Parte I: Analizando la base

El INDEC, a través de la Encuesta Permanente de Hogares (EPH), identifica a las personas desocupadas mediante la evaluación de la condición de actividad de toda la población, sin límite de edad. Una persona es considerada desocupada si es económicamente activa, no tiene trabajo y está disponible y en búsqueda activa de empleo durante el período de referencia. La EPH también registra información sobre trabajo infantil y trabajo no registrado (sin aportes jubilatorios). Esta categoría incluye tanto a aquellos que nunca han trabajado como a quienes no han tenido empleo en los últimos tres años. Además, se capturan datos de personas que participan en asociaciones sin fines de lucro, como sindicatos y ONG.

En primer lugar, para unificar las bases de datos, se utilizó la función `lower` con el objetivo de asegurar que los nombres de las columnas en ambas bases sean consistentes. Luego, se filtraron los datos, mediante la selección de únicamente aquellos provenientes del aglomerado de Buenos Aires y alrededores (es decir, donde [*'aglomerado'*] es igual a 33 o 32). Posteriormente, se concatenaron ambos Data Frames y se revisó la correspondencia de las variables entre ellos. Las columnas que no coincidían fueron eliminadas.

A continuación, se identificaron las variables relevantes para el análisis y se filtró la base de datos combinada, llamada `*df_combinado*`, para que incluyera únicamente las siguientes variables: `[*'ch04', 'ch06', 'ch07', 'ch08', 'nivel_ed', 'estado', 'cat_inac', 'ipcf', 'ano4*']`. Tras esta selección, se utilizó la función `*unique*` para verificar que las variables tuvieran valores consistentes y no contuvieran categorías no deseadas. Una vez confirmado que las variables eran pertinentes y tenían el tipo de dato correcto, se aplicó la función `*describe*` para obtener una visión general de estas. Se realizó una revisión para identificar valores incorrectos, negativos o incoherentes. Los datos negativos en todas las variables fueron eliminados, lo cual resultó en la eliminación de 182 filas.

Finalmente, para las variables *'itf'* (monto de ingreso total familiar), se generaron histogramas con el fin de comprender la distribución de las respuestas, ya que en el análisis preliminar con `describe` se observó que las medias y desviaciones estándar carecían de sentido debido a la presencia de valores de 0.0 (sin ingresos), que podía atribuirse a aquellos que decidieron no declararlos.

Después de realizar la limpieza de los datos, se elaboró un gráfico de barras para analizar la composición por sexo en los años 2004 y 2024 (Fig. 1), con el objetivo de verificar si la población evaluada es comparable en ambos años. Es fundamental que las muestras tengan una proporción similar entre hombres y mujeres, como en este caso, para asegurar la significancia y la validez de los resultados. De otra forma, los resultados podrían no ser comparables. Este análisis aporta robustez a la validez de los hallazgos y permite realizar comparaciones significativas entre los dos períodos.

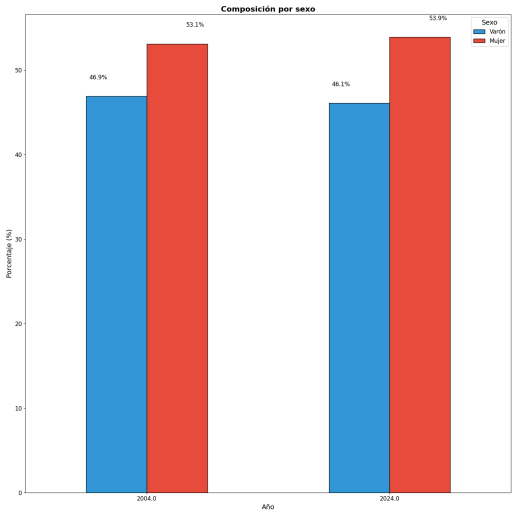


Fig. 1: gráfico de barras con la composición por sexo para 2004 y 2024.

