Informe TP3 - Ciencia de Datos

Daniela Chejfec, Esperanza Pereyra Iraola, Violeta Juliá

Parte 1: Analizando la base

Para comenzar, importamos las librerías necesarias y cargamos las bases de datos correspondientes a los años 2004 y 2024 en formato .dta y .xlsx, respectivamente. A continuación, se visualizan las primeras filas de la base de datos para entender su estructura y contenido. También identificamos los valores únicos de la columna aglomerado y la lista completa de columnas presentes, que serán útiles para la selección y manipulación de variables. Por otro lado, examinamos algunas de las variables importantes para el análisis, como *ch04* (sexo) y *ch06* (edad), entre otras. Al notar que el campo de sexo (*ch04*) está registrado como un valor de texto en 2004 (con "Varón" y "Mujer"), procedemos a homogeneizar la codificación de esta variable en ambas bases, asignando 1 a "Varón" y 2 a "Mujer", siguiendo el diccionario de variables “Diseño de registro y estructura para las bases preliminares". Este cambio facilitará la integración y el análisis comparativo de datos entre los años.

Para preparar la base de datos destinada al análisis, fue necesario filtrar las observaciones según los aglomerados de interés (Ciudad Autónoma de Buenos Aires y Gran Buenos Aires) y consolidar la información de ambas cohortes en una única estructura de datos. Comenzamos por estandarizar los nombres de las columnas, convirtiéndolos en minúsculas, dado que en la base de datos de 2024 estaban en mayúsculas. Luego, consultamos una vez más el diccionario de variables, donde se detallan los códigos correspondientes a cada aglomerado: 32 para Ciudad Autónoma de Buenos Aires y 33 para Gran Buenos Aires. En la base de 2004, estos valores aparecían como nombres de las ciudades en lugar de códigos numéricos, por lo cual reemplazamos Ciudad de Buenos Aires y Partidos del GBA por 32 y 33, respectivamente, para unificar el formato en ambas bases. Posteriormente, filtramos cada base de datos para conservar exclusivamente las observaciones de estos dos aglomerados y procedimos a unirlas mediante la función concat, consolidando así los datos de ambas cohortes en una única base (denominada base_unida). Por último, verificamos la estructura de la base unida, la cual resultó en un total de 7647 observaciones para 2004 y 7051 observaciones para 2024, lo que confirmó que el filtrado y la unión se realizaron correctamente.

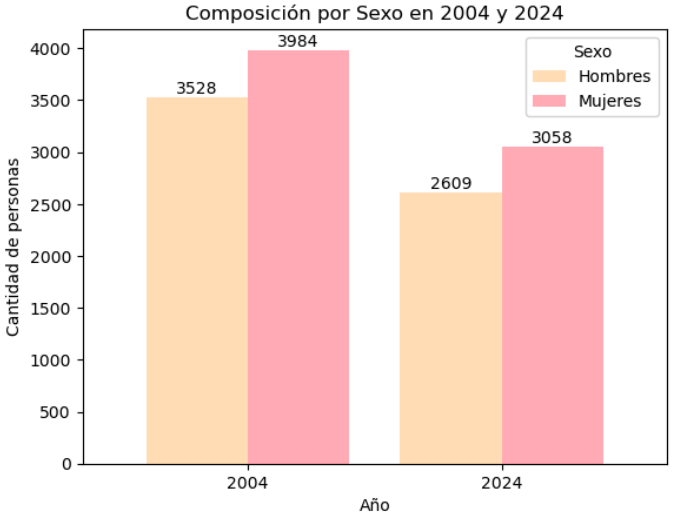
Para garantizar la calidad de los datos y eliminar valores que no tienen sentido, aplicamos un proceso de limpieza en tres etapas. Inicialmente, al inspeccionar la base unida, detectamos varias columnas con valores nulos o inconsistencias: *ch05, imputa, pondiio, pondii y pondih.* Procedimos a evaluar cada una de estas columnas:

La columna *ch05*, que contenía la fecha de nacimiento de los individuos, fue eliminada ya que la información de edad (*ch06*) era suficiente para nuestro análisis. La columna *imputa*, que indicaba si los datos habían sido imputados, contenía valores nulos que interpretamos como datos no imputados, por lo que reemplazamos los valores nulos por 0. Las columnas *pondiio, pondii* y *pondih*, que representan ponderadores de ingresos, contenían únicamente valores nulos en la base de 2004, indicando que estas variables no estaban disponibles para ese año; en consecuencia, eliminamos estas columnas para evitar inconsistencias.

