 DE BIG DATA SOBRE MÉTODOS SUPERVISADOS: REGRESIÓN & CLASIFICACIÓN USANDO LA EHP

---

Integrantes: Tomas Rodríguez Heredia,Diego Exequiel Coria, Rocco Pimenta
Fecha de entrega: Martes 2 de junio
Profesora: Romero Noelia

## Introducción:

En este cuarto trabajo estuvimos viendo los distintos métodos supervisados, usamos regresión, logic, vecinos cercanos, matrices de confusión y curvas ROC. Los aplicamos con los datos de la EPH que ya trabajamos en anteriores instancias. En particular vamos a estar tratando mucho con lo referente a “ESTADO” que dice ocupados y desocupados.

## *A. Enfoque de validación*

---

1)
Siguiendo la consigna de este primer punto usamos nuestro datos para dividir ocupados y desocupados y también si respondieron o no respondieron a la EPH. Luego dividimos nuestra base con test_split teniendo una diferencia entre 70% de entrenamiento y 30% de prueba.
Con todo esto podemos hacer la tabla de medias.

| | Media (Train) | Desvió (Train) | N (Train) | Media (Test) | Desvió (Test) | N (Test) | T-valor | P-valor |
|---|---|---|---|---|---|---|---|---|
| **CH04** | 1,44 | 0,49 | 2432 | 1,4 | 0,49 | 1043 | -0,04 | 0.86 |
| **CH06** | 39,71 | 12,84 | 2432 | 39,9 | 13,04 | 1043 | 0,1899999999999998 | 0.68 |
| **CH07** | 2,89 | 1,58 | 2432 | 2,9 | 1,59 | 1043 | 0,0099999999999979 | 0.79 |
| **CH09** | 1 | 0,06 | 2432 | 1 | 0,05 | 1043 | 0 | 0.85 |
| **CH12** | 12,84 | 4,3 | 2432 | 12,77 | 4,24 | 1043 | -0,0700000000000003 | 0.66 |
| **CH13** | 0 | 0 | 2432 | 0 | 0 | 1043 | 0 | 0 |
| **CH14** | 1,53 | 6,79 | 2432 | 1,37 | 4,63 | 1043 | -0,16 | 0.48 |
| **ESTADO** | 1,07 | 0,24 | 2432 | 1,06 | 0,23 | 1043 | -0,01 | 0.305 |
| **PP3E_TOT** | 32,57 | 21,45 | 2432 | 34,74 | 21,52 | 1043 | 2,17 | 0.0063 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **PP3F_TOT** | 1,13 | 4,6 | 2432 | 1,2 | 5,01 | 1043 | 0,07000000000000001 | 0.70 |

# *B.Metodo Supervisado 1: Modelo de Regresión Lineal*

---

2)
En este punto nos pide que con nuestra base de entrenamiento encontremos los coeficientes y la variación estándar de las variables que estuvimos construyendo en los otros trabajos, teniendo de variable dependiente a Salario_semanal.

Como referencia para la construcción de esas variable tenemos:

CH06=Edad
Estado+CH06= edad2
CH12+CH13+CH14=educ
2=mujer
CH07=estado civil
CH09=alfabeto

Importante añadimos un sexto modelo para que se resalte mas fuerte la diferencia que causa su presencia.

| Var.Dep: Salario_semanal | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 | Modelo 6 |
|---|---|---|---|---|---|---|
| Variables | (1) | (2) | (3) | (4) | (5) | (6) |
| Edad | Coef: 5686.731<br>Std: 1964.62 | Coef: 815.253<br>Std: 975.30 | Coef: 1187.123<br>Std: 985.90 | Coef: 1261.634<br>Std: 984.367 | Coef: 867.838<br>Std: 1030.26 | Coef: 970.192<br>Std: 1036.54 |
| Edad2 | | Coef: -1.97e+05<br>Std: 5.82e+04 | Coef: -1.65e+05<br>Std: 5.95e+04 | Coef: -1.50e+05<br>Std: 5.96e+04 | Coef: -1.42e+05<br>Std: 6.00e+04 | Coef: -1.61e+05<br>Std: 6.37e+04 |
| Educ | | | Coef: 1.41e+04<br>Std: 9.82e+03 | Coef: 1.75e+04<br>Std: 9.92e+03 | Coef: 1.77e+04<br>Std: 9.92e+03 | Coef: 1.88e+04<br>Std: 1.00e+04 |
| Mujer | | | | Coef: -1.708e+05<br>Std: 5.3e+04 | Coef: -1.685e+05<br>Std: 5.3e+04 | Coef: -1.704e+05<br>Std: 5.3e+04 |
| Estado civil | | | | | Coef: -2.225e+04 | Coef: -2.238e+04 |