En el proceso de análisis, se transformaron varias variables en *dummies* para simplificar su interpretación en términos binarios, facilitando la comparación entre grupos. A continuación se explican los criterios específicos aplicados para cada variable:

1. **ch06T**: La variable ch06 se transformó en ch06T, una dummy donde el valor 1 indica que la persona tiene más de 25 años y 0 si tiene 25 años o menos. El criterio de 25 años se eligió con base en el informe, que sugiere que los jóvenes generalmente ingresan al mercado laboral alrededor de esta edad.
2. **ch07T**: Se convirtió la variable ch07 en ch07T, asignando el valor 1 a aquellos que están casados o en unión, y 0 a quienes no lo están. Esto permite distinguir entre personas en una relación de convivencia formal y quienes no lo están.
3. **ch08T**: La variable ch08 se transformó en ch08T, donde 1 representa a las personas que cuentan con algún tipo de cobertura médica (obras sociales u otros), mientras que 0 agrupa a quienes no tienen cobertura o no respondieron. Este criterio permite identificar la disponibilidad de cobertura de salud en la población.
4. **nivel_edT**: La variable nivel_ed se recodificó en nivel_edT, con 0 para quienes no completaron la educación secundaria o no tienen instrucción, y 1 para quienes completaron la secundaria o tienen estudios superiores. Esta clasificación refleja el nivel educativo mínimo alcanzado por cada individuo.
5. **estadoT**: La variable estado se transformó en estadoT, asignando 0 a quienes no tienen empleo fijo (incluyendo desocupados y trabajadores informales) y 1 a quienes cuentan con un empleo estable. Este criterio permite diferenciar entre personas con y sin estabilidad laboral.
6. **cat_inacT**: La variable cat_inac se convirtió en cat_inacT, donde 0 agrupa a personas inactivas o dedicadas a actividades no remuneradas (estudiantes, amas de casa) y 1 a quienes son económicamente activos en el mercado laboral. Este criterio se utiliza para segmentar la población según su participación en actividades productivas.
7. **ipcfT**: La variable ipcf se transformó en ipcfT, donde 1 indica ingresos superiores a 368, y 0 ingresos iguales o inferiores a este valor, que representa el promedio general de ingresos en esta población. Esta división facilita la comparación entre personas con ingresos por encima o por debajo de la media.

El análisis de la matriz de correlación, presentado en un mapa de calor, revela que la mayoría de las correlaciones entre las variables son bajas, pero hay algunas excepciones significativas:

En ambos años analizados, se observa una correlación notable entre la edad y el estado civil, con un coeficiente de *r* = 0.53. Este hallazgo es lógico, ya que la edad influye en el estado civil, aumentando la proporción de personas casadas o en pareja a medida que avanza la edad.

Se identifican correlaciones cercanas a *r* = 0.40 entre las categorías de inactividad y el estado de actividad en ambos años. Esto es coherente, dado que ciertas formas de inactividad, como el estudio o el retiro, están relacionadas con el estado de empleo. Igualmente, la correlación entre la edad y el estado de ocupación o desocupación también se sitúa en torno a *r* = 0.40, lo cual tiene sentido considerando que la situación laboral varía significativamente con la edad.

Se observa una disminución exponencial en la correlación entre el ingreso per cápita familiar (ipcf) y la edad. En 2004, esta correlación era de *r* = 0.19, mientras que para 2024 ha disminuido a menos de *r* = 0.01. Este cambio sugiere que la relación entre el ingreso y la edad ha cambiado considerablemente a lo largo del tiempo.

Una tendencia similar se presenta entre el nivel de educación y el ipcf, donde la correlación era superior a *r* = 0.30 en 2004 y ha caído a menos de *r* = 0.01 en 2024. Esta reducción merece un examen detallado para entender las razones detrás de esta transformación y sus implicaciones para la población analizada.