Con esta primera limpieza completada, realizamos una segunda etapa para corregir valores fuera de los rangos esperados en variables clave. Específicamente, nos aseguramos de que *ch06* (edad) estuviera entre 0 y 100, y eliminamos valores fuera de este rango. También filtramos valores negativos en *p47t* (ingreso total individual) y *IPCF* (ingreso per cápita familiar), ya que no tienen sentido en un contexto de ingresos. Asimismo, limitamos los valores de *ch04* (sexo) a 1 y 2, correspondientes a “Varón” y “Mujer”, respectivamente, para mantener la coherencia en la variable de género. Finalmente, observamos que aún había muchas columnas con valores nulos después de la segunda limpieza, por lo cual optamos por eliminar todas las columnas que contenían al menos un

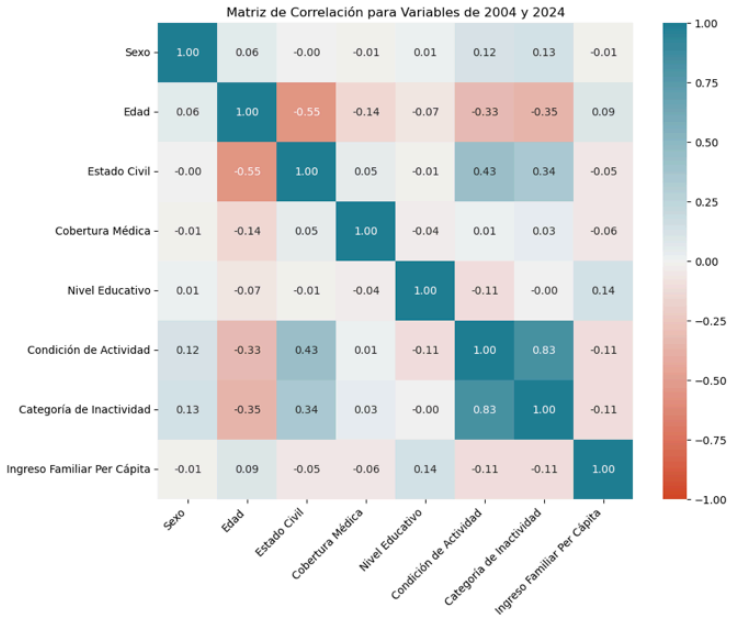
valor nulo. Esto nos dejó con una base de datos final compuesta por 13,179 observaciones y 72 columnas sin valores faltantes, asegurando así una base limpia y consistente para el análisis.

Luego de completar la limpieza de datos, realizamos un análisis de la composición de la población por sexo en los años 2004 y 2024. Filtramos los datos para incluir únicamente estos años y, posteriormente, contabilizamos la cantidad de hombres (*ch04* = 1) y mujeres (*ch04* = 2) en cada cohorte. Para mejorar la claridad visual del gráfico de barras, renombramos las columnas como "Hombres" y "Mujeres" y empleamos una paleta de colores personalizada. También añadimos etiquetas en cada barra que muestran la cantidad específica de personas por sexo. En términos de resultados, los datos revelan una notable disminución en la población total de ambos sexos, pasando de 3,528 hombres y 3,984 mujeres en 2004 a 2,609 hombres y 3,058 mujeres en 2024, lo cual indica una reducción significativa en cada cohorte. Este cambio en los valores totales podría tener implicancias importantes para la estructura poblacional y el mercado laboral de estas áreas. Además, observamos una leve feminización en la composición de la población: la proporción de hombres respecto a mujeres disminuyó de aproximadamente 0.88 en 2004 a 0.85 en 2024. Este cambio podría estar relacionado con factores como una mayor esperanza de vida femenina o una mayor mortalidad masculina en este periodo. La reducción en la población total, junto con el cambio en la proporción por sexo, podría afectar aspectos económicos y sociales, especialmente en lo relativo a la fuerza laboral y los servicios de salud pública.