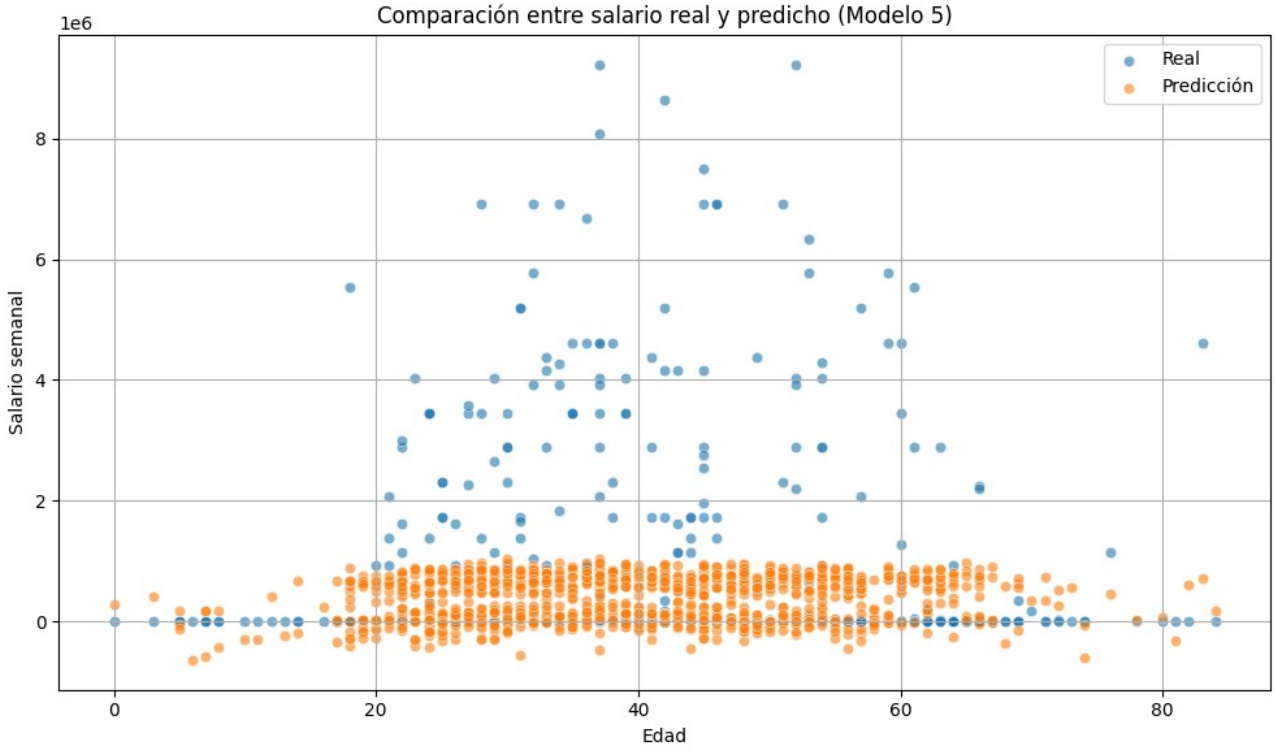
| | | | | | | |
|---|---|---|---|---|---|---|
| Alfabeto | | | | | Std: 1.72e+04 | Std: 1.72e+04 Coef: 2.11e+05 |
| **N** (observaciones) | 2567 | 2567 | 2567 | 2567 | 2567 | Std: 2.34e+05 2567 |
| R2 | | | | | | |

3)
Para este punto usamos nuestra base de testeo para sacar salario_semanal sombrerito, para lograr esto nos vamos a apoyar en los coeficientes del anterior punto y estas métricas de testeo:
MSE = Error cuadrático medio o Mean squared error
RMSE= Raíz del error cuadrático medio o Root mean squared error
MAE= Error absoluto medio o Mean absolute error

**Enfoque de Validación:**

| Var. Dep: salario_semanal | Modelo 1 | Modelo 2 | Modelo 3 | Modelo 4 | Modelo 5 |
|---|---|---|---|---|---|
| ***MSE test*** | 1617823574948,76 | 1506023676753,9 | 1502722060665,09 | 1495379093158,6 | 1487648397078,7 |
| ***RMSE test*** | 1271936,94 | 1227201,56 | 1225855,64 | 1222856,94 | 1219691,93 |
| ***MAE test*** | 703918,77 | 657471,70 | 664552,19 | 675634,28 | 680804,34 |

4)
Construimos un grafico de dispersión para mostrar nuestra predicción de salarios




El gráfico de dispersión nos permite contemplar la relación existente entre la edad y el salario semanal -tanto real como predicho. En el caso de que los puntos de la predicción, salario_semanal-hat_test, se acercan a los puntos reales, salario_semanal, esto indica que el modelo halla con éxito la relación entre la edad y el salario. Por otro lado, en el caso que hayan desviaciones, éstas pueden evidenciar sesgos o limitaciones del modelo en ciertos rangos de edad. Por ejemplo, sobre

predicción en jóvenes y subpredicción en mayores.

# *C. Métodos de Clasificación y Performance*

---

5)
En este quinto punto buscamos implementar métodos de regresión logística (Logit) y vecinos cercanos (KNN) en nuestra base de entrenamiento y para la de testeo una matriz de confusión y el AUC de curvas ROC.
Tuvimos que hacer una limpieza de datos debido a que los modelos resultaban muy perfectos, para eso utilizamos una matriz de correlación y descubrimos que nuestra variable problemática era “P21” que utilizamos para el salario_semanal. Sabiendo eso pudimos obtener resultados mas lógicos para nuestras predicciones.

Para nuestra base de entrenamiento El método de regresión lineal nos devolvió un accuracy test del: 0,12
Con vecinos cercanos la exactitud del modelo nos dio un: 0,36
en nuestra base de testeo con la matriz de confusión vemos una aproximación de: 0,11 y para nuestro ultimo planteo con curvas ROC al sacar el AUC que nos da 0,81 siendo la mejor aproximación de lejos.

Logit: 0,12
KNN:0,36
Matriz de confusión: 0,11

AUC: 0,81

6)
Para esta ultima consigna nos pide que con nuestra base de “norespondieron” hiciéramos una predicción de los desocupados en la misma. Tuvimos problemas a la hora de llamar estas funciones, para solucionarlo planteamos un nuevo train_split con los problemas filtrados.
Al final pudimos llegar a la conclusión de que hay un 83,33% de personas desocupadas en la base de datos “norespondieron”