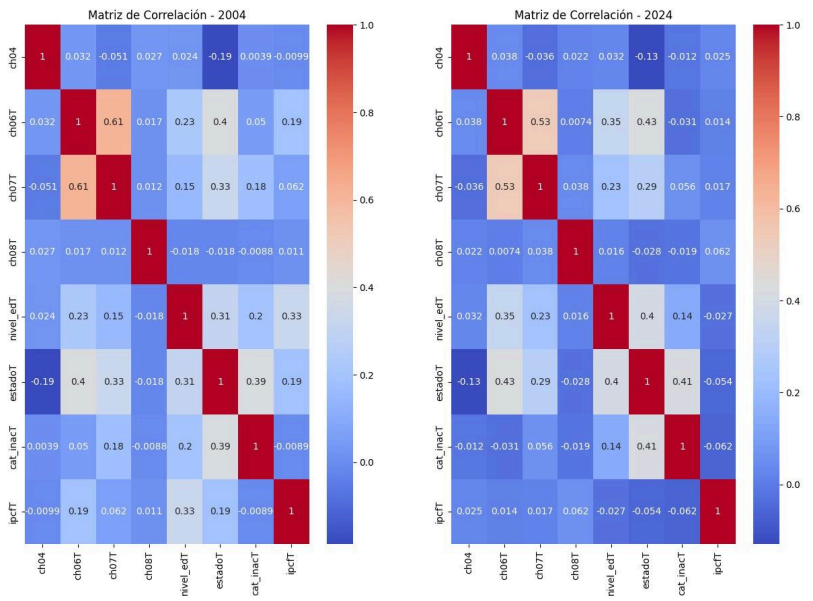


Fig. 2: Matriz de correlación para 2004 y 2024 con las variables: CH04, CH06, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC, IPCF.

La base de datos contiene un total de 6058 personas entre desocupados e inactivos entre ambos años. Mientras que en 2004 se registran 3328 personas desocupadas e inactivas, en 2024 estas suman 2730. Para realizar un análisis más exhaustivo sobre las diferencias de ingresos entre estos grupos en cada año, se calculó la media de ingreso per cápita familiar (IPCF) según el estado de ocupación (ocupado, desocupado, inactivo). (Fig. 3)

Condición de actividad: 0 = Entrevista individual no realizada (no respuesta al Cuestionario Individual); **1** = Ocupado; **2** = Desocupado; **3** = Inactivo; **4** = Menor de 10 años

En el año 2004, los ingresos per cápita familiar (IPCF) muestran valores considerablemente más bajos en comparación con los registrados en 2024, lo que sugiere posibles efectos de la inflación, cambios en el costo de vida o mejoras en el poder adquisitivo para ciertos estados de ocupación.

Los ocupados (estado 1) presentan la media de IPCF más alta en ambos años, con un aumento destacado en 2024, donde alcanzan un valor promedio de 311,698. Esto indica que, aunque todos los grupos experimentan algún crecimiento en sus ingresos, los ocupados han tenido un aumento proporcionalmente mayor. Lógicamente, además de estar influenciado por la inflación, se ve influenciado por el desarrollo y crecimiento laboral.

En contraste, el ingreso de las personas que no realizaron la entrevista (estado 0), se redujo de 52 a 0 en el período de diferencia de 20 años, posiblemente causado por campañas que enfaticen la importancia de realizar las entrevistas.

Por otro lado, los desocupados (estado 2), los inactivos (estado 3) y los menores de 10 años (estado 4) muestran variaciones en el ingreso con respecto a los ocupados, aunque también experimentan un aumento significativo en 2024 (se podría asumir que en gran parte fue causado por la inflación). Se destaca el grupo de inactivos (estado 3), con una media de 315 en 2004 y de 142,072 en 2024. Esta diferencia representa una mejora considerable en el tiempo.

