Ciencia de Datos

Trabajo Práctico № 3

Ignacio Spiousas

Noelia Romero

Ignacia Auguste

Florencia Reverendo

Trinidad Morán

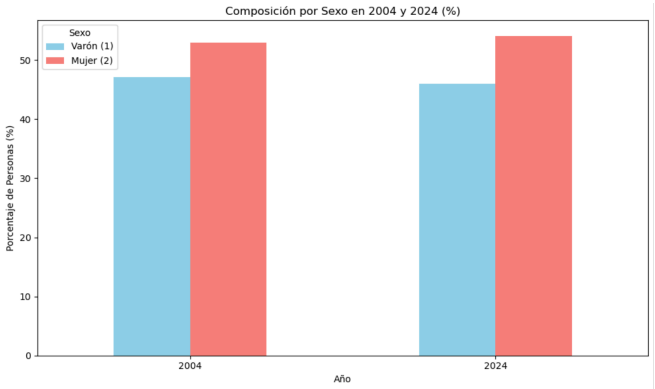
Parte I: Analizando la base

Ejercicio 1:

Según el INDEC, se considera una persona desocupada a aquella que no tiene empleo pero está en busca activa de uno y está disponible para trabajar. Asimismo, este concepto excluye a las personas con trabajos transitorios, a quienes trabajan menos horas de las necesarias y a quienes suspendieron la búsqueda de empleo.

Ejercicio 2:

Durante este trabajo, se utilizaron dos bases de datos: la Encuesta Permanente de Hogares (EPH) del primer trimestre de 2004 y 2024. Luego, se creó una base de datos conjunta, utilizando únicamente los datos correspondientes a hogares de la Ciudad Autónoma de Buenos Aires o del Gran Buenos Aires, que resultó en una base de datos de 14698 datos. A continuación, sobre esta base conjunta, se conviertieron los datos de las columnas CH04 (sexo), CH07 (estado civil), CH08 (cobertura médica), NIVEL_ED (nivel educativo), ESTADO (condición de actividad), CAT_INAC (cateogría de inactividad) a variables categóricas. Después, se eliminaron aquellos datos sin sentido de las columnas CH06 (edad), PP03D (cantidad de ocupaciones), PP08D1 (cuánto cobró por mes) y P21 (total de ingresos habituales de la ocupación de la persona) resultando en una base de 13701 datos. En este caso, por datos sin sentido se entiende a cualquier dato negativo, debido a que todas estas variables representan datos que deben ser números positivos. Se identificaron los datos faltantes a través de matrices de datos faltantes, contemplando también dentro de esta categoría las observaciones "Ns/nr". Los datos faltantes de las columnas referentes a estado civil (variable CH07, 13 datos faltantes); cobertura médica (variable CH08, 31 datos faltantes); e ingreso per cápita (IPCF, 1 dato faltante) se eliminaron al no responder a un patrón que explique su ausencia y ser relativamente pocos para el tamaño de la base. Sin embargo, la variable nivel de edad contaba con 135 datos faltantes. En el control de estos datos, para no perderlos, se calculó la mediana de la edad por nivel educativo. Luego, en cada una de las observaciones incompletas en CH06 (edad), se observó el nivel educativo y se le asignó la mediana de edad para completar la columna.

