Universidad de San Andrés Agosto de 2021

Big Data, Aprendizaje y Minería de Datos

Trabajo Práctico N°2: Clasificación aplicada a la EPH

Profesor: Walter Sosa Escudero
Asistente: Belén Michel Torino

Fecha de entrega: Lunes 13/9 a las 23:59.
Contenidos: analizar el problema de la medición de la tasa de pobreza y aplicar métodos vistos en clase para identificar individuos que caen bajo la línea de pobreza.

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 TP2. Su reporte (pdf) y código (jupyter notebook) deben estar dentro de esa carpeta.
- También deben completar el link de su repositorio -para que pueda ser clonado y corregido- en esta google sheet
- La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
  - No completen la google sheet hasta no haber terminado y estar seguros de que han hecho el commit y push a la versión final que quieren entregar. Debido a que se pueden tomar hasta 3 días de extensión a lo largo del curso, yo no corregiré sus tareas hasta no ver el link en la google sheet.
  - No hagan nuevos push después de haber entregado su versión final. Esto generaría confusión acerca de qué versión es la que quieren que se les corrija.

Reglas de formato y presentación

- El trabajo debe tener una extensión máxima de 8 páginas (no se permite Apéndice). Se espera una buena redacción en la resolución del práctico.

- El informe debe ser entregado en formato PDF, con los gráficos e imágenes pegadas en este mismo archivo.
- Adjuntar el script con los comandos utilizados, identificando claramente a qué inciso corresponde cada comando.

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 más valiosos que pueden obtenerse con los datos extraídos de esta encuesta es la tasa de pobreza.

1. Utilizando información disponible en la página del INDEC, explique brevemente cómo se identifica a las personas pobres.
2. Entre a la página https://www.indec.gob.ar/ y diríjase 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 2021 en formato xls (una vez descargada, la base a usar debería llamarse `usu_individual_T121.xls`). En la página web, también encontrará un diccionario de variables con el nombre de "Diseño de registro y estructura para las bases preliminares (hogares y personas)"; este archivo 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. Elimine todas las observaciones que **no** corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires.
    b. Si hay observaciones con valores que no tienen sentido, descártelas (ingresos y edades negativos, por ejemplo).
    c. Una vez hecha esa limpieza, realice un gráfico de barras mostrando la composición por sexo.
    d. Realice una matriz de correlación con las siguientes variables: `CH04, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC, IPCF`. Comente los resultados. Utilice alguna de los comandos disponibles en el siguiente link para realizar la matriz.
    e. ¿Cuántos desocupados hay en la muestra? ¿Cuántos inactivos? ¿Cuál es la media de ingreso per cápita familiar (IPCF) según estado (ocupado, desocupado, inactivo)?
    f. Utilizando el archivo `tabla_adulto_equiv.xlsx`, agregue una columna a su base de datos llamada `adulto_equiv` que contenga los valores de adulto equivalente de cada persona según su sexo y edad (por ejemplo, a un varón de 2 años le corresponde 0.46). Finalmente, con el comando `groupby` sume esta nueva columna para las personas que pertenecen a un mismo hogar y guarde ese dato en una columna llamada `ad_equiv_hogar`[1].

[1] Por ejemplo, si una familia está compuesta por un varón de 40 años (`adulto_equiv`=1) y su esposa de la misma edad (`adulto_equiv`=0.77) con sus mellizos varones de 5 años (`adulto_equiv`=0.60 cada uno), a todos se les deberá imputar en `ad_equiv_hogar` un valor igual a 2.97, que es la cantidad de adultos equivalentes en ese hogar.

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 cuál es su ingreso total familiar (ITF)? Guarde como una base distinta las observaciones donde respondieron la pregunta sobre su ITF bajo el nombre `respondieron`. Las observaciones con `ITF=0` guárdelas bajo el nombre `norespondieron`.
4. Sabiendo que la Canasta Básica Total para un adulto equivalente en el Gran Buenos Aires en el primer trimestre de 2021 es aproximadamente $18.914, agregue a la base `respondieron` una columna llamada `ingreso_necesario` que sea el producto de este valor por `ad_equiv_hogar`. Note que este es el valor mínimo que necesita ese hogar para no ser pobre
5. Por último, agregue a `respondieron` una columna llamada `pobre` que tome valor 1 si `ingreso_necesario` es menor al ITF que reportó esa familia, y 0 en caso contrario. ¿Cuántos pobres identificó?

Parte II: Clasificación

El objetivo de esta parte del trabajo es intentar predecir si una persona es o no pobre utilizando datos distintos al ingreso, dado que muchos hogares son reacios a responder cuánto ganan.

1. Elimine de la base las columnas `respondieron`, `norespondieron` y todas las variables relacionadas a ingresos (en el archivo `codigos_eph.pdf` ver las categorías: ingresos de la ocupación principal de los asalariados, ingresos de la ocupación principal, ingresos de otras ocupaciones, ingreso total individual, ingresos no laborales, ingreso total familiar, ingreso per cápita familiar). Elimine también las columnas `adulto_equiv`, `ad_equiv_hogar` e `ingreso_necesario`.
2. Parta la base `respondieron` en base de prueba (`test`) y base 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.
3. Establezca a `pobre` como su variable dependiente en la base de entrenamiento (vector y). El resto de las variables serán las variables independientes (matriz X). Recuerde agregar la columna de 1.
4. Implemente los siguientes métodos reportando luego la matriz de confusión, la curva ROC y los valores de AUC y de Accuracy de cada uno:
   - logit
   - Análisis de discriminante lineal
   - KNN con k=3
5. ¿Cuál de los tres métodos predice mejor? Justifique detalladamente utilizando las medidas de precisión que conoce.
6. Con el método que seleccionó, prediga qué personas son pobres dentro de la base `norespondieron`. ¿Qué proporción de las personas que no respondieron pudo identificar como pobres?

7. Note que para correr los tres métodos se utilizaron todas las variables disponibles como predictores. ¿Le parece esto correcto? ¿Qué variables habría conservado? Con las variables seleccionadas, implemente únicamente el modelo logit nuevamente y compare las medidas de precisión obtenidas con los resultados del modelo logit anterior. ¿Cambió mucho la precisión?