CODER HOUSE

Análisis socioeducativo de los habitantes de la Ciudad de Buenos Aires

Profesor: Damian Dapueto

Tutor: Héctor Alonso

Grupo de Trabajo: Lucia Buzzeo, Lucia Hukovsky, Jose Saint German, Juan Martín Carini

Introdución

Objetivos

- Descubrir las principales variables intervinientes en el nivel máximo educativo alcanzado por la población de la Ciudad Autónoma de Buenos Aires (CABA).
- Generar un modelo de predicción aplicado a nuestra variable target "Nivel Máximo Educativo", esto lo haremos implementando los siguientes modelos de clasificación:
  - **Árbol de decisión:** que construye un árbol durante el entrenamiento que es el que aplica a la hora de realizar la predicción.
  - **Bosque Aleatorio:** que es un conjunto (ensemble) de árboles de decisión combinados con bagging.

Fuente de información

Para trabajar esta problemática, se ha recurrido a la Encuesta Anual de Hogares del Gobierno de la Ciudad de Buenos Aires para el año 2019. El dataset está disponible en la base de datos abiertos del GCBA.
Esta encuesta contiene información demográfica, social, económica, educativa y de salud de 14319 habitantes de la Ciudad, la cual es una muestra representativa que permite obtener un vistazo de la población de la Ciudad.

Estructura de los trabajos

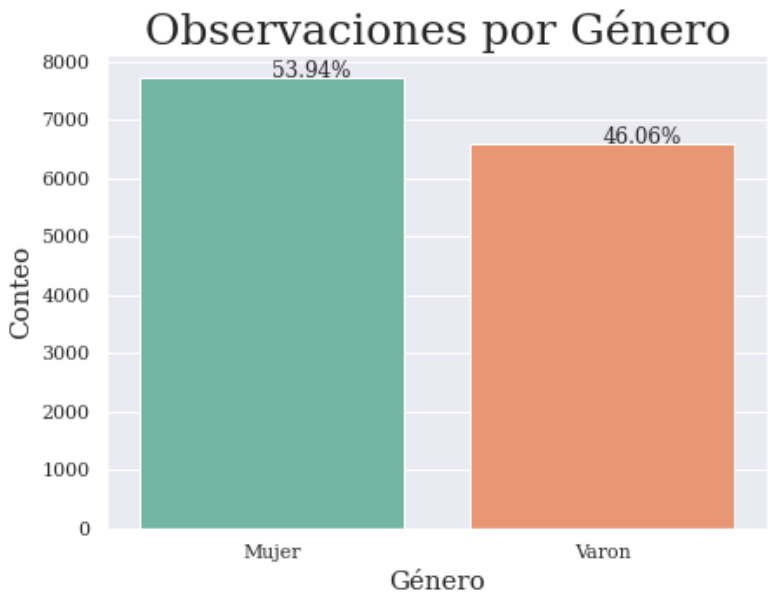
Este trabajo se ha dividido en 3 partes:

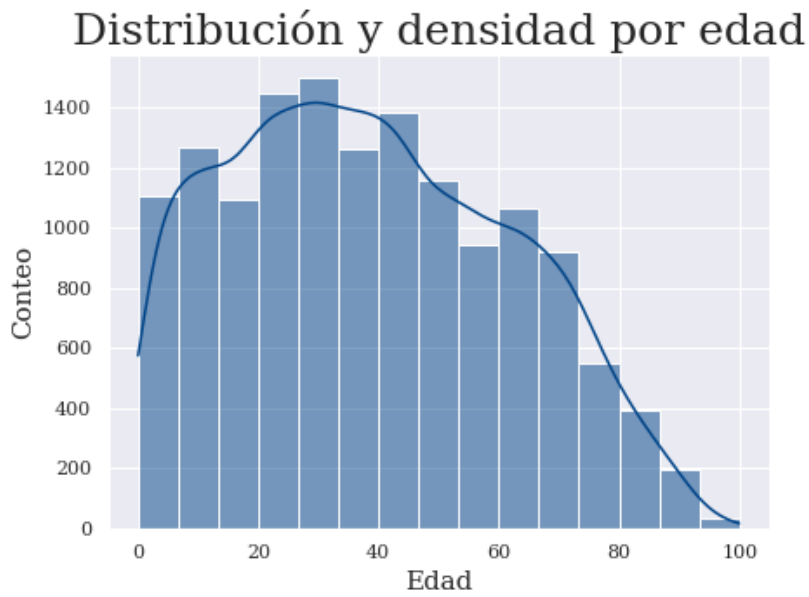
1. **Introducción a las variables del problema:** Se realiza un análisis de las variables del dataset. En el mismo se busca conocer su performance dentro del dataset. A la vez, se investiga cómo las variables interactúan entre sí. Esta parte es lo que se conoce como análisis univariado, bivariado y multivariado.
2. **Modelos analíticos:** En esta sección se entrenan diversos modelos analíticos y algoritmos que sirven para alcanzar los objetivos seteados para el presente proyecto. Como la variable objetivo es categórica, se realizan modelos de clasificación.
3. **Conclusión:** Se alcanzan conclusiones finales sobre los hallazgos. Además, se discuten posibles limitaciones y se plantean futuras líneas de análisis, a partir del análisis presente.

Análisis exploratorio de los datos

Análisis univariado

Comenzamos con un pantallazo general sobre las primeras cualidades de los datos, como muestra representativa para la EPH, sobre quiénes son los ciudadanos representados en el dataset.





En la variable género los datos parecen equilibrados en las categorías. Para el caso de la variable "edad", la distribución se asemeja a la de una normal.