Ciencia de Datos

Trabajo Práctico N° 3

Profesora: Maria Noelia Romero

Tutor: Ignacio Spiousas

Gabriel Alejandro Díaz y Gabriela Belen Sanchez

3 de noviembre de 2024

Parte I: Análisis Descriptivo

1.1 Identificación de personas desocupadas

El INDEC identifica a las personas desocupadas mediante la EPH, una encuesta que se realiza de forma telefónica o presencial. Esta se aplica sobre hogares en 31 aglomerados urbanos, considerados como una fracción representativa del país, con el fin de conocer sus características sociodemográficas y socioeconómicas que luego son generalizadas.

1.2 Procesamiento de la Base de Datos

1.2.1 Filtrado de Observaciones y limpieza de datos.

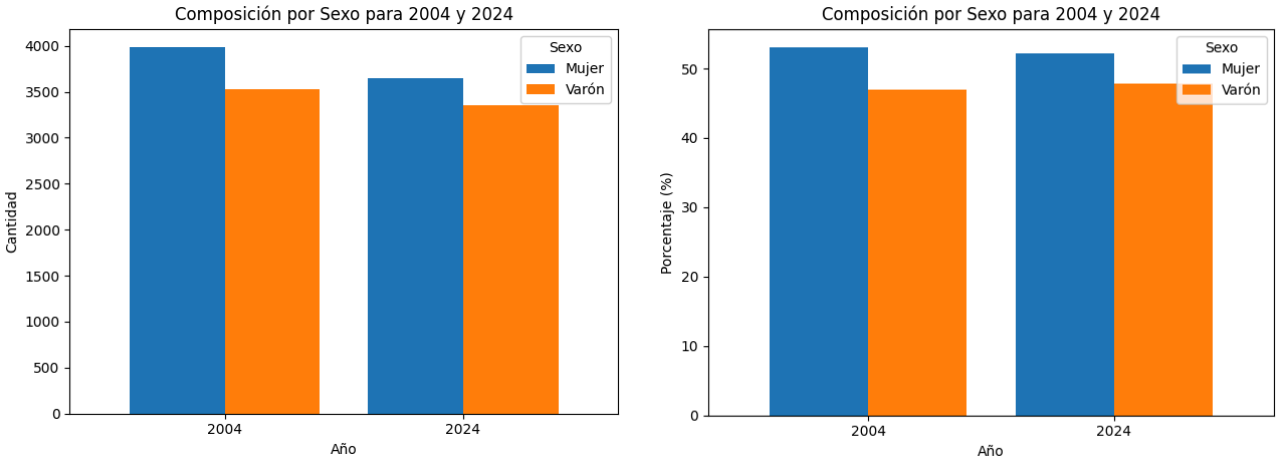
Con el fin de analizar tendencias específicas, como primer paso se filtró la base de datos para que contenga aquellas observaciones de los aglomerados: Ciudad Autónoma de Buenos Aires (CABA, "32") y Gran Buenos Aires (GBA, "33"). Se creó una sola base de datos que contenga las observaciones del primer trimestre de 2004 y 2024 (n = 14512) para el análisis de:

- **Eliminación de valores negativos o anormales**: Se verificó la presencia de valores negativos en columnas numéricas como "ch06" (edad) o "ipcf" (ingreso per cápita familiar). Aquellos encontrados fueron eliminados debido a su naturaleza ilógica. A su vez, se descartaron aquellas entradas sin valor numérico en edad por ser muy jóvenes (menores a 1 año) o de edad muy avanzada (mayores a 98). A los primeros se los descartó debido a que su análisis no agrega información sobre el desempleo, mientras que a los segundos no se los consideró por la poca exactitud de los datos.
- **Columnas con valores únicos o innecesarios**: Para reducir el tamaño de la base de datos en cuanto a sus columnas, se eliminaron aquellas que contenían valores repetidos para todas las observaciones, causado por la inclusión de solo dos aglomerados), como "región" o "mas_500"; o por el periodo de tiempo, cómo "trimestre". Similarmente, se eliminaron columnas que por su contenido no eran útiles para el trabajo actual, como "CODUSU" o "nro_hogar", ya que no se tiene por objetivo el identificar hogares particulares.
- **Manejo de valores nulos**: Se identificaron las columnas que contienen valores *Nan* y se descartaron aquellas con una gran cantidad.. Algunas de estas se trataban de columnas que no eran compartidas entre las bases de datos de 2004 y 2024, mientras que otras contenían información dependiente de otra pregunta. Posteriormente, sobre todo al plantear el modelo de regresión en la sección de *Clasificación*, se revisaron las celdas con datos faltantes. En aquellos casos donde se presentaba un Nan y este no podía ser reemplazado con información proveniente de otra columna, por un "0" (No corresponde) o por algún valor más específico, se eliminaron. También se descartaron las columnas con preguntas condicionales al estado (ya sea ocupado o desocupado), ya que no tiene sentido su introducción para la predicción.

