Parte 1 - Analizando la base

1. En la página del INDEC, las personas desocupadas se identifican a través de la Encuesta Permanente de Hogares (EPH). Según la EPH, una persona es considerada desocupada si no tiene trabajo, está buscando activamente trabajo en las últimas cuatro semanas y se encuentra disponible para trabajar.
2. **Limpieza de la base de datos**

Para realizar la limpieza de la base de datos, en primer lugar, se filtraron las bases correspondientes al primer trimestre de 2004 y 2024 para incluir únicamente los aglomerados de interés: Ciudad Autónoma de Buenos Aires y Gran Buenos Aires. Una vez filtradas, ambas bases fueron concatenadas en un único conjunto de datos. Esto facilitó el acceso y la manipulación de la información de ambos períodos en una estructura unificada.

Durante la concatenación, se detectan diferencias en los nombres de columnas y en las etiquetas de algunas variables categóricas entre ambas bases. Para asegurar la consistencia, se realizaron las siguientes acciones:

- **Estandarización de nombres de columnas** : Se unificaron los nombres de las columnas en minúsculas para ambas bases.
- **Mapeo de categorías en variables de interés** : Se procedió a un mapeo unificado de las categorías en las variables de interés (CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC e IPCF). Esto implicó asignar códigos comunes a valores equivalentes entre ambas bases.

También, se realizó una revisión para detectar y corregir valores no válidos en las variables clave de edad e ingresos:

- **Edad (CH06)** : Se revisó la columna de edad para identificar valores negativos, los cuales no serán coherentes en el contexto de esta variable. Durante este proceso, se encontraron 51 observaciones con edades negativas en CH06, las cuales fueron eliminadas.
- **Ingreso (IPCF)** : Se verificó la columna de ingresos para asegurar la ausencia de valores negativos. Esta revisión permitió confirmar que todos los valores en IPCF cumplen con los criterios de validez esperados.

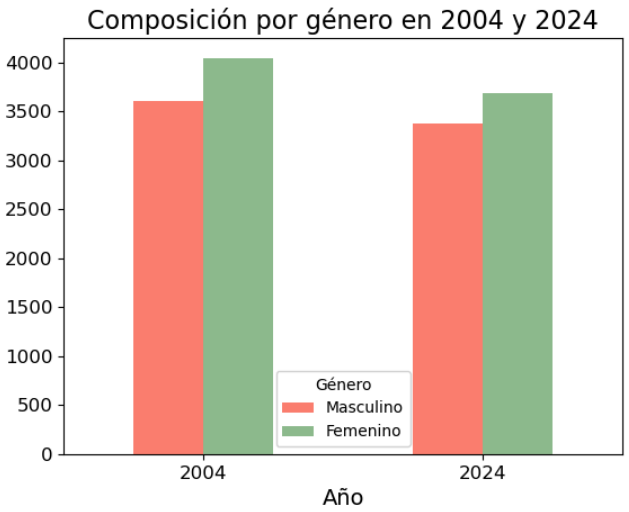

Figura 1. Composición por género en 2004 y 2024

Tras realizar la limpieza de los datos, generamos un histograma para visualizar la composición de la muestra según el sexo en los años 2004 y 2024 (ver Figura 1). En este histograma se observa una menor representación de personas del sexo masculino en comparación con el sexo femenino en ambos años.

Específicamente, en el año 2004, la distribución es la siguiente: 3,602 personas del sexo masculino y 4,045 personas del sexo femenino. En el año 2024, la composición es similar, con 3,371 personas del sexo masculino y 3,680 personas del sexo femenino. Este análisis preliminar sugiere que, en ambos años, la cantidad de personas de sexo femenino supera a la de sexo masculino en la muestra considerada.

Luego, realizamos una matriz de correlación para 2004 y 2024 con las siguientes variables: CH04, CH06, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC, IPCF. (Ver Figuras 2 y 3).