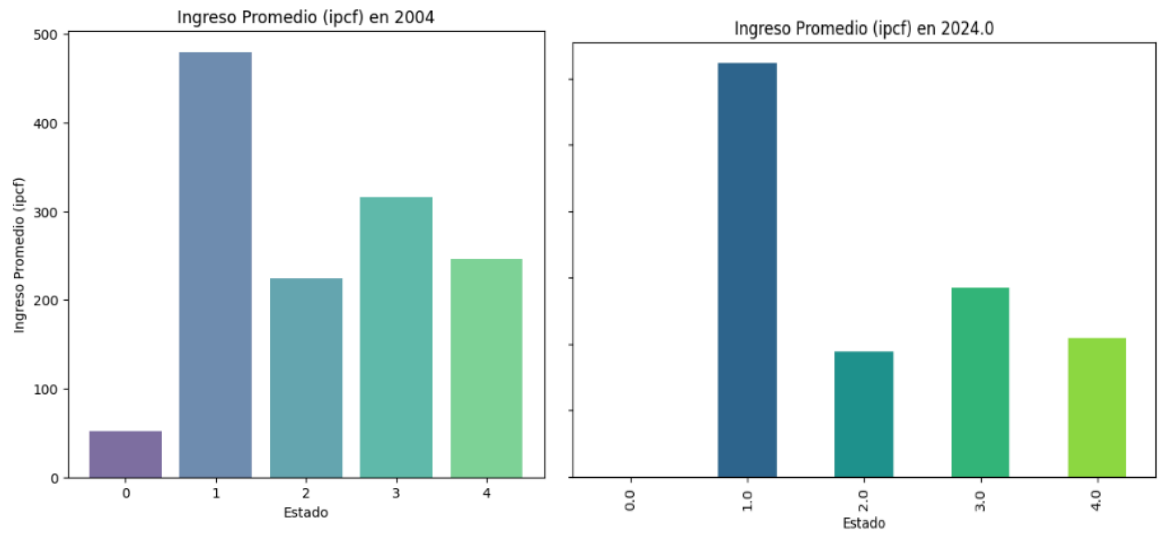


Fig. 3: Gráfico de barras con la media de ingreso per cápita familiar (IPCF) según estado (ocupado, desocupado, inactivo) en cada año 2004 y 2024.

La cantidad de personas que no respondieron sobre su condición de actividad es de 50, lo que representa una proporción del 0.38% respecto al total. En contraste, el 99.62% de las personas en la muestra, es decir, 13,065 individuos, sí respondieron esta pregunta, lo que asegura una amplia representatividad de los datos.

Se añadió una nueva variable llamada "pea" que identifica a la población económicamente activa (PEA). Esta variable agrupa a quienes antes estaban clasificados como ocupados (estado = 1) y desocupados (estado = 2), asignándoles el valor 1 en la variable "pea", mientras que el valor 0 representa a la población no activa.

Como se observa en el gráfico (Fig. 4), en 2004 la población no activa era mayor que la activa, aunque con una diferencia menor en comparación con 2024, donde la brecha se amplía considerablemente a favor de la población no activa. Además, entre las poblaciones activas, se registra una reducción de más de 1,000 personas (aproximadamente un 10% del total, calculado como 1,000/14,000), lo que podría ser un dato relevante para orientar futuras intervenciones.

Para estudios futuros, sería conveniente emplear un test estadístico, como el *t-test*, para evaluar si estas diferencias son significativas, porque ayudaría a determinar si la proporción de la población activa frente a la no activa sigue una tendencia de estabilidad, deterioro o mejora.

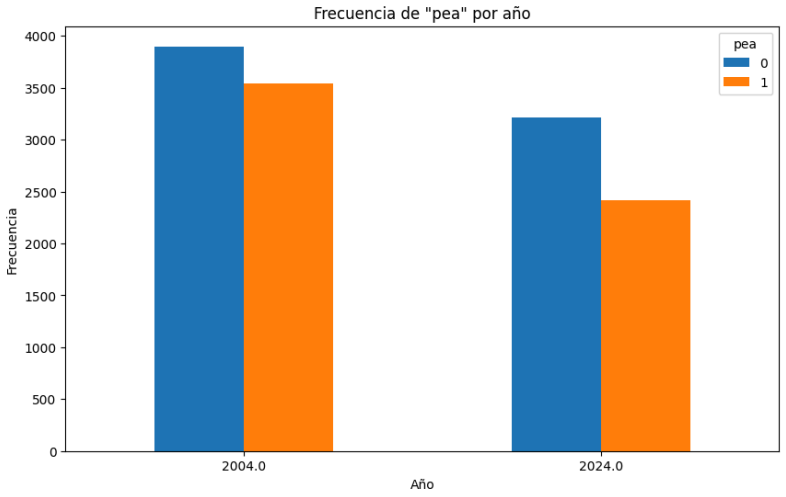


Fig. 4: Gráfico de barras de la composición por PEA para 2004 y 2024.