Para analizar las relaciones entre variables sociodemográficas y de actividad laboral, construimos una matriz de correlación a partir de las variables seleccionadas en la base de datos. Utilizamos un diccionario de etiquetas descriptivas para reemplazar los nombres de las columnas originales con términos más intuitivos, como Sexo, Edad, Estado Civil, entre otros. La matriz de correlación se generó con las variables Sexo (*ch04*), Edad (*ch06*), Estado Civil (*ch07*), Cobertura Médica (*ch08*), Nivel Educativo (*nivel_ed*), Condición de Actividad (*estado*), Categoría de Inactividad (*cat_inac*) e Ingreso Familiar Per Cápita (*ipcf*). Renombramos las filas y columnas de la matriz para asegurar una mejor interpretación y representamos las relaciones visualmente mediante un mapa de calor (Zaric, D., 2019). La Condición de Actividad y la Categoría de Inactividad muestran una alta correlación positiva (0.83), lo cual sugiere que ciertas categorías de inactividad, como la jubilación o los estudios, están fuertemente asociadas con el estado de actividad de las personas. En cuanto a la Edad y el Estado Civil, encontramos una correlación negativa de -0.55, indicando que a medida que aumenta la edad, la proporción de personas solteras disminuye, siendo más frecuente encontrar otros estados civiles en personas de mayor edad. Asimismo, la correlación moderada de 0.43 entre el Estado Civil y la Condición de Actividad indica que ciertos estados civiles, como el

matrimonio o la viudez, tienden a estar asociados con la condición de actividad o inactividad. También encontramos una correlación negativa moderada de -0.35 entre Edad y Categoría de Inactividad, lo cual refleja que, a medida que las personas envejecen, es menos frecuente que se encuentren en categorías de inactividad como estudiante o ama de casa, y más común que pertenezcan a categorías como jubilado/pensionado. Por último, observamos una correlación negativa de -0.33 entre Edad y Condición de Actividad, reflejando que la inactividad aumenta a medida que la población envejece.

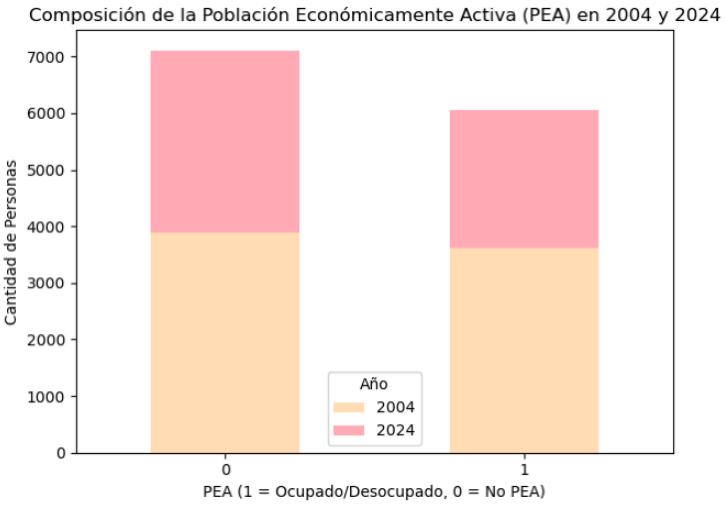


Por otra parte, para evaluar la distribución de la población según su condición de actividad y el ingreso per cápita familiar (*IPCF*) promedio en cada categoría, clasificamos las observaciones en tres grupos: ocupados, desocupados e inactivos. Según el diccionario de variables, la columna estado indica la condición de actividad de cada individuo, con valores de 1 para ocupados, 2 para desocupados y 3 para inactivos. Al analizar los datos, encontramos que hay un total de 5,253 personas ocupadas, 809 desocupadas y 5,246 inactivas en la muestra. Luego, calculamos el ingreso per cápita familiar promedio (*IPCF*) para cada grupo, y obtuvimos que el *IPCF* medio de los ocupados es de $127,719.82, significativamente mayor que el promedio de los desocupados, que es de $32,829.85, y el de los inactivos, que es de $66,492.45. Esta variabilidad en los ingresos entre los distintos grupos sugiere una relación importante entre la condición de actividad y el nivel de ingresos en el hogar, reflejando las diferencias de recursos económicos que puede experimentar cada grupo.