1.3 Composición por Sexo

Como se puede apreciar en la Figura 1, se puede apreciar una leve tendencia hacia una mayor representación femenina, presente en ambos años. Específicamente, en 2004 la muestra se encuentra compuesta por un 53.03% (n = 3984) de mujeres y un 46.97% (n = 3528) de hombres; mientras que en 2024, las primeras representan 52.15% (n = 3651) de la muestra, en contra de un 47.85% (n = 3349) masculino.

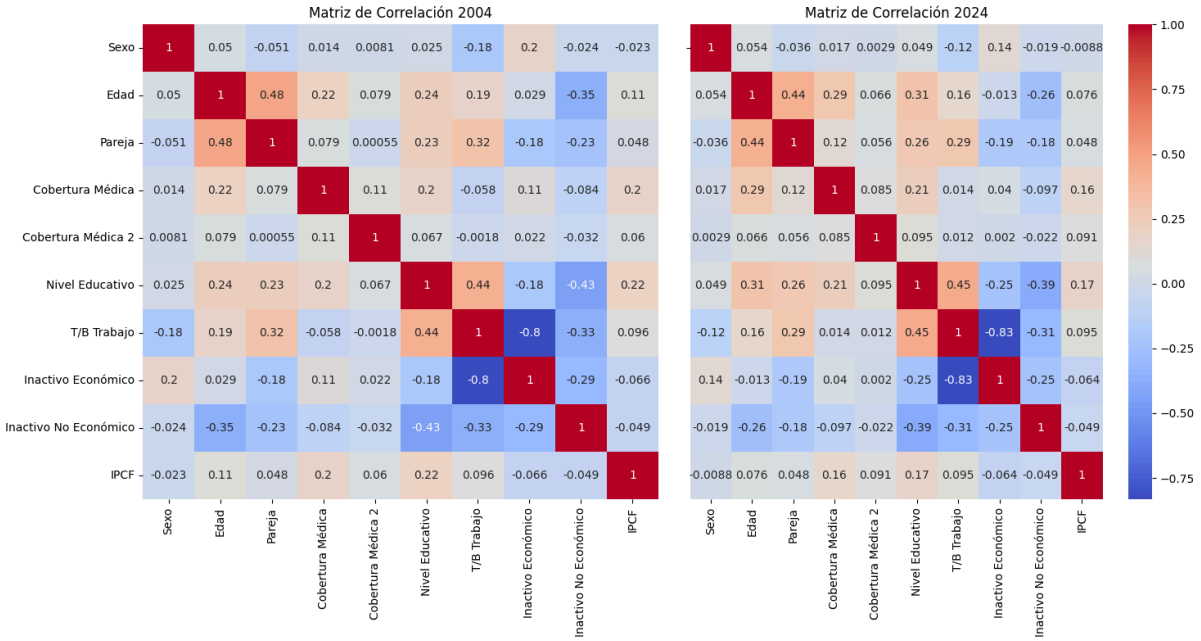
Fig 1: Composición por sexo según el año



1.4 Matriz de Correlación

A continuación se observa la matriz de correlación para variables clave en 2004 y 2024 (Figura 2). Debido a su naturaleza categórica, para facilitar su visualización y su interpretación, se optó por la creación de nuevas variables que resumieron la información. Específicamente, se utilizaron: Sexo (1 si es mujer, 0 si es varón). Edad, Pareja (1 sí "unida" o "casada"; 0 si "soltera", "viuda" o "separada/divorciada"), Cobertura Médica (1 si presenta algún tipo de cobertura médica; 0 si no), Cobertura Médica 2 (1 si tiene más de un tipo de cobertura; 0 si no), Nivel Educativo (máximo nivel educativo alcanzado, ordinal), T/B Trabajo (1 si "ocupado" o "desocupado", 0 si "inactivo" o "menor de 10 años" en Estado), Inactivo Economico (1 si "Jubilado", "Rentista o Pensionado", "Estudiante" o "Ama de casa" en Categoria Ec; 0 si no en Categoría de Inactividad), Inactivo no Económico (1 si "Discapacitado", "Menor de 6 años" u "Otro" en Categoría de Inactividad; 0 si no), y IPCF (Ingreso Per Capita Familiar, ajustada según la inflación para 2004[1]).

Fig 2: Matriz de correlación para cada año.