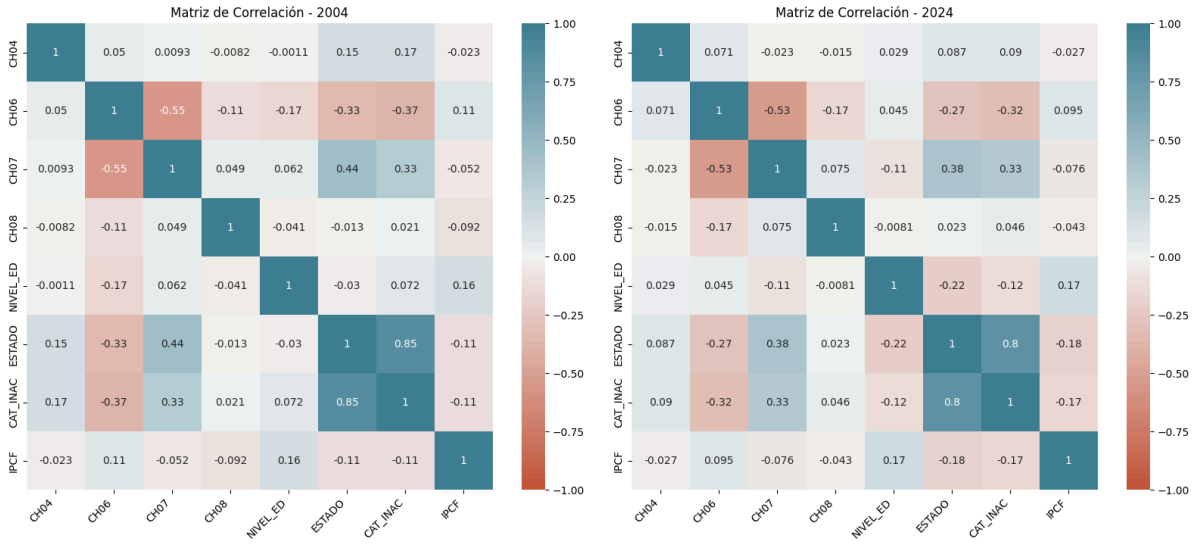


Luego de realizar la limpieza de la base conjunta, se realizó un gráfico de barras que muestra la proporción de hombres y mujeres que componen la encuesta en 2004 y 2024. La proporción se mantuvo estable, en el 2004 se compuso de 47.1% de hombres y 52.9% de mujeres, y en el 2024 la proporción fue de 46% de hombres y 54% de mujeres. Ambas encuestas están bien balanceadas en género.

A continuación, se realizó una matriz de correlación para cada año de las variables CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC e IPCF (monto de ingreso per cápita familiar). En ambas matrices, los colores cálidos indican una correlación positiva y los fríos, una negativa. Además, la intensidad del color crece a medida que la correlación es más fuerte.

De la matriz de correlación de 2004, es posible observar que las variables CH06 y CH07 están fuertemente relacionadas de forma negativa (-0,55) y que las variables ESTADO y CAT_INAC están fuertemente relacionadas de manera positiva (0,85). Luego, hay ciertas variables que están débilmente relacionadas de manera negativa, como CH06 y ESTADO (-0,33) y CH06 y CAT_INAC (-0,37). A partir de esta relación, es esperable que las variables ESTADO y CAT_INAC están débilmente relacionadas de forma positiva con CH07 (0,44 y 0,33, respectivamente). De estos resultados es posible concluir que la edad de la persona está negativamente correlacionada con su estado civil. Esto se debe a que los valores adoptados por la variable CH07 más altos son los que corresponden a la soltería, mientras que los más bajos corresponden a estar en pareja. Por lo tanto, significa que a menor edad de la persona, mayor es la probabilidad de que sea soltera. Igualmente, la correlación positiva entre ESTADO y CAT_INAC significa que la condición de actividad de una persona está relacionada con cuál es su clase de inactividad (jubilado, amo/a de casa, etc).

De la matriz de correlación de 2024, es posible observar que los mismos tipos de correlación entre las mismas variables que las del 2004, pero con coeficientes de correlación distintos. Para CH06 y CH07 es -0,55, para ESTADO y CAT_INAC es 0,8, para CH06 con ESTADO Y CAT_INAC son -0,27 y -0,32, respectivamente, y para CHO7 con ESTADO y CAT_INAC son 0,38 y 0,33, respectivamente. Por lo tanto, el análisis realizado previamente también es aplicable a la matriz de 2024.

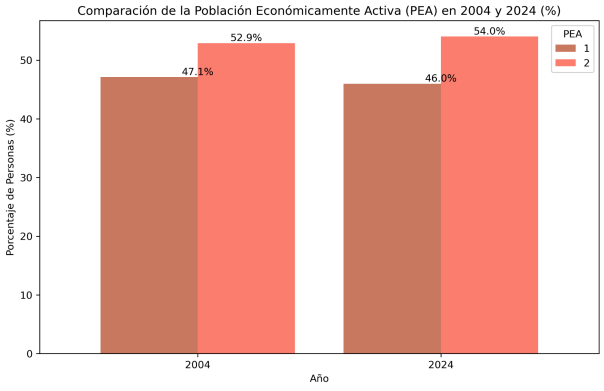


En la base conjunta, es posible observar que 839 de los datos de la muestra son "desocupados" y 5462, "inactivos". Asimismo, se calculó la media del IPCF para los desocupados, los ocupados y los inactivos de la muestra, que dió $31.655,96, $125.240,39 y $63.863,08, respectivamente.

