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 análisis 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 (Figuras 1 y 2).

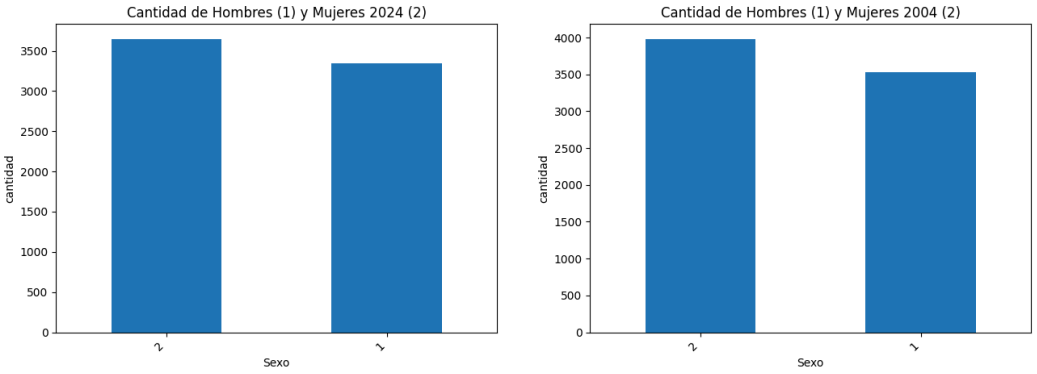
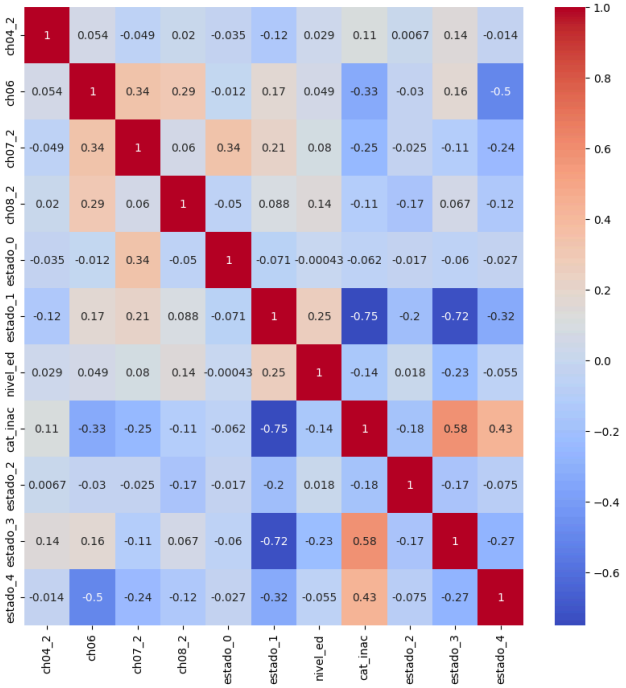


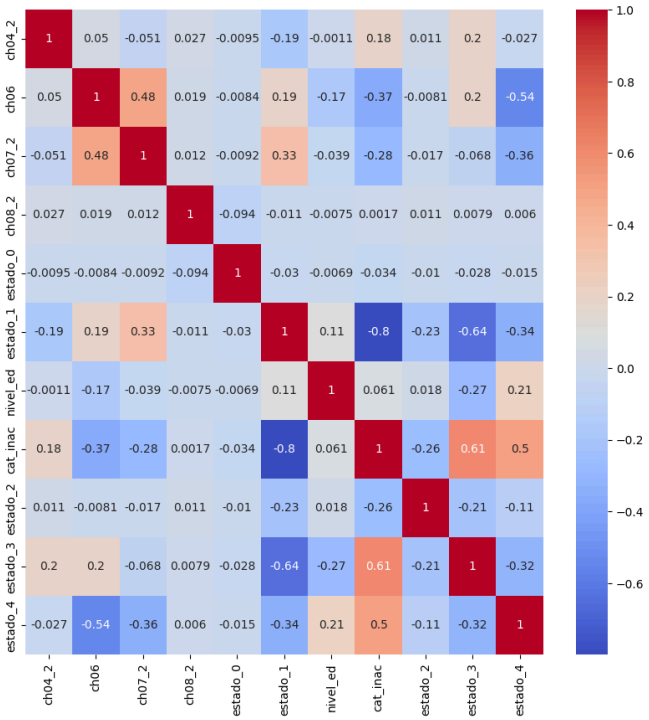
Figura 1 y 2: Gráfico de barras que representan la cantidad de personas que respondieron el cuestionario, separadas por sexo