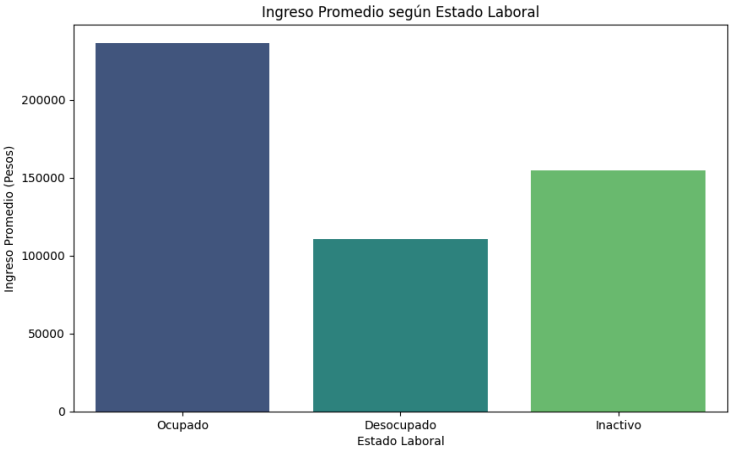
[1] La tasa de inflación utilizada (55.959,57%) fue calculada de febrero de 2004 a febrero de 2024 (https://calculadoradeinflacion.com).

De esta forma, las relaciones entre variables más relevantes observadas son:

- Se observa una correlación elevada entre Edad y Pareja en ambos años (0.48 y 0.44), lo cual indica que cuanto mayor es la edad, mayor es la chance de que la persona se encuentre casada o unida. También, se encuentra relacionada positivamente con Cobertura Médica (0.29) y Nivel Educativo (0.31), por lo que cuanto mayor es la edad, más probable es que la persona tenga cobertura media y alcance un amyor nivel educativo. Por otro lado, la Edad también se encuentra relacionada, pero negativamente, con Inactivo no Económico (-0.35 y -0.26). Esto es esperable por la categoría "Menor de 6 años". De forma similar, Nivel Educativo tiene una correlación negativa con esta última (-0.43 y -0.39), posiblemente por la misma razón.
- La variable IPCF solo presenta correlaciones relativamente altas con Cobertura Médica y Nivel Educativo en ambos años (2004: 0.2 y 0.22; 2024: 0.16 y 00.17, respectivamente), de forma tal que cuanto mayor es el ingreso familiar, más chances de tener algún tipo de cobertura médica y mayor el nivel educativo de la persona. Sorpresivamente, en ambos años, las variables T/B Trabajo y las de tipo de Inactividad tienen una correlación muy cercana a 0 con IPCF, de forma que no parece haber una relación entre ambas.
- Finalmente, se presenta una fuerte correlación entre T/B trabajo e Inactivo Económico debido a la naturaleza de las mismas variables.

1.5 Desocupados, Inactivos y Promedio de IPCF

Figura 3: IPCF promedio por Estado



Con respecto a la variable "Estado", la muestra cuenta con un total de 6302 personas con ocupación (**2004**: 3078; **2024**: 3224), 839 desocupadas (**2004**: 528; 2024: 311) y 5459 inactivas (**2004**: 2797; **2024**: 2662). Específicamente, el IPCF promedio de aquellas personas con ocupación es de $236.569,61, el de desocupados es de $110.622,63 y el de inactivos $154.491.47.