Ejercicio 3:

A continuación, se creó una variable llamada "no_respondieron" que contara la cantidad de respuestas faltantes de la variable ESTADO, para así saber cuántas personas no habían respondido su condición de actividad. Esto dio que 51 personas no habían respondido esta pregunta en la encuesta. Luego, se crearon dos bases de datos distintas, una llamada "respondieron" con todos los datos que habían respondido esta pregunta y otra llamada "norespondieron", con los que no.

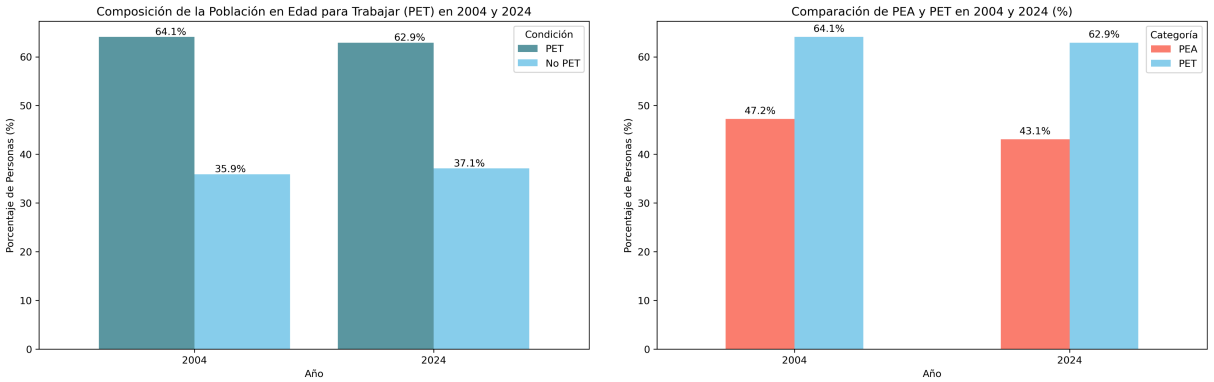
Ejercicio 4:



Luego, a la base de datos "respondieron" se le agregó una columna llamada "PEA" (Población Económicamente Activa) que toma el valor 1 si es desocupado u ocupado en ESTADO. Se expone un gráfico de barras que muestra el porcentaje de PEA del 2004 y del 2024. En él, es posible observar que en ambos años hubo una menor cantidad de PEA que de no PEA, y la proporción es muy parecida en ambos años. Esto quiere decir que la proporción de población económicamente inactiva en 2004 y en 2024 era mayor a la activa.

Ejercicio 5:

Igualmente, a la base "respondieron" se agregó una columna llamada "PET" (Población en Edad para Trabajar), que tomaba el valor 1 si la edad se encontraba entre 15 y 65, y 0 si estaba fuera. Luego, se realizó un gráfico de barras para la PET de ambos años, como también uno comparativo entre el PET y el PEA para cada encuesta.



En cuanto a la PET, en ambos años la muestra se compone de más de un 60% de población que está en edad laborable. Sin embargo, la población económicamente activa está por debajo del 50% tanto para el 2004 como para el 2024.

Ejercicio 6:

Luego, se calculó la cantidad de personas desocupadas hubo en cada año. Para ello, se agregó una columna a la base "respondieron" llamada "DESOCUPADO" que vale 1 si en la columna ESTADO decía que era desocupada y 0 si no. Después, se sumó la cantidad de desocupados que eran de 2004 y de 2024, que dió 528 y 311, respectivamente.

También, se observó la proporción de desocupados que hubo en cada nivel educativo en ambos años. En ambos casos es posible ver algo curioso, las dos mayores proporciones de desocupados son aquellos con un título secundario completo o un título superior universitario incompleto. Particularmente, en el caso de 2004 es posible ver como la proporción de desocupados aumenta a medida que el nivel educativo es mayor hasta el título superior universitario incompleto; luego, decrece. Si bien en 2024 el comportamiento es similar, tiene sus particularidades. La primera es que la proporción de desocupados con primario completo es mayor a la de secundario incompleto. La segunda, el punto donde comienza a haber un decrecimiento en la proporción de desocupados es cuando el nivel educativo es de secundario completo. Otra particularidad que cabe destacar es que la proporción de desocupados por nivel educativo en 2004 es mayor a la del 2024, con excepción de aquellos con el título secundario completo.

