Informe Trabajo práctico N°3

Alumnos: Strejilevich Francisco, Giacobelli Francisco
Profesores: María Noelia Castillo, Ignacio Spiousas
Materia: Ciencia de Datos
3/11/2024

Análisis y limpieza de base/s

La base utilizada para este trabajo consiste en los resultados de la Encuesta Permanente de Hogares (EPH) de los años 2004 y 2024. Analiza el porcentaje de sujetos empleados y desocupados en la argentina mediante una encuesta a hogares, considerando a una persona desempleada como aquella que actualmente no tiene trabajo, están disponibles para trabajar y están buscando activamente trabajo. El promedio siendo medido utilizando la población activa (INDEC 2024).

Para la limpieza y el posterior análisis de las bases de datos se utilizó Visual Studio Code y Google Colab, utilizando como lenguaje de programación Python. Se insertaron ambas bases de datos (la de 2004 y la de 2024, ambas del primer trimestre) en dos dataframes diferentes para posteriormente unirlos luego de la limpieza. Se eliminó dato que no proviniera de la provincia de Buenos Aires ni de Ciudad Autónoma de Buenos Aires. y se seleccionó las columnas ch04, ch06, ch07, ch08, nivel_ed, estado, cat_inac e ipcf dentro de las bases como columnas a utilizar y por ende fueron las únicas que fueron limpiadas.

En primer lugar, se chequeó si había datos sin información (valores nan), encontrándose que ninguna variable tenía. Se eliminó después cualquier valor que fuera negativo de las variables numéricas (en el caso de las analizadas siendo ch06 que representa la edad de los sujetos y ipcf que es el monto de ingreso per cápita familiar), ya que se trata de valores imposibles considerando que son la edad de los sujetos y su ingreso.

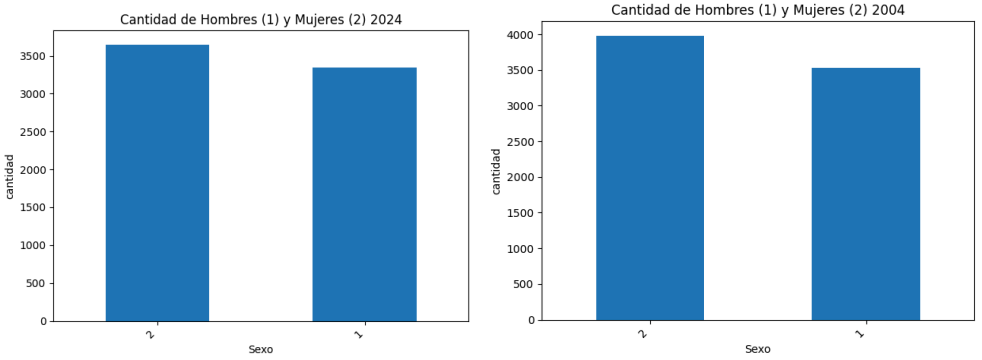
En segundo lugar, se transformó las variables categóricas de la base de datos del 2004 en variables numéricas, utilizando los valores que se les categoriza dentro de la base de 2024 a cada una de las respuestas utilizando el manual de diseño de la base de datos (INDEC, 2024) para posteriormente concatenar las bases de datos en una sola, agregando una columna “ano” (años) en la que marcamos cual era la base originaria de cada dato para facilitar el análisis posterior.

En tercer lugar, se convirtió las variables ch04, ch07, ch08 en variables binarias. ch04 representa el género de los participantes por lo que consideramos adecuado marcar como 0 a los hombres y como 1 a las mujeres. Ch07 define el estado civil del participante por lo que consideramos adecuado limitarlo a definir si está en pareja o no. Ch08 representa si el sujeto tiene algún tipo de cobertura médica clasificando si tiene y de qué tipo, decidimos limitarlo solo a definir si tiene algún tipo de cobertura médica o no. Además, para la variable estado que define cuál es el estado laboral actual del sujeto (ocupado,

desocupado, inactivo y menor de 10 años) definimos variables dummy para analizar posteriormente la correlación entre las variables debido a que consideramos que si bien se podría categorizar en personas activas laboralmente (ocupado y desocupado) y personas inactivas (jubilados y niños) consideramos que es importante poder definir el efecto que tiene cada uno de estos grupos.