1.6 Personas sin respuesta sobre Condición de Actividad

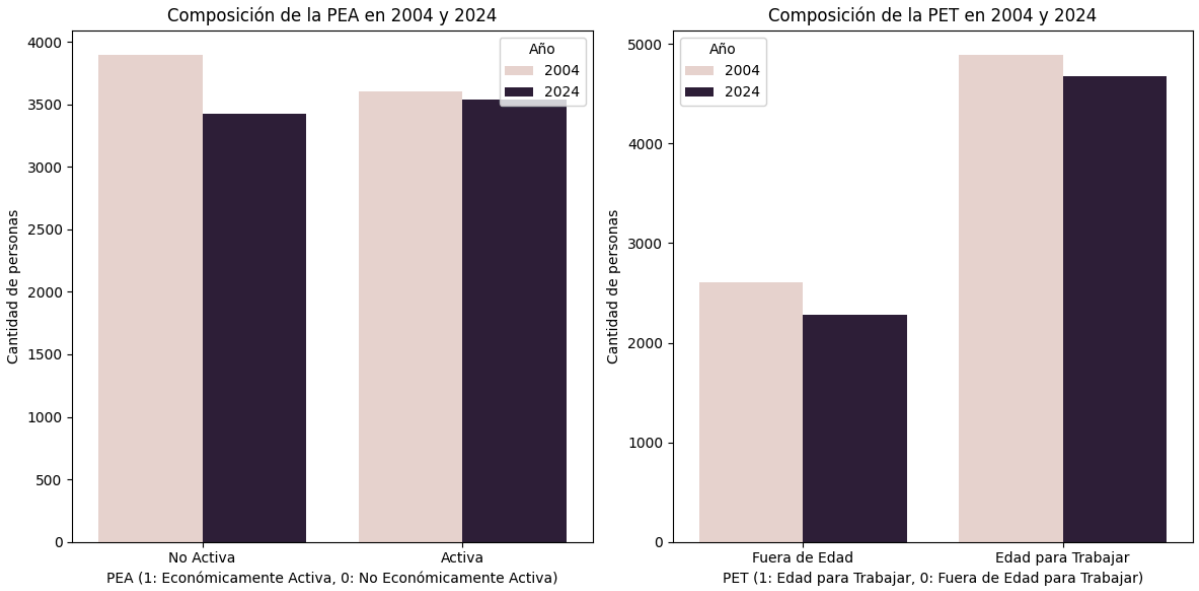
La base de datos es capaz de identificar la cantidad de personas que optaron por no responder cuál es su condición de actividad ("estado" = 0). Específicamente, en esta base de datos hubieron 51 personas sin respuesta, en comparación con 14461 que sí dieron esta información. En base a esto se creó una base de datos diferente para aquellos que respondieron ("respondieron.csv") y que no ("norespondieron.csv").

1.7 Población Económicamente Activa (PEA) vs. Población en Edad para Trabajar (PET)

El concepto de Población Económicamente Activa (o PEA) hace referencia a las personas que tienen trabajo, o que, si no lo tienen, se encuentran en su búsqueda activamente. Es decir, se encuentra compuesto por la población ocupada y desocupada, pero no por la inactiva. En este caso, la PEA en 2004 representa un 51.93% (n = 3896) de la muestra; mientras que en 2004 un 49.20% (n = 3424) (Figura 4). Por otro lado, la Población en Edad para Trabajar (PET) incluye a aquellas personas que tienen entre 15 y 65 años cumplidos. Específicamente, la proporción de PET en la muestra de 2004 es de 65.22% (n = 4893), mientras que en 2024 es de 67.17% (n = 4685) (Figura 4). De esta forma, en

ambos casos se puede apreciar como las proporciones se mantienen muy similares, con diferencias muy pequeñas en ambos indicadores.

Figura 4: Composición PEA y PET en 2004 y 2024



1.9 Desocupados por nivel educativo y por edad

Al profundizar en el nivel educativo de la muestra, es posible apreciar cómo en ambos años, aquellas personas con estudios secundarios completos y con estudios universitarios incompletos presentan las mayores proporciones de desocupación. Específicamente, en 2004 el mayor es Universidad Incompleta con un 13.57% de desocupados, seguido por Secundaria Completa con un 12.66%. Similarmente, en 2024 la de mayor proporción de desocupados es Secundaria Completa (8.11%), seguida por Universidad Incompleta (6.97%). Esta reducción de las proporciones de desocupación entre años puede apreciarse en todos los niveles educativos (Figura 5).

