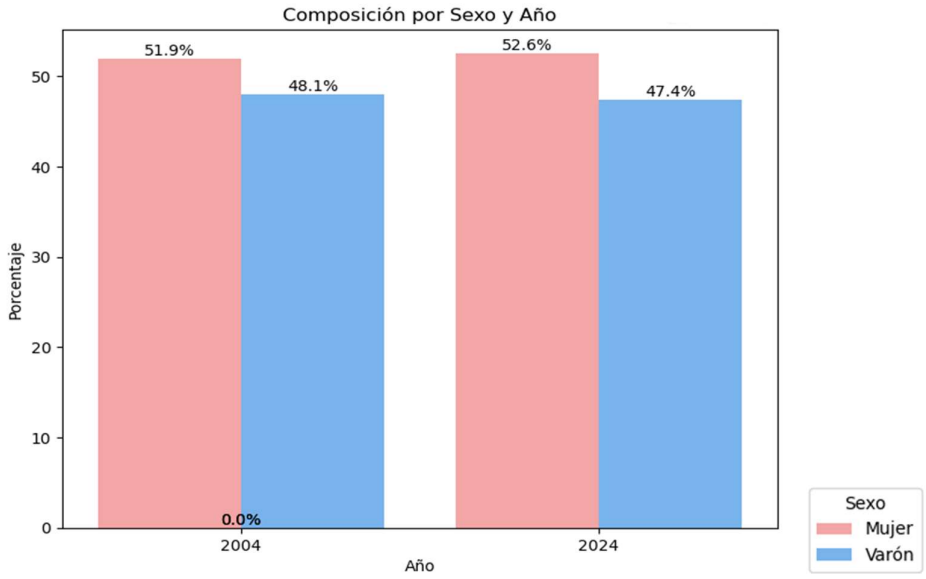
Análisis de la composición de la EPH

La base de datos para la provincia de Tucumán no contaba con observaciones de ingresos menores a cero. Respecto a las edades, sí observamos que la variable contenía valores numéricos enteros para todo el rango de observaciones excepto para los bebés menores a 1 año. En cuanto al tratamiento de estas observaciones optamos por dejarlas fuera del análisis, ya que, por un lado representaban un porcentaje bajo en la muestra y por otro, debido a que el foco lo poníamos en personas en edad de trabajar, eliminarlos no constituía un problema a los fines de este trabajo.

La EPH del primer trimestre para el año 2004 estuvo compuesta por 1.224 varones y 1.323 mujeres para el Gran Tucumán y Tafí Viejo, sumando un total de 2.547 observaciones, mientras que la EPH para el Gran Tucumán y Tafí Viejo del primer trimestre del 2024 está conformada por 2.098 personas, de las cuales 1.103 son mujeres y 995 son varones. En la encuesta del primer trimestre de 2004 las mujeres representaron aproximadamente el 52% de la muestra mientras que los varones representaron el 48% restante. En la encuesta del primer trimestre del año actual, ambos sexos tuvieron una representación porcentual idéntica a la calculada para el período de comparación (primer trimestre 2004).

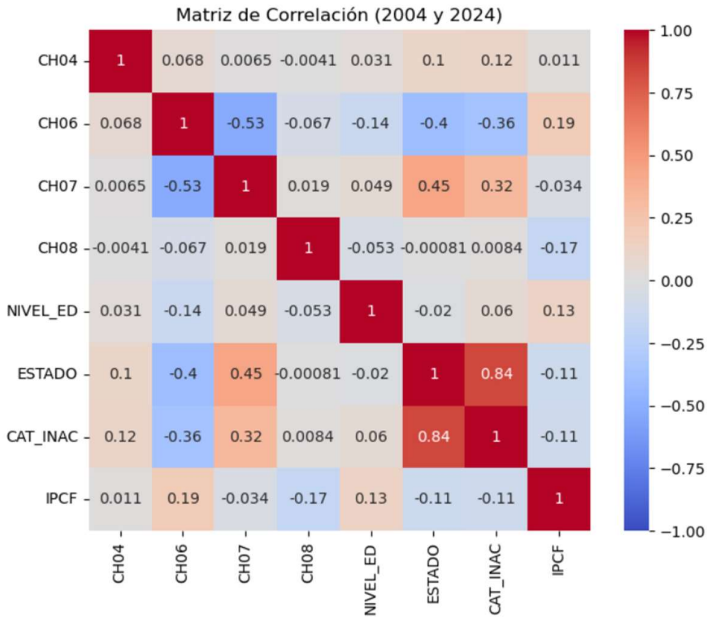
Gráfico 1: composición por sexo y año de la EPH para los primeros trimestres de los años 2004 y 2024.