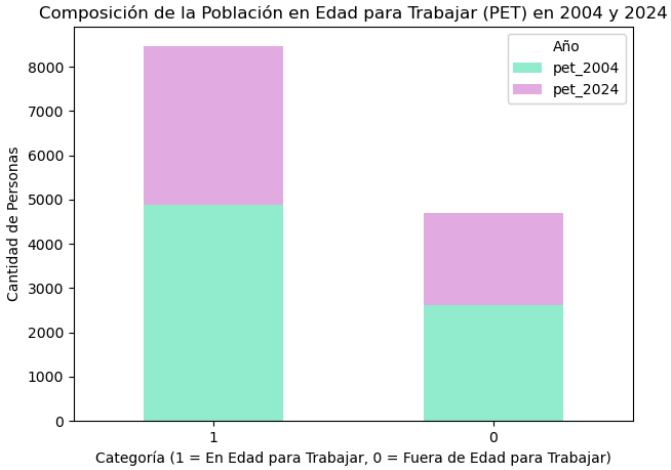
Uno de los problemas recurrentes en la Encuesta Permanente de Hogares (EPH) es la omisión de respuestas en preguntas claves como la condición de actividad laboral, lo cual limita la capacidad de analizar con precisión el estado laboral de la población. En este contexto, identificamos las personas que no respondieron a la pregunta sobre su condición de actividad, almacenada en la columna estado. Los valores 0 en esta columna indican las observaciones con datos faltantes en dicha variable. Al filtrar las observaciones, obtuvimos dos subconjuntos de datos: uno denominado *respondieron*, que incluye todas las observaciones donde se reportó la condición de actividad, y otro llamado *norespondieron*, que almacena exclusivamente las observaciones con estado igual a 0. En total, encontramos que 10 personas no respondieron su condición de actividad. Para facilitar el acceso y análisis de estos datos, guardamos ambas bases en archivos separados.

Agregamos a la base *respondieron* una nueva columna llamada *PEA* (Población Económicamente Activa), la cual identifica a las personas que están ocupadas o desocupadas en la

variable estado (valores 1 y 2, respectivamente) con un valor de 1, mientras que asigna un 0 a quienes no forman parte de la *PEA*. A continuación, realizamos un análisis de la composición de la PEA para los años 2004 y 2024 y representamos los resultados mediante un gráfico de barras. Los datos muestran una disminución significativa en la cantidad de personas en la Población Económicamente Activa, pasando de 3,606 en 2004 a 2,456 en 2024, lo cual podría reflejar una reducción en la participación laboral o cambios económicos desfavorables. Esta reducción en la *PEA* podría tener implicancias importantes, ya que una menor cantidad de personas en el mercado laboral puede derivar en una escasez de mano de obra, afectando tanto la producción como el crecimiento económico. Por otro lado, la cantidad de personas clasificadas como *No PEA* también disminuyó, de 3,896 en 2004 a 3,211 en 2024, lo que indica que en términos absolutos menos personas se encuentran en situaciones de inactividad. Estos cambios en ambas categorías sugieren una evolución en la dinámica laboral y resaltan la importancia de monitorear la participación en el mercado laboral para anticipar y mitigar los posibles impactos en el sistema económico.

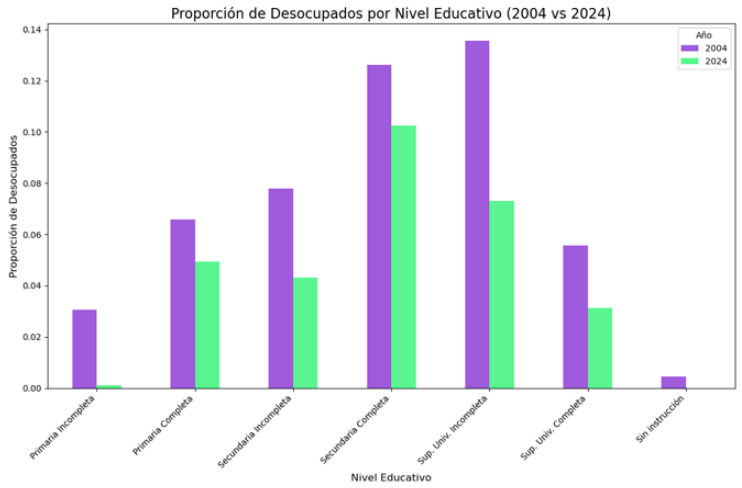


Para identificar a la Población en Edad para Trabajar (*PET*), agregamos una columna a la base respondieron que toma el valor de 1 si la persona tiene entre 15 y 65 años cumplidos y 0 en caso contrario. Realizamos un análisis de la composición de la *PET* para los años 2004 y 2024, comparando la cantidad de personas dentro y fuera de esta categoría en cada cohorte, y visualizamos los resultados mediante un gráfico de barras. En cuanto a los resultados, observamos que la cantidad total de personas en edad para trabajar ha disminuido significativamente de 7,502 en 2004 a 5,667 en 2024, lo que representa una reducción aproximada del 24%. Específicamente, en 2004 había 4,893 personas en edad laboral, cifra que se redujo a 3,582 en 2024, lo que implica una disminución de alrededor del 27% en el grupo considerado disponible para el mercado laboral. Al mismo tiempo, la cantidad de personas fuera de edad para trabajar también se redujo, pasando de 2,609 en 2004 a 2,085 en 2024. Esta disminución es menor, aproximadamente del 20%, indicando que la proporción de personas en edad laboral ha caído de manera más pronunciada en comparación con aquellos fuera de la edad laboral. La disminución observada en la *PET*, especialmente entre aquellos en edad laboral, puede tener implicaciones significativas para la economía, dado que una menor cantidad de personas potencialmente disponibles para trabajar puede reducir la oferta de mano de obra, afectando la productividad y el crecimiento económico. Este cambio en la estructura demográfica resalta la importancia de monitorear la composición de la fuerza laboral y desarrollar estrategias que mitiguen las posibles consecuencias de una población laboralmente activa en disminución.