En ambos casos se presentó una mayor cantidad de participantes del género femenino. Con 3651 mujeres en 2024 y 3984 en 2004

Posteriormente se realizó dos matrices de correlaciones entre las variables filtradas para el año 2004 y 2024 (Figuras 3 y 4)



(Figura 3: Matriz de correlación del año 2024)



(Figura 4: 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 la mayoría de otras correlaciones significativas resultan de interacciones similares entre variables por ejemplo la correlación positiva entre estar inactivo y tener una categoría de inactividad (estado_4 y cat_inac)

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, para este análisis se multiplicó el ingreso de 2004 por 633.97, para poder compararlo con los resultados de 2024. Se encontró que en promedio los ocupados (Estado = 1) ganan 253629 pesos, los inactivos (Estado = 3) 166360 pesos y los desocupados (Estado = 2) 120976 pesos, lo que resulta interesante debido a que implica que aquellos que están inactivos ganan en promedio una mayor cantidad de dinero que aquellos que están activamente buscando trabajo (los desocupados).

Para los posteriores análisis se separó la base en aquellos que respondieron su estado civil y aquellos que no (Estado = 0). Los siguientes análisis son con la población que respondieron la pregunta.

En primer lugar, se analizó qué tantas personas estaban económicamente activas (PEA), 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 (3896 personas inactivas y 3606 personas activas) mientras que en el 2024 había una mayor cantidad de personas activas (3424 personas inactivas y 3535 personas activas) (Figura 5).

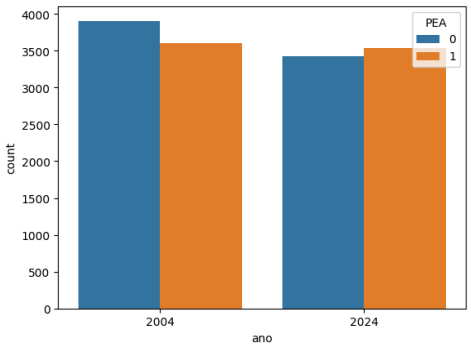


Figura 5: Gráficos de barras de PEA anuales

Se analizó el total de personas que estaban en edad para poder trabajar (PET) 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 4893 personas en edad en 2004 y 4675 en 2024. (Figura 6)