Figura 5: Proporción de Desocupados por Nivel Educativo

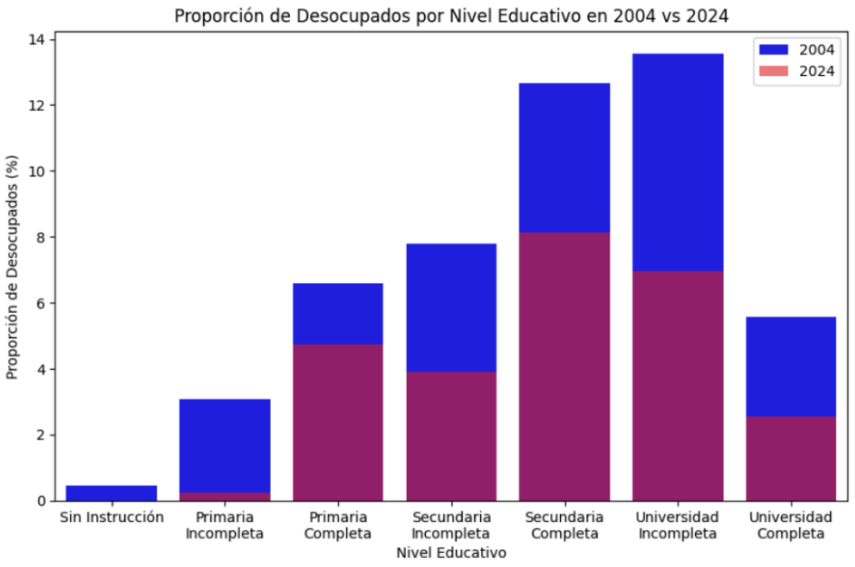
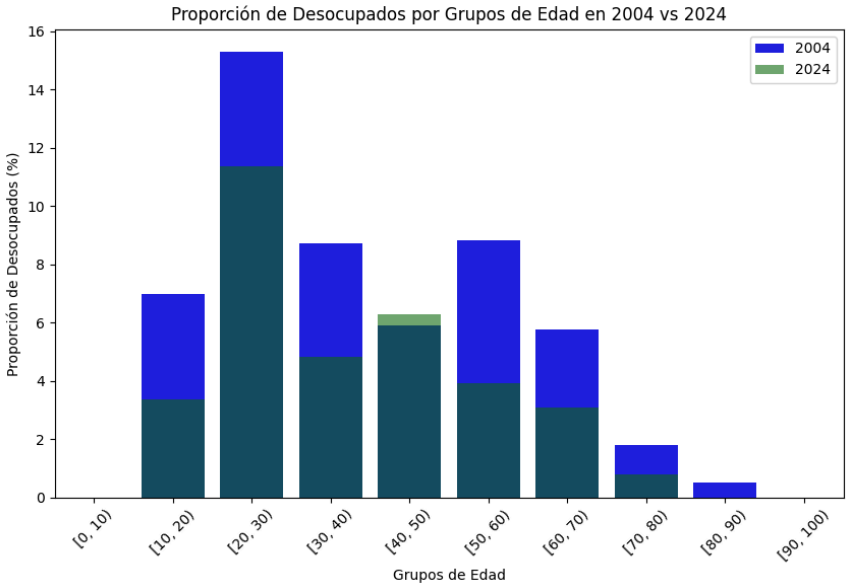


Figura 6: Proporción de Desocupados por Rango Etario en 2004 y 2024

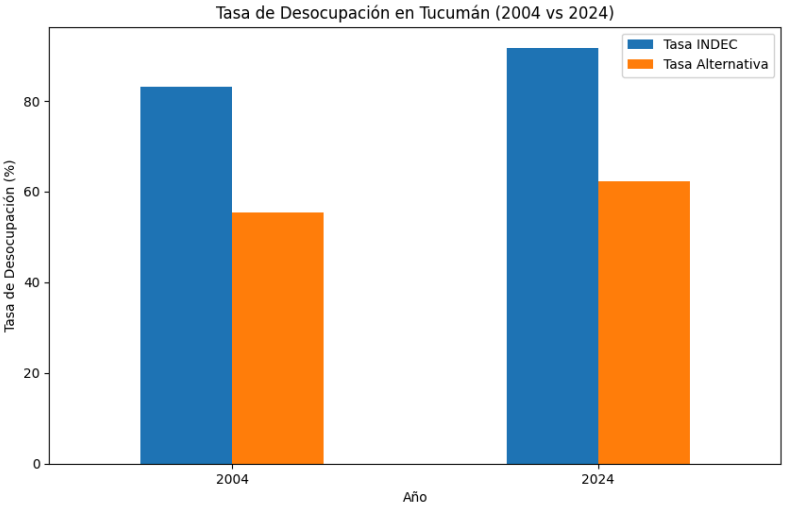


Por otro lado, al observar la proporción de desocupación en cada rango etario, es posible apreciar como el mayor porcentaje se encuentra en ambos casos en el rango de 20-30 años. En el 2004, los desocupados representan un 15.29%, mientras que en 2024 se ve reducido a un 11.35%. Nuevamente, se puede apreciar una reducción del desempleo según el año en la mayoría de los rangos, excepto en uno. En la población perteneciente al rango de 40-50, puede apreciarse un incremento de la desocupación de un 0.39% (pasó de 5.90% en 2004 a un 6.29% en 2024).

1.9.2. Desocupación Alternativa en Tucumán

La tasa de desocupación alternativa, se diferencia de la tasa del INDEC al considerar la PET en lugar de la PEA en el denominador. De esta forma, para una misma muestra, es posible obtener una proporción diferente de desocupación según la definición utilizada. En este caso, al tomar como referencia a la muestra perteneciente a Tucumán, podemos apreciar cómo en 2004 la Tasa INDEC fue de 83.11% y la alternativa de 55.44%; mientras que en 2024 la primera fue de 91.71% y la segunda de 62.33% (Figura 7). Como se puede apreciar, en ambos casos la Tasa INDEC fue más alta que la alternativa y en ambos casos subieron de 2004 a 2024.