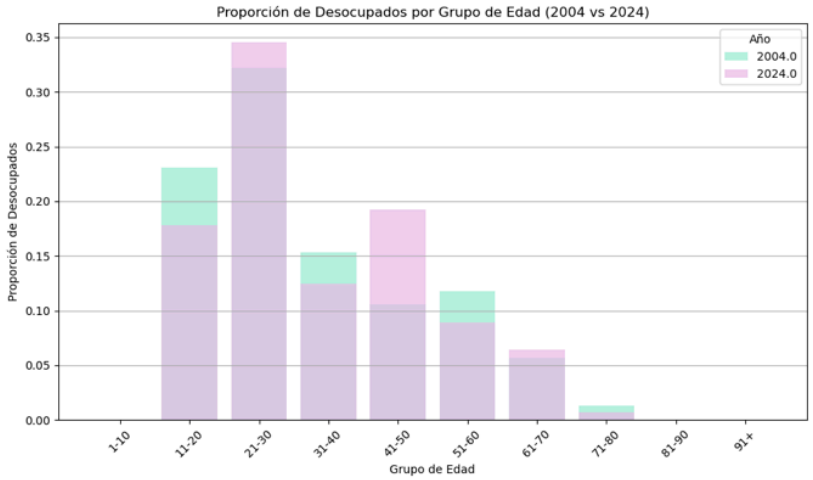
Para analizar la evolución de la desocupación entre 2004 y 2024, agregamos una nueva columna llamada *desocupado* a la base *respondieron*, en la cual se asigna el valor 1 si el individuo está desocupado (según la variable estado, valor 2) y 0 en caso contrario. Así, observamos que en 2004 había 528 personas desocupadas, mientras que en 2024 esta cifra se redujo a 281, representando una disminución significativa de aproximadamente el 47%. Esta reducción sugiere una mejora en el mercado laboral, que podría estar vinculada a un aumento de oportunidades de empleo o a cambios en la dinámica de la oferta y demanda de trabajo.

A continuación, analizamos la desocupación según el nivel educativo para los años 2004 y 2024. Observamos que la proporción de desocupados ha disminuido en todos los niveles educativos entre ambos años. En 2004, las mayores tasas de desocupación se encontraban en los niveles de Secundaria Completa y Superior Universitario Incompleto. En 2024, estas proporciones son menores, destacando una mejora especialmente en el grupo de Secundaria Completa. Sin embargo, los niveles de desocupación entre quienes cuentan con Primaria Completa y Secundaria Incompleta siguen siendo relativamente altos en comparación con otros grupos, lo cual indica que, a pesar de las mejoras generales en el mercado laboral, quienes no completan sus estudios secundarios continúan enfrentando mayores barreras para acceder al empleo.



Adicionalmente, creamos una variable categórica que agrupa a los individuos en intervalos de edad de 10 años, para comparar la proporción de desocupación entre estos grupos en 2004 y 2024.

Los resultados indican que los jóvenes (21-30 años) son los más afectados por la desocupación en ambos años, presentando la mayor proporción de desocupados (32.2% en 2004 y 34.5% en 2024), lo que sugiere la persistencia de dificultades para ingresar al mercado laboral en los primeros años de la vida laboral. Observamos también una disminución en el grupo de 11-20 años, que pasa del 23.1% en 2004 al 17.8% en 2024, lo cual podría ser resultado de una mayor permanencia en el sistema educativo o de programas de apoyo al empleo juvenil. Un dato relevante es el aumento de la desocupación en el grupo de 41-50 años, que crece del 10.6% al 19.2%, lo cual podría indicar dificultades para la adaptación de los adultos con experiencia a las nuevas demandas del mercado laboral. En contraste, la desocupación en el grupo de 51-60 años experimenta una leve mejora, bajando del 11.7% en 2004 al 8.9% en 2024, posiblemente debido a una mayor estabilidad laboral o a la tendencia hacia jubilaciones anticipadas. Finalmente, los grupos de mayores de 71 años presentan niveles de desocupación bajos o nulos en ambos años, lo cual es coherente con su baja participación en el mercado laboral.

