Big Data - UNT

PRIMAVERA 2024

TRABAJO PRÁCTICO N 3

ANÁLISIS DESCRIPTIVO Y PREDICCIÓN DE DESOCUPACIÓN

Reglas de Formato y Presentación

Fecha de entrega: jueves 10 de octubre a las 13.59 hs.

Contenido: Análisis descriptivo y problema de clasificación de empleo entre cohortes.

Modalidad de entrega

Al finalizar el trabajo práctico deben hacer un último `commit` en su repositorio de GitHub con el mensaje Entrega final del TP.

- Asegúrense de haber creado una carpeta llamada TP3. Deben entregar un reporte (pdf) y el código (Jupyter notebook). Ambos deben estar dentro de esa carpeta.
- Deberán enviar el link a su repositorio -para que pueda ser clonado y corregido- al canal de Slack #tp-entregas
- La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
    - No envíen el mensaje 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` después de haber entregado su trabajo. Esto generaría confusión acerca de qué versión es la que quieren que se les corrija.

Modalidad de entrega

- El informe debe ser entregado en formato PDF, con los gráficos e imágenes en este mismo archivo. Puede tener una extensión máxima de hasta 8 paginas (no se permite Apéndice). Se espera una buena redacción en la resolución.

- Entregar el código con los comandos utilizados, identificando claramente a que inciso corresponde cada comando.

- **Importante**: Todos los miembros del equipo deben haber hecho al menos un `commit` durante la realización del TP para asegurar que todos hayan aportado a su resolución.

Parte I: Analizando la base

La Encuesta Permanente de Hogares (EPH) es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el Instituto Nacional de Estadística y Censos (INDEC), que permite conocer las características sociodemográficas y socioeconómicas de la población. Uno de los indicadores mas valiosos sobre el mercado laboral que pueden obtenerse con los datos de esta encuesta es la tasa de desocupación.

1. Utilizando información disponible en la pagina del INDEC, expliquen brevemente como se identifica a las personas desocupadas.

2. Entren a la pagina https://www.indec.gob.ar/ y vayan a la sección Servicios y Herramientas -> Bases de datos. Descarguen la base de microdatos de la Encuesta Permanente de Hogares (EPH) correspondiente al primer trimestre de **2004** y **2024** en formato .dta y .xls, respectivamente (una vez descargadas, las bases a usar deberán llamarse `usu_individual_T104.dta` y `usu_individual_T124.xls`). En la pagina web, también encontrara un diccionario de variables con el nombre de "Diseño de registro y estructura para las bases preliminares (hogares y personas)". Descarguen el diccionario de cada año. En estos archivos se les indica qué significa cada variable que aparece en la base de datos, en particular, en la sección de Diseño de registros de la base Personas.

    a. Eliminen todas las observaciones que **no** corresponden a los aglomerados de Gran Tucumán - Tafí Viejo y unan ambos trimestres en una sola base.

    b. Si hay observaciones con valores que no tienen sentido, descártenlas (por ejemplo, ingresos y edades negativos). Expliquen las decisiones tomadas.

    c. Una vez hecha esa limpieza, realicen un gráfico de barras mostrando la composición por sexo para 2004 y 2024. Comenten los resultados.

    d. Realicen una matriz de correlación para 2004 y 2024 con las siguientes variables: `CH04, CH06, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC, IPCF`. Utilicen alguno de los comandos

disponibles en este link o este link para graficar la matriz de correlación. Comenten los resultados.[1]

e. ¿Cuántos desocupados hay en la muestra? ¿Cuántos inactivos? ¿Cuál es la media de ingreso per cápita familiar (IPCF) según estado (ocupado, desocupado, inactivo)?

3. Uno de los grandes problemas de la EPH es la creciente cantidad de hogares que no reportan sus ingresos (ver por ejemplo el siguiente informe). ¿Cuántas personas no respondieron cual es su condición de actividad? Guarden como una base distinta llamada `respondieron` las observaciones donde respondieron la pregunta sobre su condición de actividad (`ESTADO`). Las observaciones con `ESTADO=0` guárdenlas en una base bajo el nombre `norespondieron`.

4. Agreguen a la base `respondieron` una columna llamada PEA (Población Económicamente Activa) que tome 1 si están ocupados o desocupados en `ESTADO`. Realicen un gráfico de barras mostrando la composición por PEA para 2004 y 2024. Comenten los resultados.

5. Agreguen a la base `respondieron` una columna llamada PET (Población en Edad para Trabajar) que tome 1 si están la persona tiene entre 15 y 65 años cumplidos. Realicen un gráfico de barras mostrando la composición por PEA para 2004 y 2024. Comenten los resultados y compare PET con PEA.

6. Por ultimo, agreguen la base `respondieron` una columna llamada desocupado que tome 1 si esta desocupada. ¿Cuántas personas están desocupadas en 2004 vs 2024?

   a. (Opcional) Muestre la proporción de desocupados por nivel educativo comparando 2004 vs 2024. ¿Hubo cambios de desocupados por nivel educativo?

   b. (Opcional) Cree una variable categórica de años cumplidos (`CH06`) agrupada de a 10 años. Muestre proporción de desocupados por edad agrupada comparando 2004 vs 2024. ¿Hubo cambios de desocupados por edad?

7. (Opcional) *Dos tasas de desocupación*: calcule la tasa de desocupación para Tucumán en 2004 y 2024 siguiendo la definición del INDEC. En economía laboral, muchas veces se argumenta que participar del mercado laboral (PEA=1) es una decisión endógena de los individuos y no quisiéramos que el indicador de la tasa de desocupación varíe por dicha decisión. Calcule la tasa de desocupación alternativa como el

[1] Para todos los gráficos que presente, recuerde tener presentes los tres principios de visualización de datos discutidos en la Clase 1. Referencia: † Schwabish, J. A. (2014). An economist's guide to visualizing data. *Journal of Economic Perspectives*, 28(1), 209-234.

porcentaje de desocupados respecto de la PET. Presente en una tabla o gráfico la tasa de desocupación del INDEC y de economía laboral para 2004 y 2024 y comente los resultados. ¿Cuáles son las ventajas y desventajas de dichas mediciones?

Parte II: Clasificación

El objetivo de esta parte del trabajo es intentar predecir si una persona está desocupada o no utilizando distintas variables de características individuales.

1. Para cada año, partan la base `respondieron` 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 101. Establezca a desocupado como su variable dependiente en la base de entrenamiento (vector y). El resto de las variables serán las variables independientes (matriz X). Recuerden agregar la columna de unos (1).

2. Implementen los siguientes métodos reportando luego la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy de cada uno:

    - Regresión logística
    - Análisis discriminante lineal
    - KNN con k=3
    - Naive Bayes

3. Compare los resultados de 2004 versus 2024. ¿Cuál de los métodos predice mejor en cada año? Justifiquen detalladamente utilizando las medidas de precisión que conocen.

4. 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?