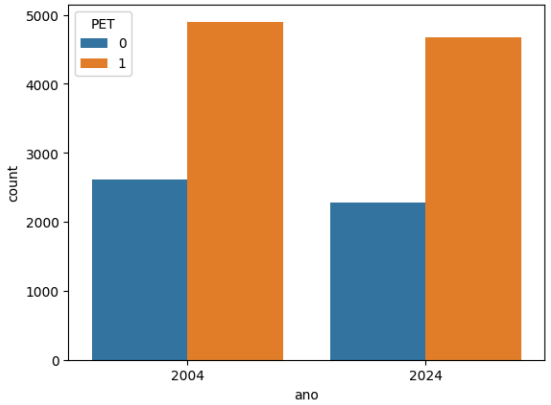
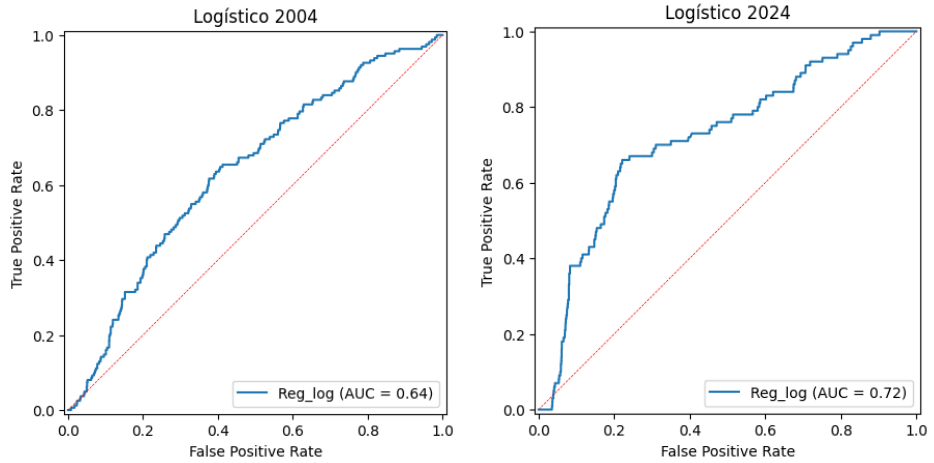


Figura 6: Gráficos de barras de PET anuales

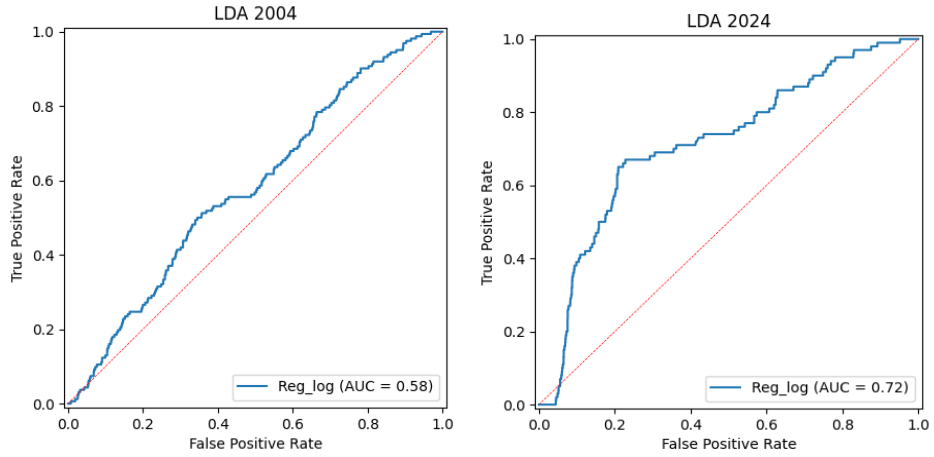
Finalmente se comparó cuántas personas había desocupadas en 2024 y 2004. Se encontró que en 2004 había 528 personas desocupadas y en 2024 había 311. Esta diferencia se utilizará para los posteriores análisis estadísticos

Análisis estadístico

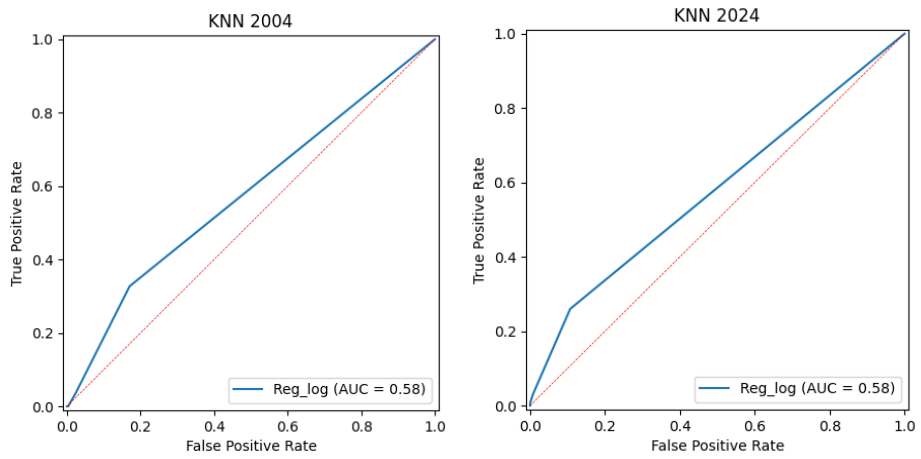
Para el análisis estadístico se realizaron cuatro modelos diferentes para los dos años dentro de la base de datos utilizando las variables utilizadas para la matriz de correlación excluyendo cat_inac y estado, la primera debido a que solo diferencia la razón por la cual una persona está inactiva y la segunda debido a que se trata de la variable con la que se formó Y (desocupados), con cada uno de los valores de estado no desocupados siendo 1 dentro de la variable y, por ende es irrelevante. Se realizaron cuatro modelos diferentes por año: 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 (Figuras 7,8,9,10,11,12,13,14).