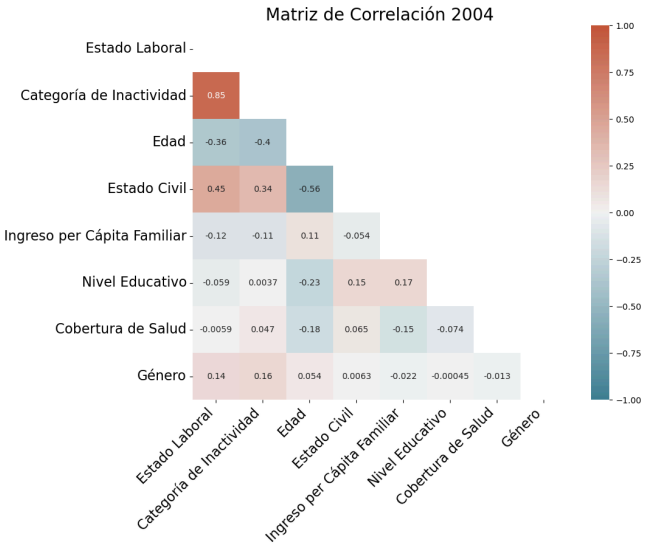

Figura 2. Matriz de correlación de 2004

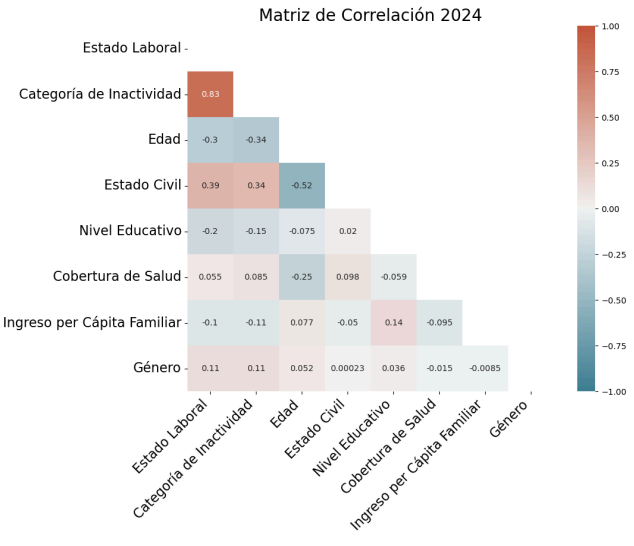

Figura 3. Matriz de correlación de 2024

Las matrices indican que las correlaciones entre variables se mantienen parecidas en ambos años, presentando mayor correlación entre variables como categoría de inactividad y estado laboral, y edad y estado laboral y categoría de inactividad. La matriz muestra las variables más correlacionadas en la parte superior izquierda, y divide por colores (rojo y azul) si la correlación es positiva o negativa.

Por un lado, en ambas matrices vemos una alta correlación positiva (corr = 0.85 en 2004 y corr = 0.83 en 2024) entre categoría de inactividad y estado laboral. La segunda correlación más significativa en ambos años es el estado civil y la edad, en la que se presenta una correlación negativa (corr = -0.56 en 2004 y corr = -0.52 en 2024). En el caso de la edad y el estado laboral se ve una correlación negativa moderada (corr = -0.36 en 2004 y corr = -0.3 en 2024) y en el caso de la edad y la categoría de inactividad se presenta una correlación negativa moderada similar a la anterior (corr = -0.4 en 2004 y corr = -0.34 en 2024). Variables como el estado civil y el género, el ingreso per cápita familiar y el nivel educativo presentan una correlación casi nula (cercana a 0).

Por otra parte, en la muestra hay en total 839 desocupados y 5462 inactivos. La media de IPCF para ocupados es de 106.443,40, de desocupados de 31.655,95 y de inactivos 63.863, 08.

3. En la base de datos, observamos que 51 personas no respondieron a la pregunta sobre su condición de actividad, en contraste con las 14,467 personas que sí proporcionaron esta información.

4. La composición por PEA (Población Económicamente Activa) para 2004 y 2024 es similar para ambos años, aunque para 2024 aumentó un 3.2%. Específicamente, en 2004 la cantidad de PEA es de 3607 y en 2024 de 3535 (Ver Figura 4).

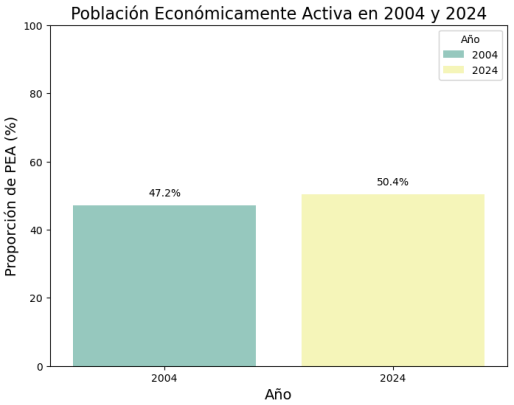

Figura 4. Composición por PEA para 2004 y 2024

5. La composición por PET para 2004 y 2024 también es similar entre ambos períodos y aumentó un 2,6% de 2004 a 2024 (Ver Figura 5).