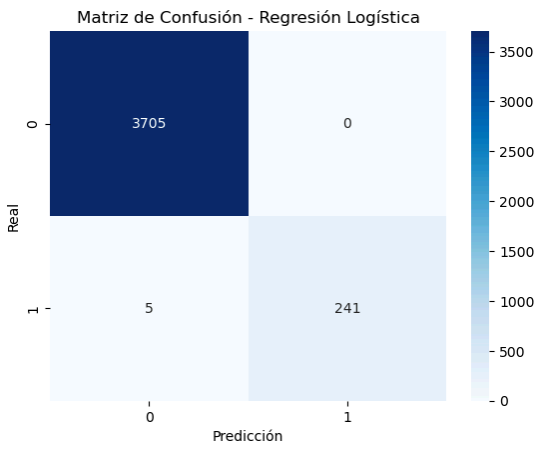


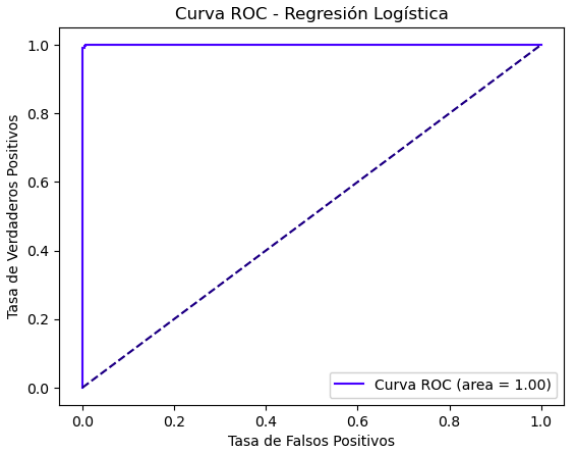
Parte 2: Clasificación

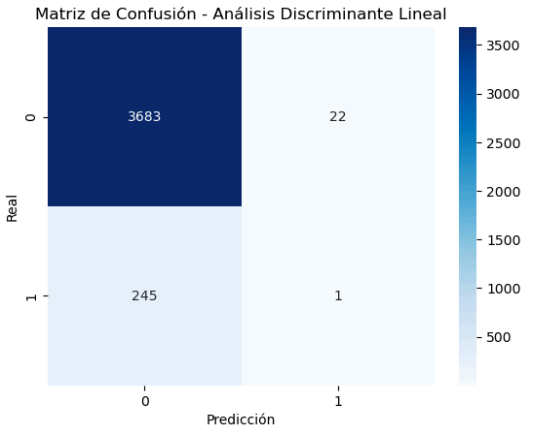
Para llevar a cabo el análisis predictivo, dividimos la base *respondieron* en un conjunto de entrenamiento (*train*) y un conjunto de prueba (*test*), asignando el 70% de los datos al entrenamiento y el 30% a la prueba. Utilizamos el comando train_test_split de sklearn con una semilla fija (*random_state* = 101) para garantizar la reproducibilidad de los resultados. Definimos la variable *desocupado* como nuestra variable dependiente (vector y) y seleccionamos como variables independientes (matriz X) una serie de características sociodemográficas y laborales: Sexo (*ch04*), Edad (*ch06*), Estado Civil (*ch07*), Cobertura Médica (*ch08*), Nivel Educativo (*nivel_ed*), Condición de Actividad (*estado*), Categoría de Inactividad (*cat_inac*) e Ingreso Familiar Per Cápita (*ipcf*). Para facilitar el ajuste de modelos lineales, añadimos una columna de unos (1) a ambas matrices independientes (X_train y X_test) como intercepto. La distribución de los datos resultante fue de 9,218 observaciones en el conjunto de entrenamiento y 3,951 en el conjunto de prueba, asegurando una base robusta para entrenar y validar los modelos predictivos. Estos conjuntos se guardaron en archivos CSV.