Figura 7: Tasa de desocupación INDEC vs. Alternativa



Estas diferencias se deben a la propia naturaleza de las variables. Por una lado, la PEA al tener en cuenta aquellos individuos trabajando o en busca del empleo, tiene en cuenta la fuerza laboral efectiva. En este caso, es posible identificar mejor a los *desempleados*, el cual se relaciona con la actividad económica; en lugar de los *desocupados*. Pero, por la misma razón se deja de lado a aquellas personas en edad de trabajar que no desean buscar un empleo (son inactivas).

Por otro lado, la PET, al considerar a las personas en edad de trabajar (nuevamente, de 15 a 65 años), presenta una base mayor sobre la cual calcular la tasa, por lo que los valores tenderán a ser menores. En este sentido, si tiene en cuenta a las personas inactivas y su impacto en la tasa, pero por lo mismo diluye el enfoque en los problemas de empleo y desempleo.

Parte II: Clasificación

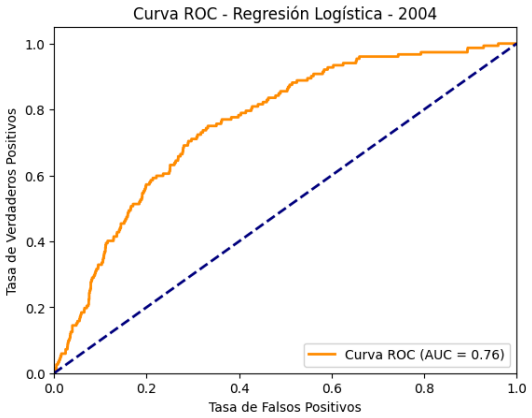
2.1 Preparación y División de la Base de Datos

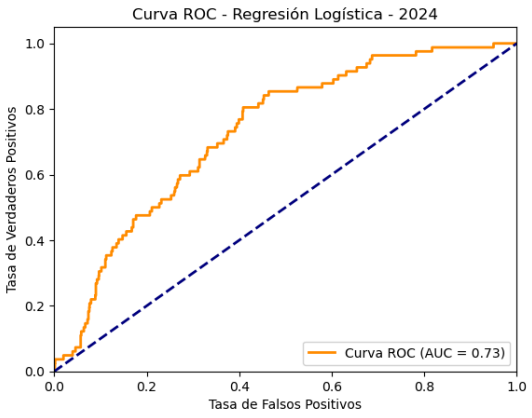
Con el fin de clasificar en base a la variable "desocupado" (toma el valor de 1 si el individuo está desocupado y 0 si no lo está), se dividió a la base de datos según el año en que se realizó la encuesta. Posteriormente, se creó un conjunto de entrenamiento (70%) y de prueba (30%) por cada año (**2004**: *n train* = 5156 y *n test* = 2210; **2024**: *n train* = 3950 y *n test* =1693)[2]. Las variables predictoras (X) incluyen: Aglomerado, Sexo, Edad, Pareja, Cobertura M, Alfabetizado, Nivel Educativo e IPCF. En este caso, se excluyó a la variable Categoría de Inactividad debido a que para ambos casos (ocupados y desocupados) está tomaba el mismo valor ("0", es decir, no aplica).

2.2 Implementación de los Modelos de Clasificación

2.2.1 Regresión Logística

Con la matriz de confusión se observó que en ambos años (2004 y 2024), el modelo mostró una buena capacidad para identificar correctamente a los no desocupados. Sin embargo, no logró identificar a ningún desocupado en 2004 (152 falsos negativos) y en 2024 fueron 82 falsos negativos. Por otra parte, la curva ROC mostró un área bajo la curva (AUC) de aproximadamente 0.76 para 2004 y 0.73 para 2024. Asimismo, se vio que la precisión fue de 93% en 2004 y 95% en 2024, lo que indicaría un rendimiento estable en ambos años. La consistencia de los resultados entre 2004 y 2024 sugiere que el modelo tiene dificultades en identificar correctamente a los desocupados en ambos años.