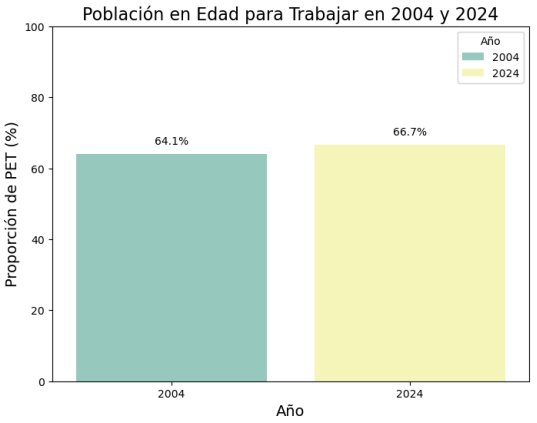

Figura 5. Composición por PET para 2004 y 2024

Al comparar PEA vs. PET podemos interpretar que la composición por PEA es menor respecto a la composición por PET y, además, se mantiene tanto en PET como PEA un aumento desde 2004 a 2024 (Ver Figura 6).

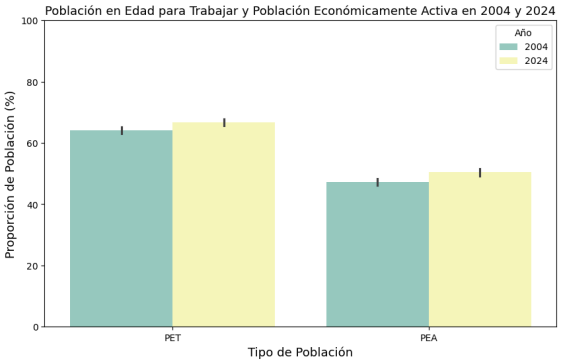

Figura 6. Comparación PET y PEA para 2004 y 2024

6. En 2004 hay en total 528 personas desocupadas y en 2024 un total de 311, por lo tanto se puede interpretar que disminuyó la cantidad de desocupados de 2004 a 2024 (Ver Figura 7).

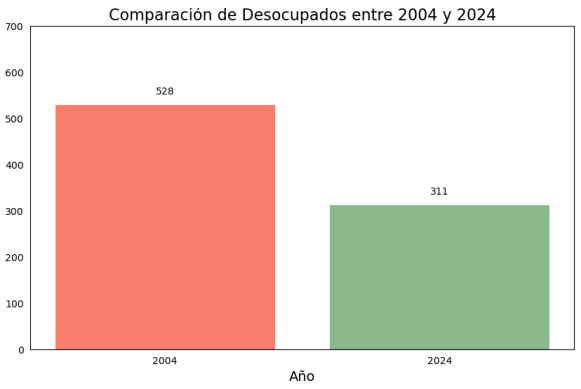

Figura 7. Comparación de desocupados entre 2004 y 2024

a. La proporción más elevada de desocupados se encuentra en el nivel de secundaria completa y educación superior universitaria incompleta, siendo mayor en el año 2004 en comparación con 2024. Mientras que en el nivel “sin instrucción” y primaria incompleta es menor la cantidad de personas desocupadas (Ver Figura 8).

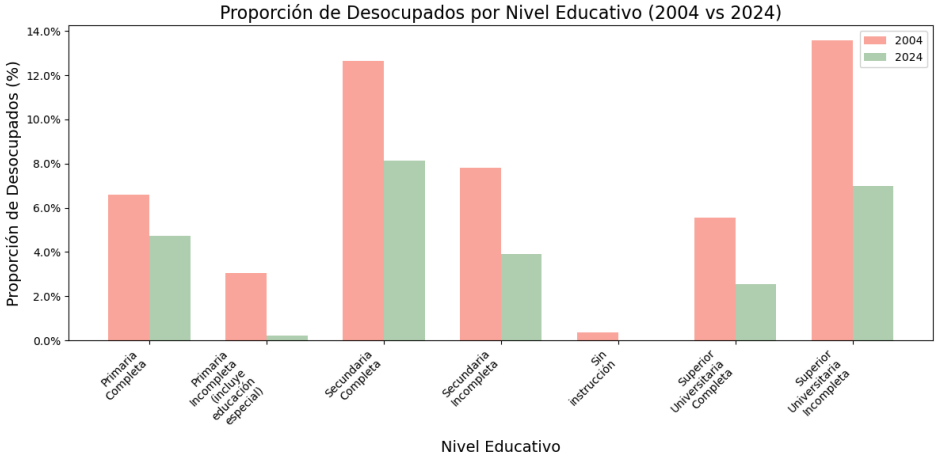

Figura 8. Proporción de Desocupados por Nivel Educativo (2004 vs 2024)