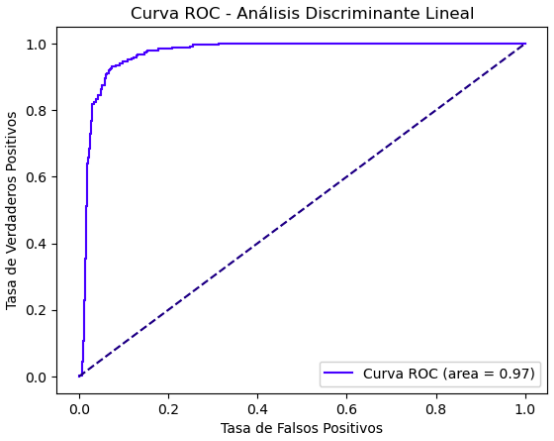
Para evaluar la capacidad predictiva de nuestro modelo de desocupación, implementamos cuatro métodos de clasificación: Regresión Logística, Análisis Discriminante Lineal, K-Nearest Neighbors (K=3) y Naive Bayes. Estos métodos fueron elegidos para explorar enfoques diversos en la predicción del estado de desocupación (variable dependiente) a partir de características sociodemográficas y laborales (variables independientes) presentes en el conjunto de datos. Cada modelo fue entrenado en el conjunto de datos de entrenamiento (*train*) y evaluado en el conjunto de prueba (*test*), y los resultados se analizaron a través de tres métricas principales. En primer lugar, calculamos la matriz de confusión, que nos permite observar los valores verdaderos frente a las

predicciones, evaluando así la capacidad de cada modelo para clasificar correctamente a ocupados y desocupados. En segundo lugar, analizamos la curva ROC y el AUC (Área Bajo la Curva), donde la curva ROC muestra el balance entre la tasa de verdaderos positivos y falsos positivos, mientras que el AUC cuantifica la capacidad discriminativa general del modelo. Por último, calculamos la precisión (accuracy), que indica la proporción de predicciones correctas en el conjunto de prueba, brindando una medida general del rendimiento de cada modelo. En términos de desempeño, la Regresión Logística se destacó con un AUC de 1.00 y una precisión de 100%, demostrando un equilibrio ideal entre sensibilidad y especificidad y clasificando correctamente tanto a ocupados como desocupados. El Análisis Discriminante Lineal (LDA), por su parte, obtuvo un AUC de 0.97 y una precisión del 93%, mostrando un rendimiento preciso y consistente aunque con un leve sesgo en la clasificación de desocupados.

