factor~data
EIDAES_UNSAM

Diploma Universitario
en Ciencias Sociales Computacionales y Humanidades Digitales

Consigna Trabajo FInal
Opción 1 (Módulos 3 y 4)

Introducción

El trabajo final de los módulos 3 y 4 consiste en tomar un subconjunto de los datos de la Encuesta Permanente de Hogares y construir modelos que predigan los valores de la variable objetivo (ingresos de la ocupación principal). Si bien es importante la capacidad predictiva y las métricas del modelo final, prestaremos especial atención en la corrección a la explicación y justificación de las decisiones tomadas, en tanto muestren manejo de los temas vistos en clase.

Dataset

El dataset presentado es un subconjunto de datos de la tabla de individuos de la Encuesta Permanente de Hogares correspondiente al III trimestre del 2021. Solamente se presentan algunas variables (que se detallan a continuación) y se han seleccionado las personas ocupadas (las personas que trabajaron al menos una hora durante la semana anterior al operativo). Es decir, no se encuentran en esta base de datos las personas inactivas ni tampoco las desocupadas. Para el presente ejercicio no se utilizarán los ponderadores ni los factores de expansión de la Encuesta.

Variables

- NIVEL_ED: nivel educativo
- CH03: relación de parentesco con el jefe de hogar
- CH04: sexo
- CH06: edad en años cumplidos
- CH07: estado conyugal
- PP04A: sector del establecimiento donde trabaja (público-privado)
- CAT_OCUP: categoría ocupacional
- INTENSI: intensidad de la ocupación (ocupado pleno, sobreocupado, subocupado)
- PP3E_TOT: cantidad de horas que trabajó la semana anterior en la ocupación principal
- CATEGORIA: carácter de la ocupación principal (basado en el CNO)
- CALIFICACION: calificación de la ocupación principal (basado en el CNO)
- P21: Ingreso de la ocupación principal

Pueden descargar el dataset desde este link.

Para mayores detalles sobre estas variables y sus definiciones conceptuales y operativas pueden consultar la siguiente documentación:

- Diseño de registro contiene todas las variables de las tablas de hogar e individual y los sistemas de categorías
- Descripción general de la EPH (definiciones conceptuales, operativas, diseño del operativo, etc.)
- Clasificador Nacional de Ocupaciones (definiciones conceptuales)

Consignas

- Realizar un análisis exploratorio con visualizaciones y métricas del dataset. ¿Qué conclusiones útiles para la modelización pueden sacarse?
- Construir al menos dos modelos para predecir los ingresos de la ocupación principal (P21). Comparar los resultados entre ambos y elegir el enfoque más útil y justificar la elección. Podrá utilizar tanto modelos de regresión lineal o logística como modelos basados en ensamble learning.
- Interpretar el output del modelo final en base a las métricas obtenidas, por un lado, y por el otro en base a la información que brinda respecto del fenómeno en estudio.

Entregables

Se esperan dos entregables:

- Un documento en formato word, google doc (no pdf) en el que se desarrollan las respuestas a las consignas, se presentan los principales resultados (tablas, visualizaciones) y las interpretaciones.
- Un notebook/script en el que se realiza el procesamiento del texto y se generan los modelos y visualizaciones correspondientes.

La entrega se hará mediante el sistema Google Classroom

Modalidad de trabajo

El trabajo deberá hacerse en grupos de 2 personas. El trabajo deberá ser subido a Google Classroom *por solamente uno de los integrantes* del grupo. Deberá constar en el trabajo el nombre de ambos.

Fecha de entrega

La fecha última de entrega será el 17/10/2023 a las 23:59. No se concederán prórrogas. (A CONFIRMAR)