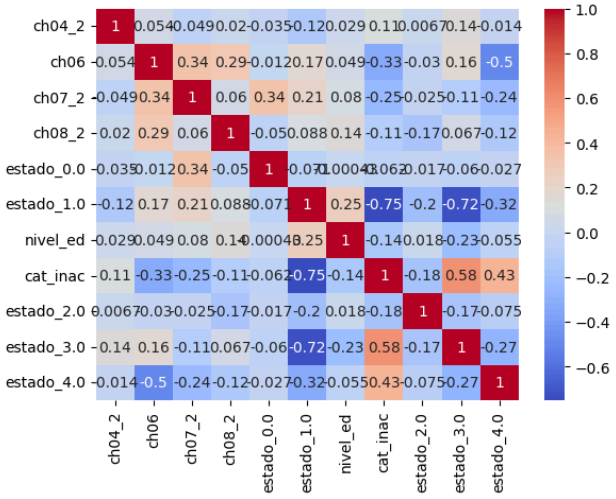
Gráficos y analisis de frecuencia/densidad

Se realizó un gráfico de barras para comparar la cantidad de hombres y mujeres que participaron en la encuesta en cada uno de los años

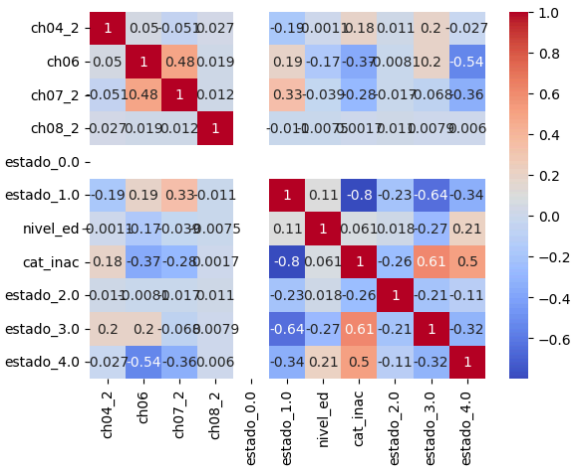


En ambos casos se presentó una mayor cantidad de participantes del género femenino.

Posteriormente se realizó dos matrices de correlaciones entre las variables filtradas para el año 2004 y 2024



(Matriz de correlación del año 2024)



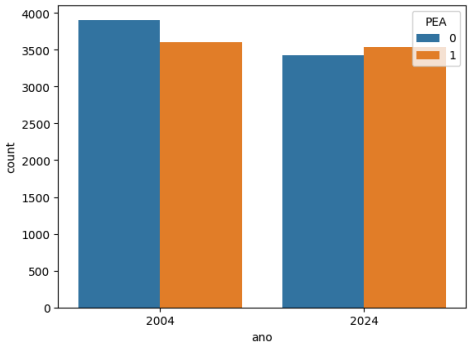
(Matriz de correlación del año 2004)

En ambos casos se puede observar una correlación negativa entre cat_inac y estar en un estado de ocupado (-0.75 en 2024 y -0.8 en 2004), en ambos es esperable debido a que cat_inac porque una persona está inactiva laboralmente, por lo que es esperable que si una persona está actualmente trabajando no debería tener una explicación de porqué está inactiva laboralmente. Por otro lado, en la matriz de correlación de 2004 se puede observar que no hay datos de “estado 0.0” esto implica que todos respondieron la encuesta en 2004.