Se añadió una nueva columna a la base de datos llamada "PET" (Población en Edad para Trabajar), que toma el valor 1 si la persona tiene entre 15 y 65 años cumplidos. El gráfico correspondiente (Fig. 5), muestra que existe una diferencia superior al 20% en la proporción de personas que en 2004 estaban en edad para trabajar (PET) y también se identificaban como parte de la PEA, en comparación con la encuesta realizada en 2024, en la que este grupo disminuyó. Esto podría indicar un índice de envejecimiento en la población, y puede evidenciar que el recambio de mano de obra al mercado laboral no se produce de forma proporcional.

Además, es importante considerar que algunas personas pueden sentirse reticentes a declarar que están trabajando, debido a temores relacionados con la AFIP y las obligaciones tributarias, especialmente en contextos de trabajo informal. Para profundizar en esta situación, sería útil correlacionar estos datos con la cantidad de empleos no registrados (trabajo en negro) para evaluar el impacto real de este fenómeno en la participación laboral.

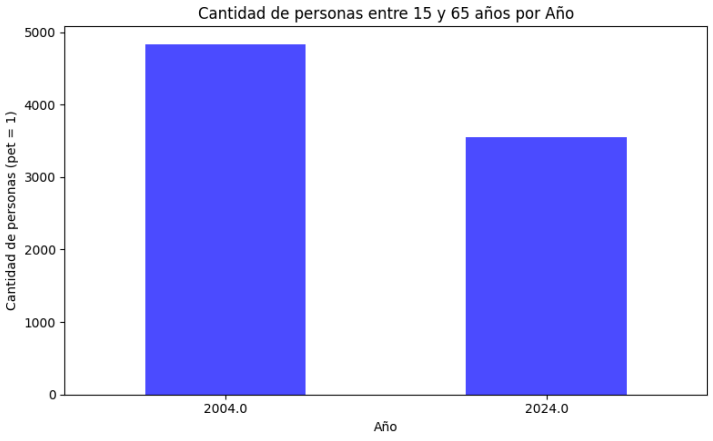


Fig. 5: Gráfico de barras de la composición por PET para 2004 y 2024.

Para evaluar la cantidad de personas desocupadas en cada año, se creó una columna llamada “desocupado” que toma el valor de 1 para aquellas personas cuya respuesta en la encuesta indicaba un estado de ocupación “desocupado” ('estado' = 2). A continuación, como se muestra en el gráfico (Fig. 6), se calculó el total de personas identificadas como desocupadas y se realizó una subdivisión por año. Los resultados mostraron que la cantidad de personas desocupadas en el año 2004 fue de 528, mientras que en el año 2024 fue de 311.

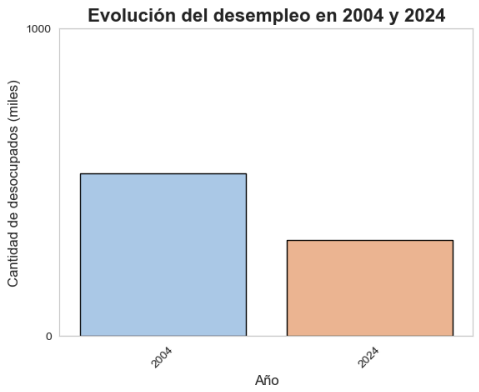


Fig. 6: Gráfico de barras de la proporción de desocupados comparando 2004 vs 2024.

El gráfico (Fig. 7) muestra la proporción de personas desocupadas según su nivel educativo en los años 2004 y 2024. En 2004, la desocupación era significativamente mayor entre aquellos que no habían completado la educación primaria (*nivel_ed*= 1.0) en comparación con 2024. Lo mismo ocurre con quienes reportaron haber completado la educación primaria (*nivel_ed*= 2.0), por lo que la proporción de desocupados en 2004 es considerablemente mayor que en 2024. Estas diferencias podrían ser investigadas en futuros análisis, ya que es importante comprender por qué la proporción de desocupación entre quienes no completaron el nivel primario es menor que la de aquellos con el nivel primario completo.