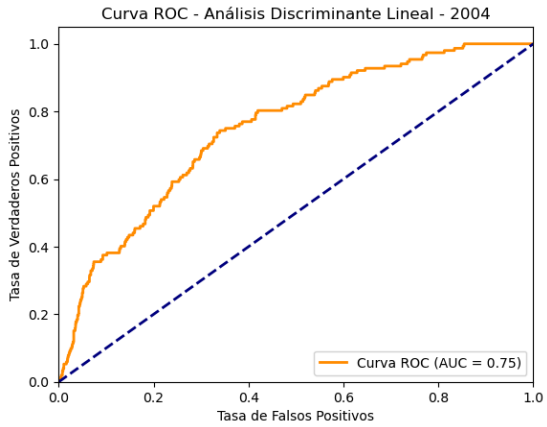


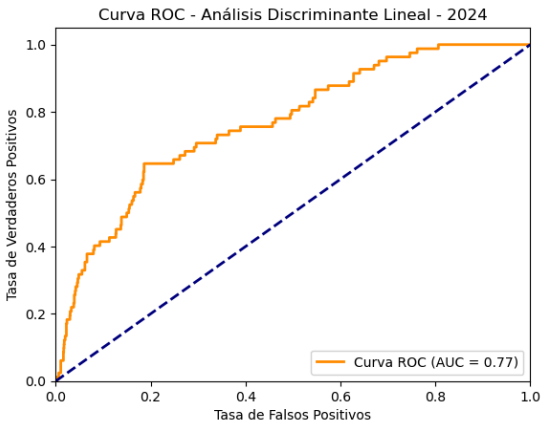
2.2.2 Análisis Discriminante Lineal (LDA)

En la matriz de confusión se observó que los resultados fueron similares a la regresión logística, con una alta tasa de precisión en ambas clases. La clasificación de la clase desocupada no tiene predicciones correctas en 2004 (152 falsos negativos), ni en 2024. Respecto a la curva ROC, se noto que la AUC fue ligeramente superior a la de la regresión logística, donde alcanzó 0.75 para 2004 y

[2] Creados mediante el número de semilla 101

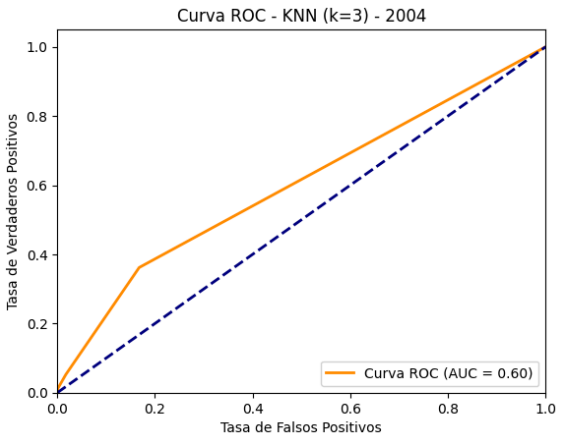
0.77 para 2024. En cuanto a la precisión, se vio que fue de 94% en 2004 y 95% en 2024. A pesar de esto, la mejora en AUC sugiere una mejor capacidad de separación entre clases.

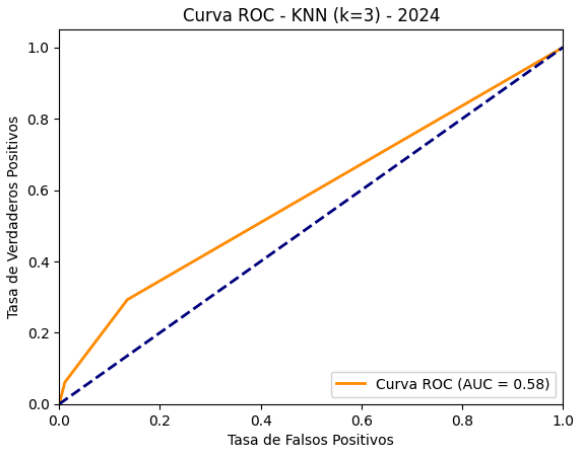




2.2.3 K-Nearest Neighbors (KNN) con k=3

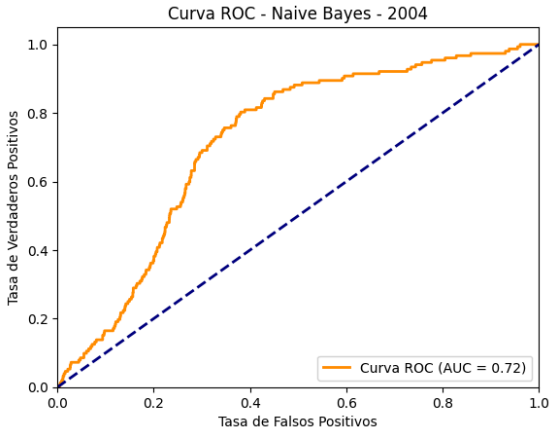
El modelo de K-Nearest Neighbors (KNN) lo que hace es clasificar a cada observación en función de sus tres vecinos más cercanos en el espacio de variables, sin suponer ninguna relación funcional entre las variables y la clase objetivo. En la matriz de confusión se vio que KNN logró identificar a algunos desocupados en 2004 (8 verdaderos positivos) y en 2024 (5 verdaderos positivos), aunque en ambos años hubo falsos negativos y positivos. En la curva ROC se noto que la AUC alcanzó el valor de 0.60 en 2004 y 0.58 en 2024, lo que indica una capacidad discriminativa menor en este ultimo año. Por otro lado, la precisión fue del 93% en 2004 y de 94% en 2024. Los resultados de KNN sugieren que este modelo es sensible a la estructura de los datos (mayor variabilidad), y es menos eficaz en la base de 2024. A su vez, es menos preciso que los modelos lineales.