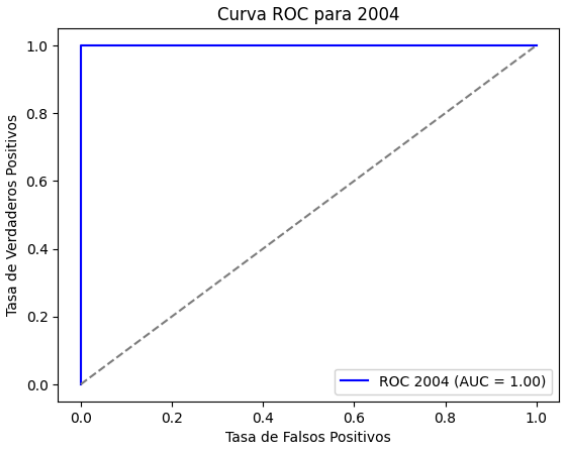
Asimismo, se realizó el mismo análisis pero por rango de edad, enfocándose en si hubo cambios en la proporción entre los años. Se creó una columna llamada "EDAD_GRUPO" que creaba separaciones por edad en intervalos de 10 y se calculó la proporción de desocupados en cada uno de

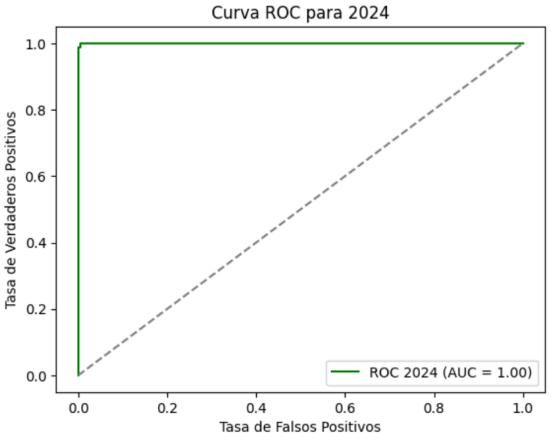
estos intervalos según el año. Al realizar las proporciones, es posible ver que la proporción de desocupados en 2004 en todos los rangos de edad es mayor a la de 2024, con excepción del rango de 40 a 49, en el que es mayor en 2024, y los rangos de 0 a 9 y de 90 a 99, donde son iguales en ambos años (en ambos es cero).

Parte II: Clasificación

Se realizaron cuatro métodos distintos para intentar predecir los datos del año 2004 y 2024 de la base de datos "respondieron": regresión logística, análisis discriminante lineal, KNN y Naive Bayes. Con ellos se calcularon la matriz de confusión, la curva de ROC, los valores de AUC y las de *accuracy* para cada base.

3. A partir de estos cuatro cálculos se decidió que el mejor método de predicción para cada año fue el de regresión logística. Esto se debe a que la curva de ROC está en los ejes, el AUC dio 1, la *accuracy* dio 1 y la matriz de confusión para el 2004 no tuvo errores y para el 2024 tuvo muy pocos errores. Como el valor máximo del AUC y de la *accuracy* son 1, esto muestra que la regresión logística para este caso es muy buena. Además, la curva de ROC ideal debería estar tocando el eje de arriba a la izquierda, que en ambos casos lo hace. Por último, la matriz de confusión muestra la cantidad de casos que este método clasificó de manera correcta, que en el caso del 2004 no hubo errores y en el 2024 hubo tan solo 6 de 1804 datos a clasificar, que es el 0,33% de los casos.





4. Para evaluar el rendimiento del modelo de clasificación en la base de datos "norespondieron", se aplicó el modelo previamente entrenado para cada uno de los conjuntos de datos por año, con el objetivo de predecir la variable objetivo "DESOCUPADO". Con este fin, se creó la variable "DESOCUPADO" de manera análoga a la creada para la base "respondieron" en la primera parte del trabajo. Los resultados obtenidos incluyen una matriz de confusión que muestra únicamente un valor en la clase negativa ("[[41]]"), lo cual indica que todos los registros fueron clasificados como "no desocupado". El valor de *accuracy* alcanzó el 100%, ya que todas las predicciones coinciden con los valores reales de la variable dependiente. Sin embargo, debido a que no hubo ninguna muestra

positiva (individuos desocupados) en los datos, no fue posible calcular el área bajo la curva ROC (AUC), generando un valor indefinido (`nan`). Estos resultados reflejan una limitación en la evaluación del modelo para esta muestra en particular, ya que la ausencia de casos positivos impide un análisis más completo de su capacidad predictiva para ambas clases. Esto sucedió para las muestras de ambos años, ya que la base de datos no contenía ninguna observación en la que "DESOCUPADO" valiera 1.