En las categorías de secundaria incompleta y completa (*nivel_ed*= 3.0 y *nivel_ed*= 4.0, respectivamente), se observan diferencias significativas en la tasa de ocupación entre aquellos que finalizaron la secundaria y quienes no lo hicieron, y este patrón se repite en ambos años. También es de suma relevancia evaluar por qué en ambos años, la tasa de desocupación fue mayor en él 2024. Esto podría correlacionarse con algún índice de pobreza. De igual manera, el mismo patrón se observa en los niveles de universidad completa e incompleta (*nivel_ed*=5.0 y *nivel_ed*=6.0, respectivamente), donde existen diferencias significativas en la tasa de ocupación entre quienes terminaron sus estudios universitarios y quienes no. Este comportamiento se mantiene en ambos años. Asimismo, resulta crucial analizar las razones por las cuales la tasa de desocupación fue más alta en 2024 en comparación con 2004. Por último, las proporciones de los que no tuvieron instrucción (*nivel_ed=7.0)* fueron similares en ambos años.

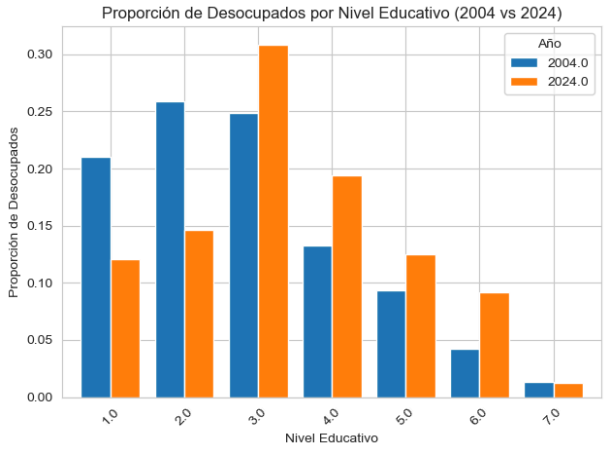


Fig. 7: Gráfico de barras de la proporción de desocupados por nivel educativo comparando 2004 vs 2024.

El gráfico (Fig. 8) muestra la proporción de personas desocupadas según grupo etario.

Es razonable que entre los 10 y los 20 años la proporción de desocupados sea significativamente mayor al promedio, ya que muchos aún no han ingresado plenamente al mercado laboral. En el grupo de 20 a 30 años, la tasa de desocupación aumenta, alcanzando casi el 13%, esto se puede deber a que algunos jóvenes siguen siendo estudiantes. Por otro lado, entre los 30 y los 60 años, se observa el pico de actividad laboral, con la menor proporción de desocupados.

Después de los 60 años, la tasa vuelve a subir y alcanza otro máximo local entre los 70 y 80 años, para luego caer a menos del 5% de desocupación en los años posteriores. Es de suma relevancia destacar que no se observan diferencias significativas entre los años 2004 y 2024, lo que indica que este patrón de desocupación por edad se mantiene estable entre ambos periodos.

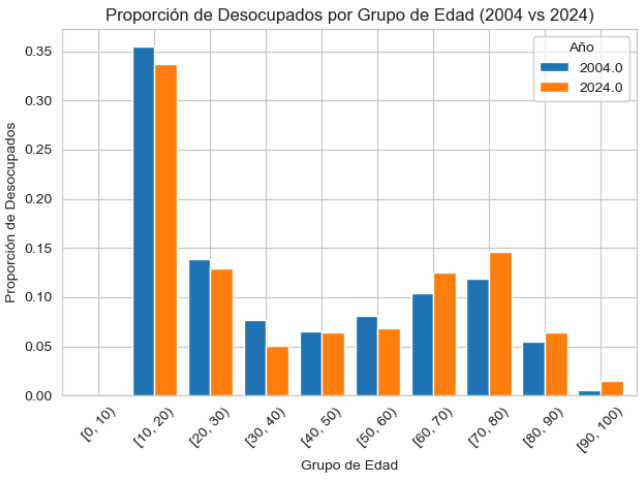


Fig. 8: Gráfico de barras de la proporción de desocupados por edad comparando 2004 vs 2024.