Fuente: elaboración propia a partir de los datos de las EPH (Indec)

En este trabajo realizamos una matriz de correlación utilizando datos para ambos años de la EPH (2004 y 2024) y las variables: años cumplidos (CH06), sexo (CH04), estado civil (CH07), cobertura médica (CH08), nivel educativo (NIVEL_ED), estado de actividad (ESTADO) y categoría de inactividad (CAT_INAC).

Gráfico 2: matriz de correlación entre las variables mencionadas previamente usando los datos relevados por la EPH para el primer trimestre de los años 2004 y 2024.



Fuente: elaboración propia a partir de los datos de las EPH (Indec)

Las variables que están más fuertemente correlacionadas son la categoría de inactividad con el estado de actividad, esto de hecho se explica por construcción de las variables ya que una de las categorías de Estado es Inactivo. Es decir, conocer la variable Estado nos da mucha información sobre la categoría de inactividad de la persona.

Otras variables que tienen una correlación negativa bastante elevada son edad (CH04) y estado civil (CH07). Entendemos que este coeficiente indica que cuanto menor es la persona, mayor es el número que toma la variable estado civil, ya que esta variable asigna el valor más alto de su rango (5) a las personas solteras.

Conforme a las definiciones del Indec, el grupo de los desocupados está compuesto por todas aquellas personas que sin tener trabajo se encuentren disponibles para trabajar y han buscado activamente una ocupación en un período de referencia determinado.

Utilizamos la variable ESTADO que toma el valor 2 cuando la persona está desocupada y 3 cuando está inactiva. Encontramos que para el Gran Tucumán y Tafí Viejo, para el primer trimestre de 2004 habían 178 personas desocupadas y 973 personas inactivas. Mientras que, para mismo aglomerado y trimestre, para el año 2024, el número de desocupados disminuyó a 81 (un 45% del valor que tomó en 2004) y el número de personas inactivas disminuyó a 869.

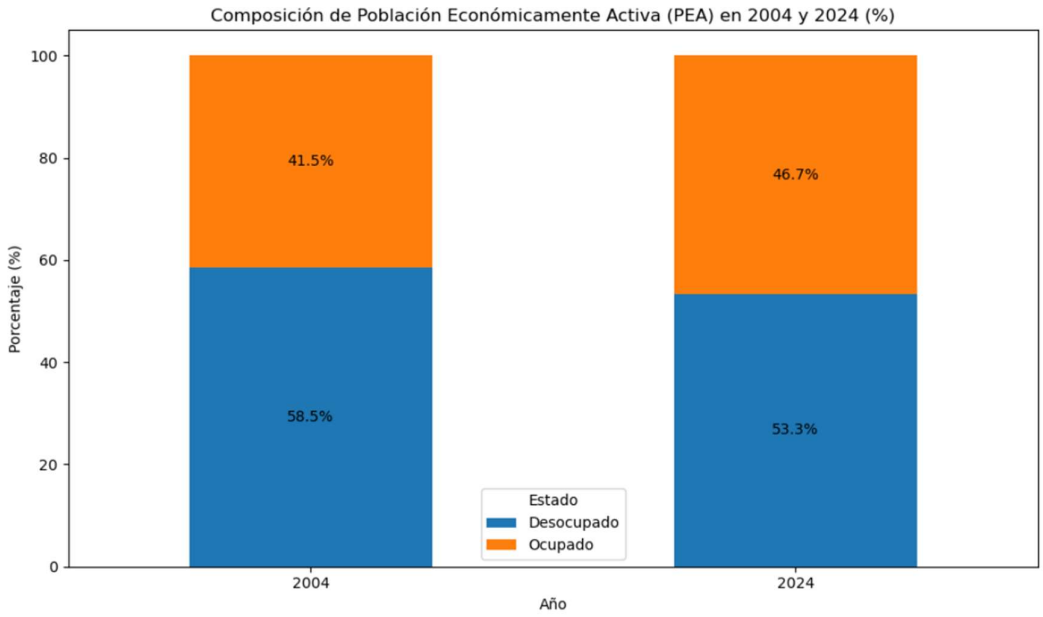
Cuadro 1: Ingreso medio en pesos por estado de actividad para las categorías desocupado, inactivo y ocupado. Datos EPH primer trimestre de 2004 y 2024.