b. La proporción de desocupados más elevada se encuentra entre los 20 y 29 años, y no presenta grandes diferencias en la comparación entre 2004 y 2024. En el caso del grupo etario agrupado de 40-49 años, en cambio, se vio el mayor cambio en desocupados, ya que en 2004 la proporción era de aproximadamente 10% y en 2024 se encuentra en un 20% aproximadamente, por lo que se puede observar un aumento en la proporción de aproximadamente un 100%. En el caso del grupo etario de 10-19 años, se puede observar una disminución de aproximadamente 30% de los desocupados, ya que en 2004 estos se encontraban en más de un 17% y en 2024 esta proporción disminuyó a aproximadamente 12%.

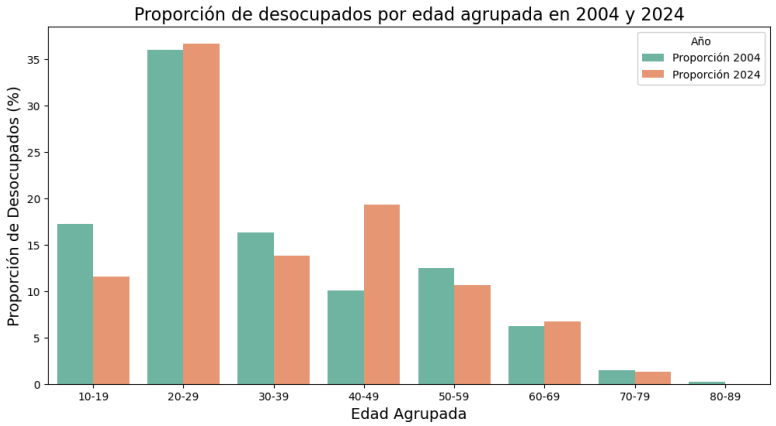

Figura 9. Proporción de Desocupados por edad agrupada (2004 vs 2024)

Parte 2: Clasificación

En esta sección del informe, el objetivo es desarrollar un modelo predictivo para determinar la probabilidad de que una persona esté desocupada, utilizando diversas

variables de características individuales. Para ello, se implementarán diferentes métodos de clasificación (regresión logística, análisis discriminante lineal, KNN con k = 3 y naive Bayes) y se evaluarán sus desempeños a través de diversas métricas: la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy.

Modelo de Regresión Logística

Resultados 2004

- ROC AUC = 0.671
- Accuracy = 0.942

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	4138	0
Real: Ocupado	257	0

Figura 10. Matriz de Confusión Regresión Logística 2004

Resultados 2024

- ROC AUC = 0.651
- Accuracy = 0.960

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	2019	0
Real: Ocupado	84	0

Figura 11. Matriz de Confusión de Regresión Logística 2024

Modelo de Análisis Discriminante Lineal

Resultados 2004

- ROC AUC = 0.969
- Accuracy = 0.936

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	4110	28
Real: Ocupado	255	2

Figura 12. Matriz de Confusión de Análisis Discriminante Lineal 2004

Resultados 2024

- ROC AUC = 0.972
- Accuracy = 0.958

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	2013	6
Real: Ocupado	82	2

Figura 13. Matriz de Confusión de Análisis Discriminante Lineal 2024

KNN con k = 3

Resultados 2004

- ROC AUC = 0.819
- Accuracy = 0.935

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	4073	65
Real: Ocupado	222	35

Figura 14. Matriz de Confusión de KNN = 3 para 2004

Resultados 2024

- ROC AUC = 0.592
- Accuracy = 0.953

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	1997	22
Real: Ocupado	76	8

Figura 15. Matriz de Confusión de KNN = 3 para 2024

Naive Bayes

Resultados 2004

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	4139	0
Real: Ocupado	257	0

Figura 16. Matriz de Correlación de Naive Bayes 2004

- ROC AUC = 0.708
- Accuracy = 0.942

Resultados 2024

- ROC AUC = 0.703
- Accuracy = 0.960

	Predicción: Desocupado	Predicción: Ocupado
Real: Desocupado	2019	0
Real: Ocupado	84	0

