UNIVERSIDAD DE BUENOS AIRES

BIG DATA & MACHINE LEARNING

Cuarto Trabajo Práctico

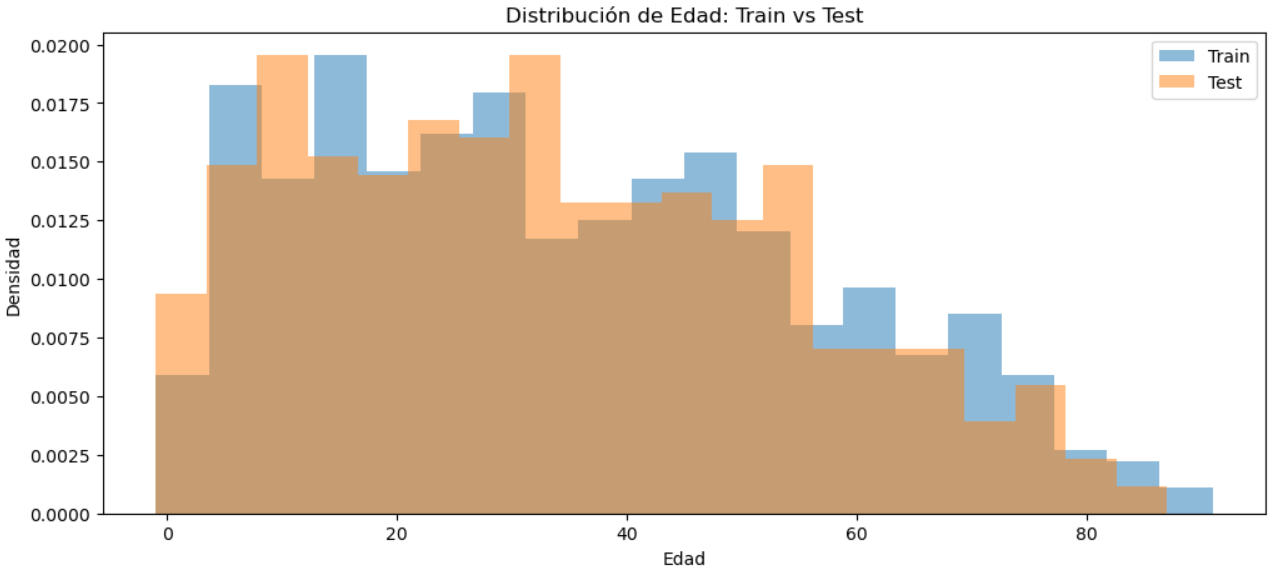
Grupo #27

Integrantes:

Ramiro Sanchez

Francisco Barrial

1. Enfoque de validación
2. En el primer punto, hicimos un gráfico de densidad de población por edad, para eso definimos variables como desocupación, edad al cuadrado, analizamos el nivel de educación por edad y calculamos el salario semanal para poder tener datos significativos para predecir los resultados de la mejor manera.



Para mejor entendimiento modificamos la opacidad de la figura para poder ver superposiciones.

La figura deja ver un ligero sesgo hacia la derecha, lo cual si bien ambos apuntan al mismo tipo de población señalando que la predicción funcionó correctamente, apunta a que el test indicaría a una población más jóven que el train. Por otra parte, ambos observan un pico entre los 20 y los 40 años de edad, lo que corresponde típicamente con los datos laborales.

B. Método supervisado 1: Modelo de Regresión Lineal

1. Para el segundo punto estimamos el salario semanal en 5 modelos diferentes,

edad

edad+edad²

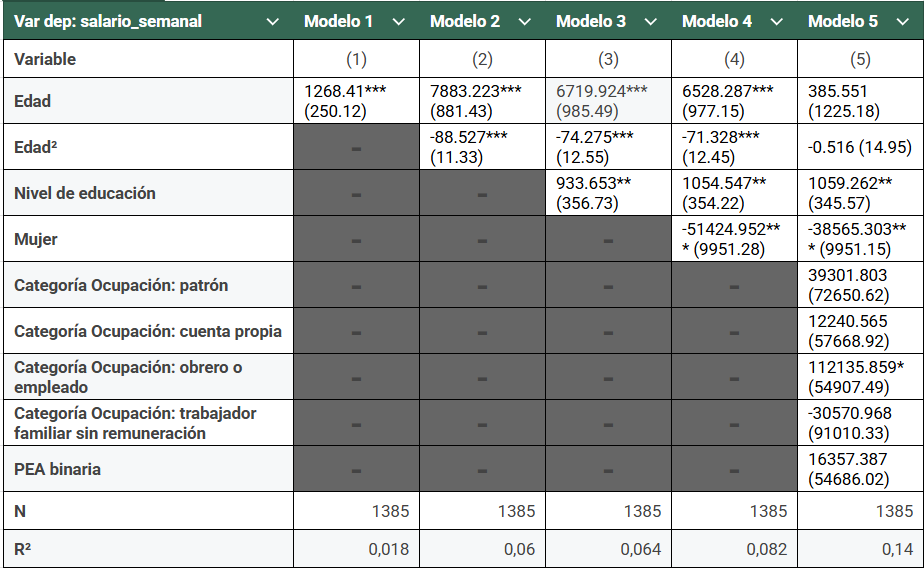
edad+edad²+nivel\_ed

edad+edad²+nivel\_ed+mujer

y para el quinto elegimos tipo de ocupación que tiene 4 categorías distintas y como venía de la mano con una de esas 4 variables elegimos actividad/inactividad de la PEA como segunda, dejando el modelo de la siguiente forma:

edad+edad²+nivel\_ed+mujer+categoría de ocupación+PEA binaria

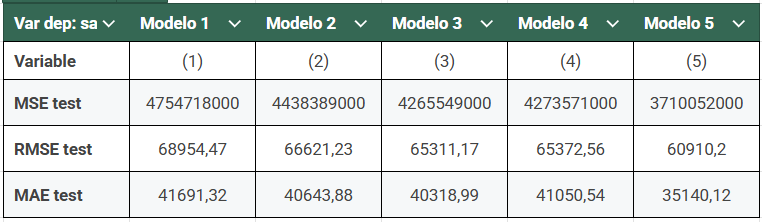
Dando como resultado el siguiente cuadro.



La cantidad de asteriscos hace referencia a la significatividad de la variable, siendo 3 asteriscos el máximo, por ejemplo podemos ver que la edad es muy significativa en todos los modelos menos llegado al último, al igual que la edad². El nivel de educación si bien no tiene el máximo nivel de significatividad, mantiene el propio en todos los modelos presentes al igual que la variable mujer. Por último de las variables seleccionadas podemos ver como la única con cierto nivel de significatividad es la categoría de ocupación de obrero/empleado.

Podemos ver un R² bajo, que si bien marca que la diferencia de ingreso no está siendo explicada mayormente por las variables independientes, si deja resaltar variables como la de género y da visibilidad al hecho de que según el modelo las mujeres perciben entre $38,565 y $51,424 menos que los hombres en salario semanal. Además indica que hay más variables incluidas en el motivo de la variación de ingresos semanales.

1. En esta instancia del trabajo calculamos el error medio cuadrático, su raíz y el error absoluto medio de todos los modelos, dando como resultado lo siguiente.



Podemos observar que a excepción del modelo 4, mientras más variables posea, más bajo es el error, siendo el quinto modelo, donde se agrega categoría de ocupación y actividad en la PEA el que más progreso aporta, bajando el error en un 15% dando un desvío de la predicción de alrededor de $35.000 frente al salario real en la mitad de los casos.