Estado	Ingreso Medio
Desocupado	31834.561216
Inactivo	60272.543137
Ocupado	77187.089531

Fuente: elaboración propia a partir de los datos de las EPH (Indec)

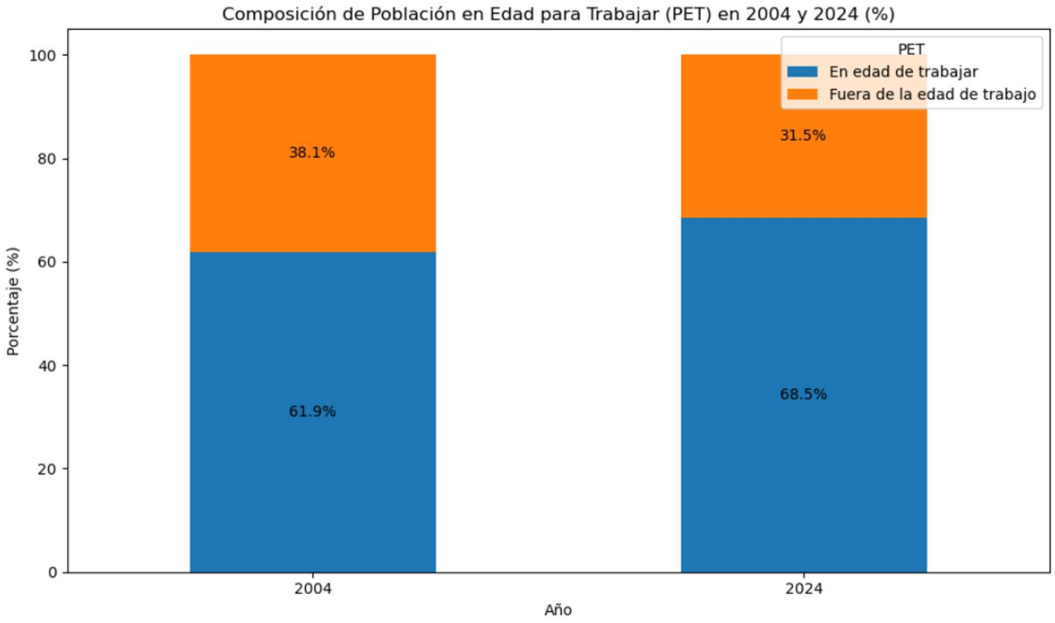
Para el trimestre y los años analizados encontramos 10 observaciones que no respondieron a la pregunta sobre su estado de actividad, 6 de ellas en el año 2004 y las restantes 4 en el año 2024. Generamos el siguiente gráfico sólo con las observaciones que declararon un ESTADO.

Gráfico 3: composición por PEA para los años 2004 y 2024 usando los datos de la EPH.



Fuente: elaboración propia a partir de los datos de las EPH (Indec)

Gráfico 4: composición por PET para los años 2004 y 2024 usando los datos de la EPH.



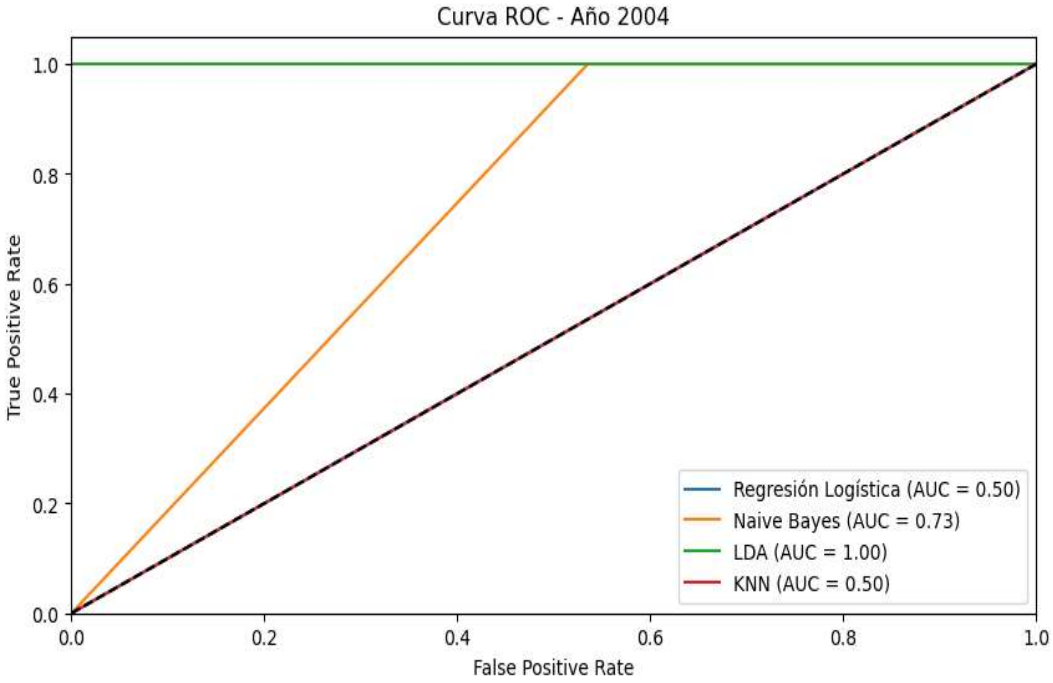
Fuente: elaboración propia a partir de los datos de las EPH (Indec)

Podemos ver que hay un porcentaje mayor en el grupo de personas en edad de trabajar (62% en el 2004 a 68,5% en el año 2024) que de personas económicamente activas en los años considerados (58,5% y 53% respectivamente). Esta brecha es menor para 2004 que para 2024.