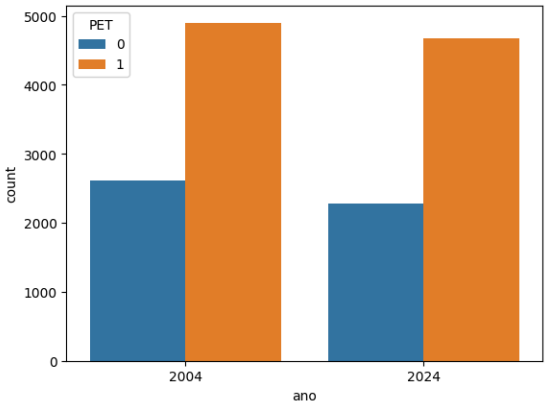
Luego de unir las bases de datos de 2004 y 2024 se buscó la cantidad total de personas desocupadas e inactivas dentro de la base. Se encontró que había un total de 839 desocupados y un total de 5459 inactivos dentro de la muestra. Posteriormente se analizó el promedio de ingreso per cápita familiar dentro de las categorías de estado laboral encontrando que en promedio los ocupados ganan 106460 pesos, los inactivos 63898 pesos y los desocupados 31655.

Para los posteriores análisis se separó la base en aquellos que respondieron su estado civil y aquellos que no. Los siguientes análisis son con la población que respondieron la pregunta.

En primer lugar, se analizó que tantas personas estaba económicamente activa, separando a aquellos que están ocupados o desocupados y aquellos que no, analizando anualmente. Se encontró que en 2004 había una cantidad mayor de personas inactivas mientras que en el 2024 había una mayor cantidad de personas activas. En el total de la muestra se encontró 7330 inactivos y 7141 activos



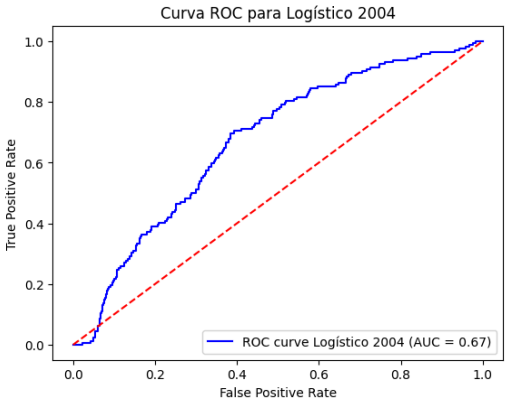
Se analizó el total de personas que estaban en edad para poder trabajar en cada año, definiendo la edad laboral entre 15 a 65 años. Se encontró que en ambos años una mayor cantidad de personas en edad de trabajar, con un total de 9577 personas en edad y 4894 que no.

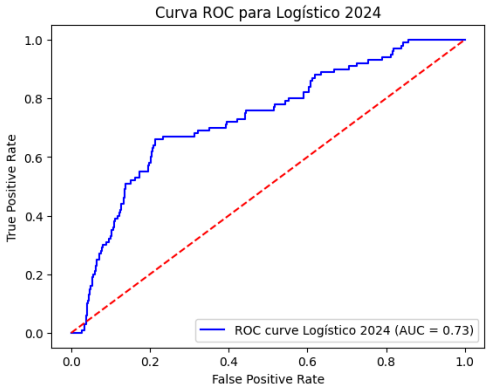


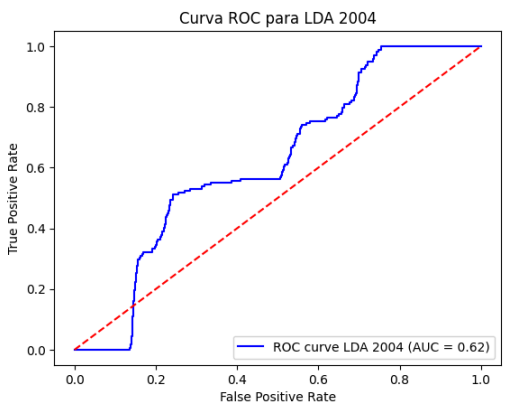
Finalmente se comparó cuantas personas había desocupadas en 2024 y 2004. Se encontró que en 2004 había 528 personas desocupadas y en 2024 había 311.