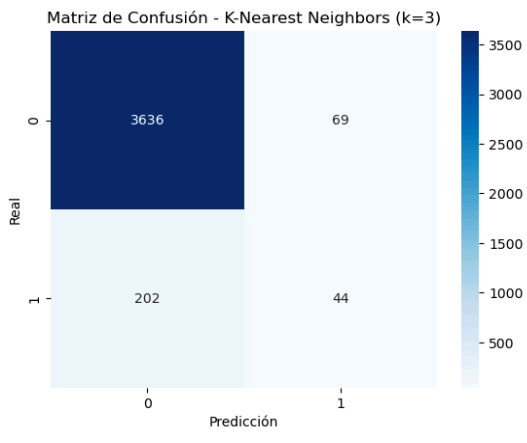


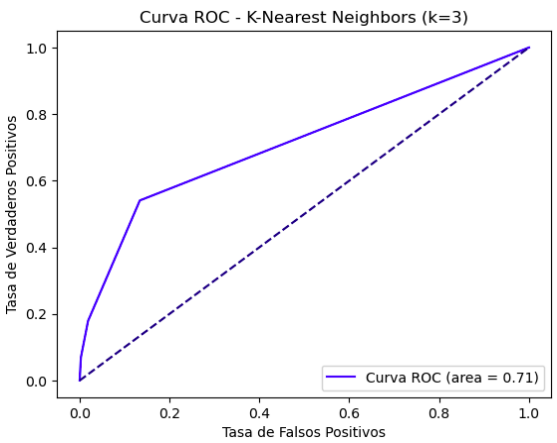


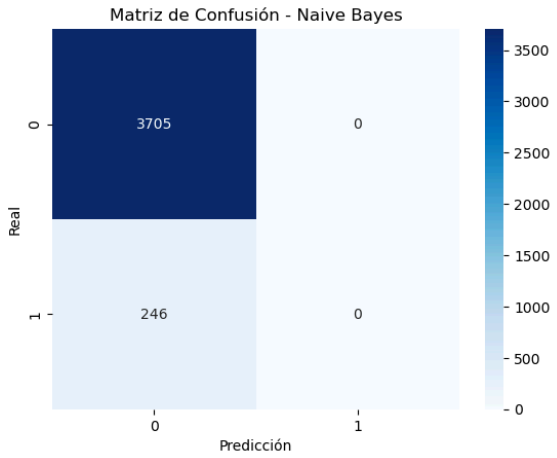


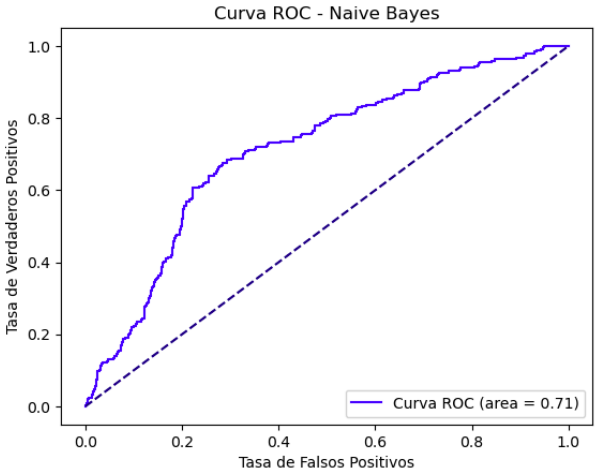


El modelo K-Nearest Neighbors (K=3) alcanzó un AUC de 0.71 y una precisión de 93%. Si bien esta técnica mostró buena precisión, su menor capacidad discriminativa en el AUC refleja que es más sensible a la variabilidad en el conjunto de prueba, lo que puede afectar su estabilidad en la clasificación de desocupados. Por último, el modelo Naive Bayes obtuvo un AUC de 0.71 y una precisión de 94%, mostrando buenos resultados en la clasificación de ocupados, pero algunas limitaciones en la identificación de desocupados.









Al comparar el desempeño de los modelos en la predicción de desocupación para los años 2004 y 2024, encontramos variaciones en la precisión y en el poder discriminativo de cada método. Para ambos años, evaluamos la Regresión Logística, el Análisis Discriminante Lineal, el K-Nearest Neighbors (K=3) y el Naive Bayes en términos de precisión (*accuracy*) y AUC (Área Bajo la Curva ROC), para así determinar cuál modelo se ajusta mejor en cada periodo. En el año 2004, la Regresión Logística y Naive Bayes alcanzaron los mejores resultados, con una precisión del 100% y un AUC de 1.00 cada uno, lo cual indica una capacidad óptima para diferenciar entre ocupados y desocupados sin errores de clasificación. El Análisis Discriminante Lineal (LDA) también mostró un buen desempeño, con una precisión de 92% y un AUC de 0.97, aunque su clasificación fue algo menos precisa que la Regresión Logística y Naive Bayes. Por otro lado, el K-Nearest Neighbors (K=3) tuvo un desempeño más limitado, con una precisión del 92% y un AUC de 0.70, lo que sugiere que este modelo es menos efectivo en la discriminación de las clases en comparación con los otros métodos. Para el año 2024, la Regresión Logística nuevamente se destacó, alcanzando una precisión del 99% y un AUC de 1.00, manteniendo así su efectividad en la clasificación de la variable dependiente. El Análisis Discriminante Lineal también mostró un rendimiento sólido con una precisión de 94% y un AUC de 0.97, resultados similares a los obtenidos en 2004. El K-Nearest Neighbors (K=3) y Naive Bayes también lograron una precisión de 94%, pero sus valores de AUC fueron menores, de 0.64 y 0.71 respectivamente, lo que indica una menor capacidad para discriminar entre ocupados y desocupados en 2024 en comparación con los otros modelos.

Para identificar posibles desocupados en la base de datos *norespondieron*, utilizamos el modelo de Regresión Logística entrenado previamente en la base *respondieron*. Este modelo, que había mostrado el mejor desempeño en la clasificación de ocupados y desocupados, fue ajustado con las características sociodemográficas y laborales disponibles. Al aplicar el modelo sobre *norespondieron*, los resultados indicaron que ninguna de las personas en esta base fue clasificada como desocupada, resultando en una proporción predicha de desocupación del 0.00%. Este valor tiene sentido dado que, al revisar las 10 observaciones presentes en *norespondieron*, observamos que ninguna persona fue etiquetada inicialmente como desocupada. Esta ausencia de desocupados en la muestra de *norespondieron* explica el resultado del modelo y refleja que la base carece de ejemplos de personas desocupadas para clasificar.

Bibliografía

- Zaric, D., (2019). Better Heatmaps and Correlation Matrix Plots in Python. Towards Data Science. April, 15.