PARTE II

Análisis curvas ROC

CURVA ROC 2004



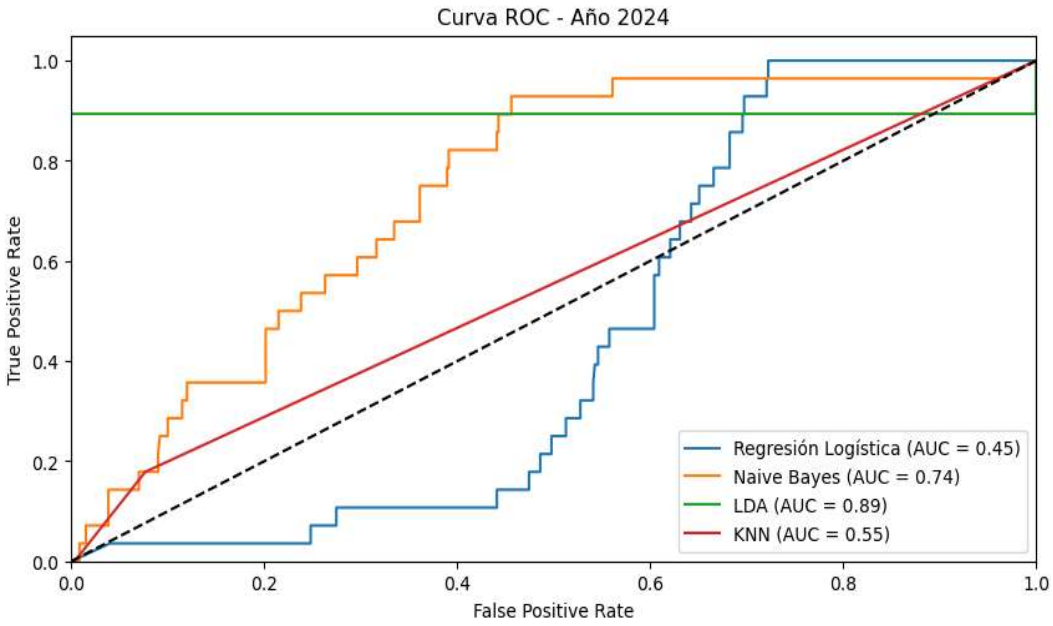
1. **LDA (Análisis Discriminante Lineal):**
    - **AUC = 1.00:** Este es el mejor resultado posible y sugiere que el modelo LDA clasifica perfectamente los datos. Debemos tener en cuenta que un AUC de 1.00 podría indicar sobreajuste.
2. **Naive Bayes:**
    - **AUC = 0.73:** Un valor de AUC de 0.73 indica un buen desempeño, aunque no tan bueno como el LDA. Este modelo parece ser capaz de distinguir entre las clases con una precisión razonable.
3. **Regresión Logística y KNN:**
    - **AUC = 0.50:** Un valor de AUC de 0.50 es equivalente al rendimiento de un clasificador aleatorio. Esto sugiere que estos dos modelos no están aportando ninguna información útil para la clasificación en este conjunto de datos.

Conclusiones

- **LDA es el mejor modelo:** Basado en el AUC, el modelo LDA es claramente el mejor de los cuatro modelos evaluados. Sin embargo, como mencionamos antes, es importante verificar si no hay sobreajuste.

- **Otros modelos no son adecuados:** Tanto la Regresión Logística como el KNN tienen un rendimiento muy pobre, lo que sugiere que estos modelos no son adecuados para este conjunto de datos en particular.

CURVA ROC 2024



Basándonos en la gráfica, podemos hacer las siguientes observaciones:

- **LDA es el mejor modelo:** La curva del LDA está más cerca de la esquina superior izquierda y tiene el AUC más alto (0.89), lo que indica que tiene un excelente equilibrio entre sensibilidad y especificidad.
- **Naive Bayes es el segundo mejor modelo:** También tiene un buen desempeño, con un AUC de 0.74.
- **KNN y Regresión Logística tienen un desempeño más bajo:** Sus curvas están más cerca de la línea diagonal y tienen valores de AUC menores, lo que sugiere que estos modelos no son tan efectivos para esta tarea de clasificación.