UNIVERSIDAD DE BUENOS AIRES - 2025

BIG DATA & MACHINE LEARNING

TRABAJO PRÁCTICO N° 4

MÉTODOS SUPERVISADOS: REGRESIÓN & CLASIFICACIÓN USANDO LA EHP

Fecha de entrega: 3 de junio a las 12:59 hs.

Contenido: Regresión lineal de salarios y aplicar los métodos de clasificación vistos en clase usando la EPH.

Modalidad de entrega

- Asegurense de haber creado una carpeta llamada TP4 en su repositorio de GitHub del grupo.

- El informe debe subirse a dicha carpeta en repositorio del grupo en formato PDF con el nombre **Big_Data_TP4_Grupo#.pdf** (donde **#** es el número de grupo), incluyendo gráficos e imágenes dentro del mismo archivo. La extensión máxima es de 8 páginas (sin apéndices) y se espera una redacción clara y precisa.

- Se debe publicar el código con los comandos utilizados, indicando claramente a qué inciso corresponde cada uno. El nombre del archivo deberá ser **Big_Data_TP4_Grupo#**.
    - Al finalizar el trabajo práctico deben hacer un último commit en su repositorio de GitHub llamado "Entrega final del tp".
    - El Jupyter Notebook y el correspondiente al TP4 deben estar dentro de esa carpeta.
    - La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
        - No envien el correo hasta no haber terminado y estar seguros de que han hecho el *commit* y *push* a la versión final que quieren entregar.
        - No hagan nuevos *push* despues de haber entregado su versión final. Esto generaría confusión acerca de que versión es la que quieren que se les corrija.

- También deben enviar el link de su repositorio -para que pueda ser clonado y corregido- a mi correo 25RO35480961@campus.economicas.uba.ar . Usar de asunto de email

"**Big Data - TP 4 - Grupo #"** donde # es el número de grupo que le fue asignado.

- En resumen, la carpeta del repositorio debe incluir:
    - El codigo
    - Un documento Word or PDF donde esten las figuras, tablas y una breve descripción de las mismas. Tenga en cuenta la devolución general de TP2. Si no incorpora dichos lineamientos el nuevo reporte, entonces se le descontara puntos por mala presentación de la información.
- **Cualquier detección de copia o plagio será sancionada.**

El objetivo de este trabajo práctico es intentar predecir si una persona está desocupada o no utilizando distintas variables de características individuales y los distintos clasificadores vistos en clase. Recuerden que en los trabajos prácticos anteriores crearon dos bases de datos distintas: `respondieron`, que tiene datos de personas que sí respondieron su condición de empleo y `norespondieron`, que tienen aquellas personas que no declaran estar desempleadas o empleadas.

A. Enfoque de validación

Utilicen la base `respondieron`. Para cada año, dividan las observaciones en una base de prueba (test) y una de entrenamiento (train) utilizando el comando `train_test_split`. La base de entrenamiento debe comprender el 70% de los datos, y la semilla a utilizar (*random state instance*) debe ser 444. Establezca a desocupado como su variable dependiente en la base de entrenamiento (vector y). El resto de las variables seleccionadas serán las variables independientes (matriz X). Recuerden agregar la columna de unos (1).

1. Cree una tabla de diferencia de medias entre la base de entrenamiento y la de testeo de las características seleccionadas en su matriz X. Comente la tabla de la diferencia de medias de sus variables entre entrenamiento y testeo.

B. Metodo Supervisado 1: Modelo de Regresión Lineal

2. Para los ocupados de la EPH en su region seleccionada usando la **base de entrenamiento**, estime los siguientes modelos usando como variable dependiente *salario_semanal* ***(y)*** y como predictores las variables creadas en el TP3:

    i. *salario_semanal* en *edad*
    ii. *salario_semanal* en *edad* y *edad2*
    iii. *salario_semanal* en *edad, edad2* y *educ*
    iv. *salario_semanal* en *edad, edad2, educ* y *mujer* (donde es una dymmy que toma *mujer*=1 si CH04==2)
    v. *salario_semanal* en *edad, edad2, educ, mujer* y dos variables que haya creado y limpiado en el TP2 o TP3 que usted crean son relevantes para predecir salarios semanales.

    Complete y comente brevemente la siguiente tabla, reportando los coeficientes (hasta 3 decimales luego de la coma) y desvío estandar (sd con 2 decimales despues de la coma) de cada coeficiente entre parentesis:

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**)**					
R²					

Nota: destaque con *, **, y *** cuando el p-valor de los coeficientes reportados sean menor que 0.1, 0.05 y 0.001 respectivamente.

3. *Enfoque de Validación:* Ahora para cada modelo estime el salario predicho de testeo (*salario_semanal_test* sombrerito) usando las observaciones separadas de **testeo** y los coeficientes estimados en el apartado anterior. Reporte y comente las siguientes métricas de testo para cada modelo:

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					

4. *Opcional:* Para el modelo de mejor performance, ilustre la predicción de salarios (*salario_semanal_hat_test*) en un grafico de dispersión con *salario_semanal* en el eje vertical y *edad* en el eje horizontal usando la base de testeo. Comente brevemente el gráfico.

C. Métodos de Clasificación y Performance

5. Implementen los métodos de **regresión logistica** (logit) y **vecinos cercanos (KNN)** con K=5 usando en la *base de entrenamiento*. Luego, usando la *base de testeo*, reporte la matriz de confusión para p>0.5, la curva ROC, los valores de AUC y de Accuracy de testeo de cada uno.

   ¿Cuál de los métodos predice mejor en cada año? Justifiquen detalladamente utilizando las medidas de perfomance mencionadas arribas.

6. Con el método que seleccionaron, predigan qué personas son desocupadas dentro de la base `norespondieron`. ¿Qué proporción de las personas que no respondieron pudieron identificar como desocupadas?