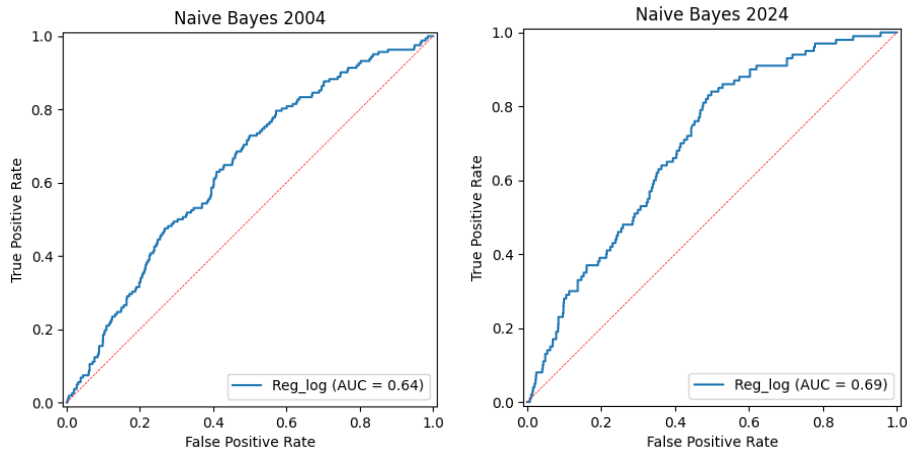
(*Figura 7 y 8: Curvas ROC de los modelos logísticos de 2004 y 2024*)



(*Figura 9 y 10: Curvas ROC de los LDA de 2004 y 2024*)



(*Figura 11 y 12: Curvas ROC de los KNN de 2004 y 2024*)



(*Figura 13 y 14: Curvas ROC de los Naive Bayes de 2004 y 2024*)

Comparación de resultados

Entendiéndose a la Accuracy cómo la que mide la proporción de instancias clasificadas correctamente y el AUC (Área bajo curva ROC) cómo la capacidad de distinción entre clases. Utilizando estas variables se analizó cuál era el mejor modelo para cada año, con los resultados siendo presentados en la siguiente figura (Figura 15)

2004:

	Modelo	Precisión	AUC
0	Regresión Logística	0.928032	0.637645
1	LDA	0.928032	0.579811
2	KNN	0.908041	0.577472
3	Naive Bayes	0.928032	0.637367

2024:

	Modelo	Precisión	AUC
0	Regresión Logística	0.952107	0.722382
1	LDA	0.952107	0.716577
2	KNN	0.946360	0.577077
3	Naive Bayes	0.952107	0.694384

(*Figura 15:* Tabla de comparación anuales de los modelos)

Debido a los resultados obtenidos se puede observar que en ambos años el modelo logístico resulta el mejor, ya que tienen la mejor accuracy (el valor de precisión con 0.928032 en 2004 y 0.952107 en 2024) o la misma que los otros modelos, sin embargo en ambos casos el AUC resultó superior a la del resto de los modelos (0.637645 en 2004 y 0.722382 en 2024)