Figura 17. Matriz de Confusión de Naive Bayes 2024

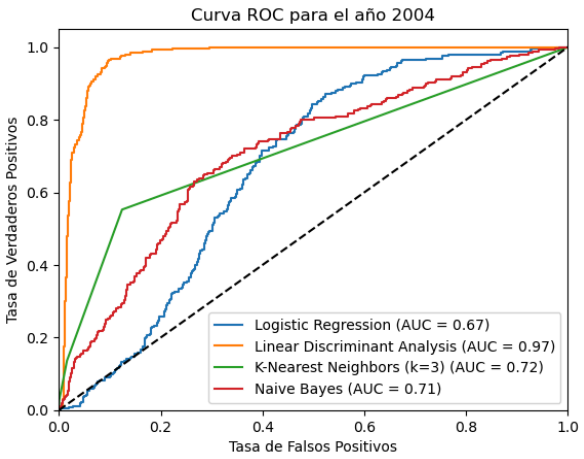

Figura 18. Curva ROC de los 4 modelos en 2004

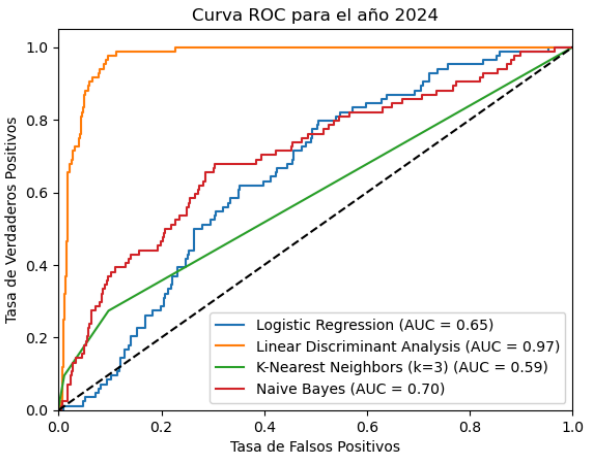

Figura 19. Curva ROC de los 4 modelos en 2024

3. Compare los resultados de 2004 versus 2024. ¿Cuál de los métodos predice mejor en cada año? Justifiquen detalladamente utilizando las medidas de precisión que conocen.

De acuerdo al análisis de comparación de métricas de los diferentes métodos, encontramos que el Análisis Discriminante lineal es el que mejor predice en 2004 y 2024. Si bien el nivel de Accuracy es similar respecto a los demás modelos, el AUC es significativamente alto comparado con los demás. Además, en cuanto a las matrices de confusión, en 2004, el LDA logra identificar correctamente 2 verdaderos positivos en la clase positiva, mientras que otros modelos (como la Regresión Logística y Naive Bayes) no logran predecir ningún positivo. En 2024, el LDA también tiene detecciones en ambas clases, lo cual mejora su capacidad de discriminación y lo convierte en un modelo más equilibrado. Aunque la precisión es similar entre LDA y otros modelos, el hecho de que el LDA sea capaz de identificar algunos positivos ayuda a que su AUC sea significativamente más alto (0.969 en 2004 y 0.972 en 2024), lo cual refleja un mejor equilibrio entre sensibilidad

(verdaderos positivos) y especificidad (falsos positivos). Además, en los dos años (2004 y 2024), el LDA mantiene un desempeño superior, lo cual muestra su robustez y consistencia en ambos conjuntos de datos.

En cuanto a la interpretación de las curvas ROC (Figura 18 y Figura 19), la curva ROC de LDA está más cerca de la esquina superior izquierda del gráfico en comparación con los demás modelos. Esto significa que el modelo está logrando una alta tasa de verdaderos positivos (sensibilidad) y una baja tasa de falsos positivos (especificidad alta). Por lo tanto, esto indica que LDA tiene una excelente capacidad para clasificar correctamente las observaciones en ambas clases.

4. Con el método que seleccionaron, predigan qué personas son desocupadas dentro de la base norespondieron. ¿Qué proporción de las personas que no respondieron pudieron identificar como desocupadas?

Utilizando el modelo seleccionado, se procedió a predecir la condición de actividad de las personas que no respondieron a la encuesta. En el análisis de la base de datos correspondiente a los 51 individuos que no brindaron información sobre su situación laboral, se logró identificar que 9 de ellos son clasificados como desocupados. Esto representa un 17.65% del total de las personas que no respondieron.