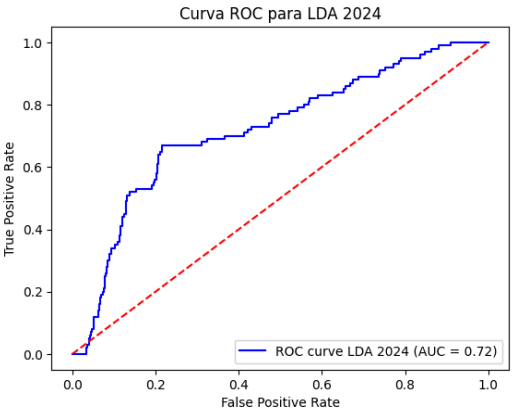
Análisis estadístico

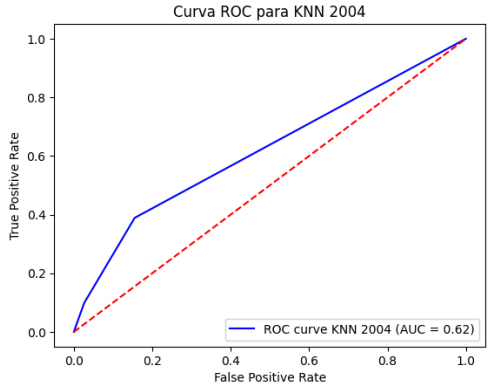
Para el análisis estadístico se realizaron cuatro modelos diferentes para los dos años dentro de la base de datos, una regresión logística, un análisis discriminante lineal (LDA), un knn utilizando como k a 3, y un naive bayes. Para cada uno de ellos se realizó las curvas roc, la matriz de confusión, los valores de AUC y la accuracy.

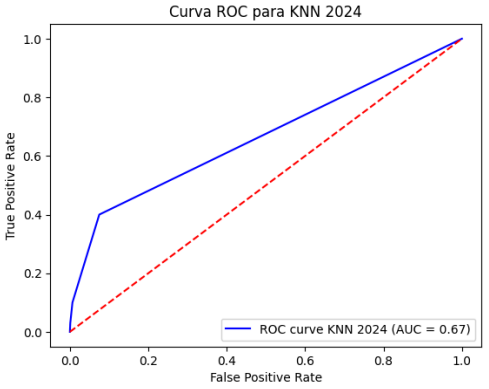


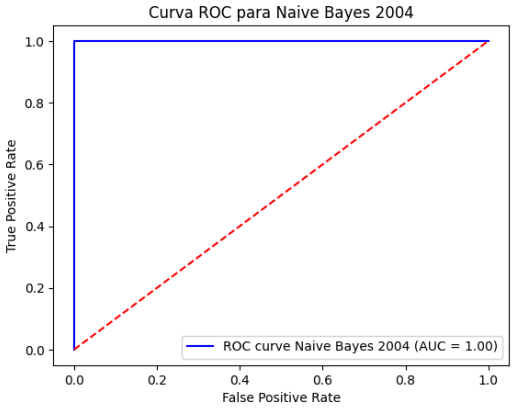


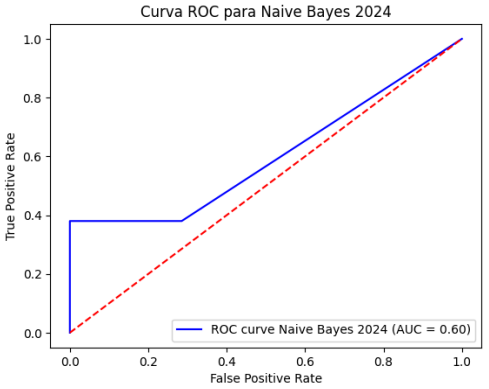





Curva ROC para KNN 2004
True Positive Rate
False Positive Rate
ROC curve KNN 2004 (AUC = 0.62)


Curva ROC para KNN 2024
True Positive Rate
False Positive Rate
ROC curve KNN 2024 (AUC = 0.67)


Curva ROC para Naive Bayes 2004
True Positive Rate
False Positive Rate
ROC curve Naive Bayes 2004 (AUC = 1.00)


Curva ROC para Naive Bayes 2024
True Positive Rate
False Positive Rate
ROC curve Naive Bayes 2024 (AUC = 0.60)