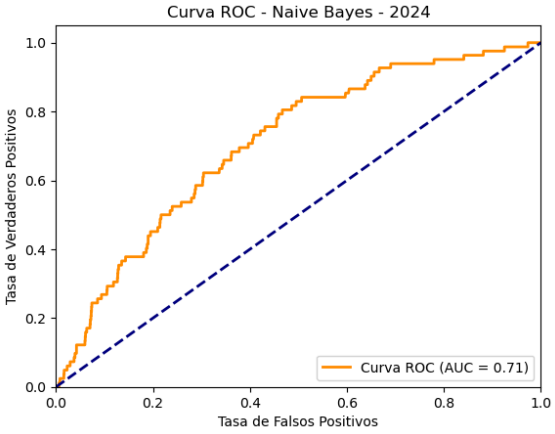




2.2.4 Naive Bayes

El modelo asume independencia condicional entre las variables predictoras y clasifica en función de probabilidades. Con la matriz de confusión se observó que Naive Bayes presentó un rendimiento muy similar al de la regresión logística, tanto en 2004, como en 2024, donde hubo 0 identificaciones precisas para desocupados y clasificación precisa para los no desocupados. Con la curva ROC se vio que la AUC fue baja, aunque algo mayor que en KNN, donde se alcanzó un 0.72 en 2004 y 0.71 en 2024. La precisión fue de 93% en 2004 y 95% en 2024. Aunque Naive Bayes es consistente en su rendimiento en ambos años, su capacidad para clasificar clases es limitada. La suposición de independencia condicional entre variables podría no cumplirse en este conjunto de datos.





2.3 Comparación de Resultados entre 2004 y 2024

Al comparar los modelos para los años 2004 y 2024, se observa que todos los modelos muestran mejoras en 2024, especialmente en términos de exactitud y AUC. Sin embargo, la capacidad de los modelos para identificar correctamente a los desocupados sigue siendo limitada. En ambos años, la exactitud es alta para los modelos de regresión logística, análisis discriminante lineal y Naive Bayes, donde se alcanza 0.93 en 2004 y 0.95 en 2024. Sin embargo, la exactitud no es suficiente para evaluar adecuadamente el rendimiento de los modelos, ya que todos muestran un sesgo hacia la clase mayoritaria (no desocupados). Por otra parte, se vio que el modelo de análisis discriminante lineal tiene el AUC más alto en ambos años (0.75 en 2004 y 0.77 en 2024), lo que indica una mayor capacidad para distinguir entre desocupados y no desocupados. La regresión logística baja en 2024 (AUC de 0.73 frente a 0.76 en 2004) y sigue estando por debajo del análisis discriminante lineal (aunque por poco). Asimismo, la matriz de confusión muestra que, en general, los modelos no logran identificar correctamente a los desocupados. KNN es el único modelo que identifica a algunos desocupados en ambos años, aunque con una precisión y discriminación más baja en comparación con los modelos lineales. En términos de AUC y capacidad discriminativa, el análisis discriminante lineal es el modelo más adecuado en ambos años. Aunque todos los modelos presentan limitaciones en la clasificación de desocupados, el análisis discriminante lineal muestra una mejor capacidad para separar las clases.

2.4 Predicción de Desocupados en la Base No Respondieron

Con el mejor modelo identificado (en este caso, LDA), se predijo la condición de desocupación para la base norespondieron, que contiene aquellas entradas que no dieron respuesta sobre su estado laboral (n = 51). En primer lugar, al aplicar el modelo en los datos del año 2004 (n = 10), se predijo un 0% de desempleo, es decir, todas las observaciones fueron catalogadas como poseedoras de una ocupación. En cambio, al predecir la desocupación en 2024 de estos datos (n = 41), todos ellos se clasificaron como desocupados (100% de desempleados). Específicamente, se cree que este último resultado pudo ser causado porque todas las entradas de 2024 de este dataset contenían un “0” en IPCF.

En general, se cree que este resultado puede indicar que las características de los individuos en "no respondieron" son relativamente homogéneas o que no presentan diversidad suficiente para que el modelo LDA pudiera diferenciar entre ocupados y desocupados. A su vez, debe interpretarse estos resultados con cautela ya que la matriz de confusión no logró identificar ningún desocupado.