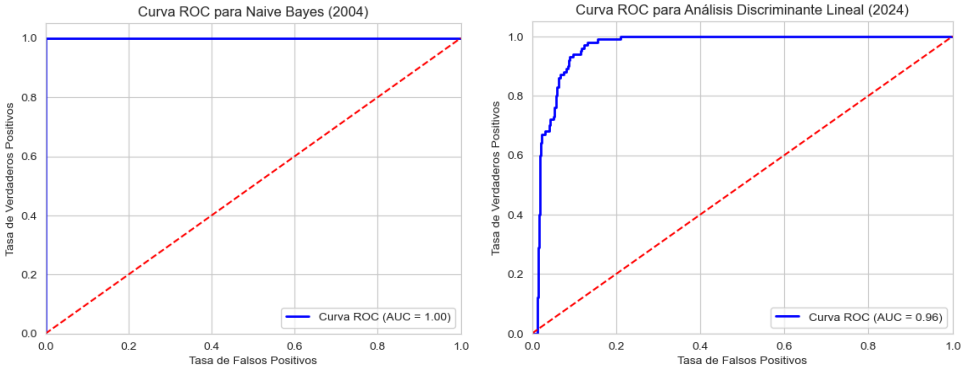
Parte II: Clasificación

En esta parte del trabajo, buscamos armar un modelo de predicción para estimar si una persona está o no desocupada a partir de ciertas características individuales. Para esto, para cada año, partimos la base de datos en una base *test* y una base *train* con el código X_train_2004, X_test_2004, y_train_2004, y_test_2004 = train_test_split (X_2004, y_2004, test_size= 0.3, random_state= 101)

(En este caso, para el año 2004). El test_size= 0.3 determina que vamos a utilizar el 70% de los datos para entrenar el modelo, y lo vamos a testear en el otro 30%. En la base de entrenamiento establecimos la variable [‘desocupado’] como variable dependiente, y el resto de las variables son consideradas como variables independientes.

Las variables que utilizamos como variables independientes son ['ch04', 'ch06', 'ch07', 'ch08', 'nivel_ed', 'estado', 'cat_inac', 'ipcf']. Consideramos que estas son las variables determinantes para predecir si una persona está o no desocupada.

Posteriormente, aplicamos cuatro métodos de análisis: regresión logística, análisis discriminante lineal, KNN con k=3 y Naive Bayes. Se reportaron los valores de la curva ROC, la matriz de confusión, los valores de Auc y Accuracy de cada uno (Fig. 9).



2004	Exactitud	Matriz de confusión	Auc
Regresión logística	0.983	[[2087 10] [27 128]]	0.995
Análisis discriminante lineal	0.927	[[2084 13] [150 5]]	0.975
KNN con k=3	0.923	[[2047 50] [122 33]]	0.689
Naive Bayes	0.996	[[2097 0] [7 148]]	1.0

2024	Exactitud	Matriz de confusión	Auc
Regresión logística	0.952	[[1988 0] [100 0]]	0.879
Análisis discriminante lineal	0.945	[[1975 13] [100 0]]	0.963
KNN con k=3	0.952	[[1975 13] [86 14]]	0.660
Naive Bayes	0.952	[[1988 0] [100 0]]	0.732

Fig. 9: Curva ROC, matriz de confusión, valores de Auc y Accuracy en los años 2004 y 2024 respectivamente.

Por un lado, se encontró que para 2004 el método que mejor predijo fue Naive Bayes. Predijo con una exactitud de 0.996 y un Auc de 1.0. Se puede considerar que es un modelo casi perfecto. Respecto a la matriz de confusión, esta muestra solo 7 falsos negativos y 0 falsos positivos.

Por otro lado, para 2024 el modelo que mejor predijo fue el análisis discriminante lineal, con una exactitud de 0.945 y un Auc de 0.963. Sin embargo, la matriz de confusión muestra 13 falsos positivos, pero tiene un número notable de falsos negativos (100). Esto implica que el modelo falla en la predicción de clases

Por último, se utilizó el método que seleccionamos para cada año, y se encontró que los modelos predicen que todas las personas que no contestaron están desocupadas.