

Trabajo Práctico N°3

El desempleo (D) es un problema macroeconómico grave y representa a la proporción de población que no encuentra trabajo, pero está en la búsqueda y tiene disponibilidad. Forman parte de la población activa ($PA = E + D$), por lo tanto la tasa de paro se define como D/PA (Mankiw, 2008, p.275).

En Argentina se utilizan varios indicadores para obtener la composición de la tasa de desocupación. Algunos de ellos son de carácter demográfico, por ejemplo, el sexo, la edad, el rol que ocupa en su familia o el nivel educativo, mientras que otros se refieren a cuestiones más puntuales como el tiempo de búsqueda de empleo, la calificación y la rama de actividad de la última ocupación, la categoría ocupacional o el tipo de establecimiento del último empleo.

PARTE 1: Limpieza de la base y gráficos de exploración

Para comenzar con el filtrado de datos, se seleccionó la columna “región” de cada una de las bases (2004 y 2024) y los datos correspondientes al Gran Buenos Aires y CABA. Posteriormente se utilizó el comando `str.lower()` para pasar cada uno de los nombres de las columnas a imprenta minúscula, de modo que estas pudieran ser comparables y, para observar si había columnas en un dataframe que no estaban en el otro, se utilizó el mecanismo `if/else` que arrojó 9 columnas no coincidentes. Estas fueron eliminadas para evitar datos faltantes al momento de unir las bases pertenecientes a diferentes años.

Los dos dataframes, correspondientes al 2004 y 2024, fueron unidos de forma vertical utilizando la función `concat()`. Se reemplazaron los datos de algunas columnas pertenecientes a la primera base, que estaban en formato string y se les otorgó un código indicado en el documento [Diseño de registro y estructura para la base de datos](#). Posteriormente se utilizó un `for()` para recorrer las columnas numéricas y detectar los valores inferiores cero (negativos). Estos datos también fueron eliminados ya que la mayoría representaba algún tipo de ingreso, que se podría interpretar como un nivel de deuda, y otros carecían de interpretabilidad,

Una vez realizada la limpieza de datos, se realizó un gráfico de barras (Gráfico 1) comparativo entre ambos años y segmentado según el sexo. A simple vista se puede observar una mayor cantidad de datos en el año 2004. Asimismo, en ambos grupos la participación de las mujeres es mayor.

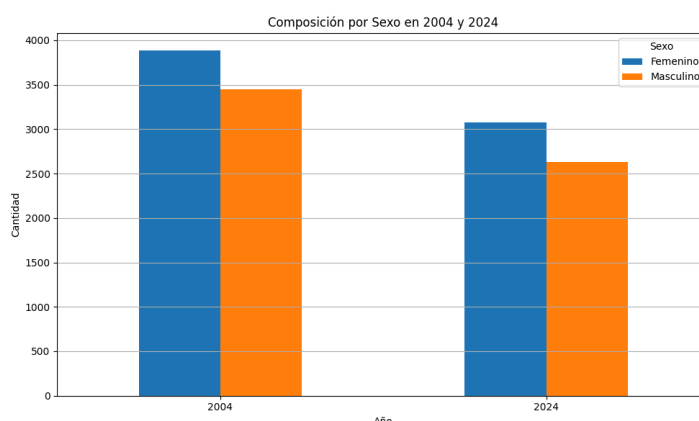


Gráfico 1: comparación de datos por sexo y año

En segundo lugar, se realizaron dos matrices de correlaciones (mapas de calor) para cada uno de los años analizados. Se consideraron las variables sexo, edad, estado civil, cobertura médica, nivel educativo, condición de actividad (estado), categoría de inactividad y monto de ingreso per cápita familia. Para permitir una mejor interpretabilidad de las correlaciones, se categorizaron subvariables *dummies* (por ejemplo, de la variable sexo se desprendieron las subvariables varón y mujer). Como resultado, se obtuvo una matriz de 27x27, que destacó algunos paralelismos interesantes.

Gráfico 2: matriz de correlación 2004

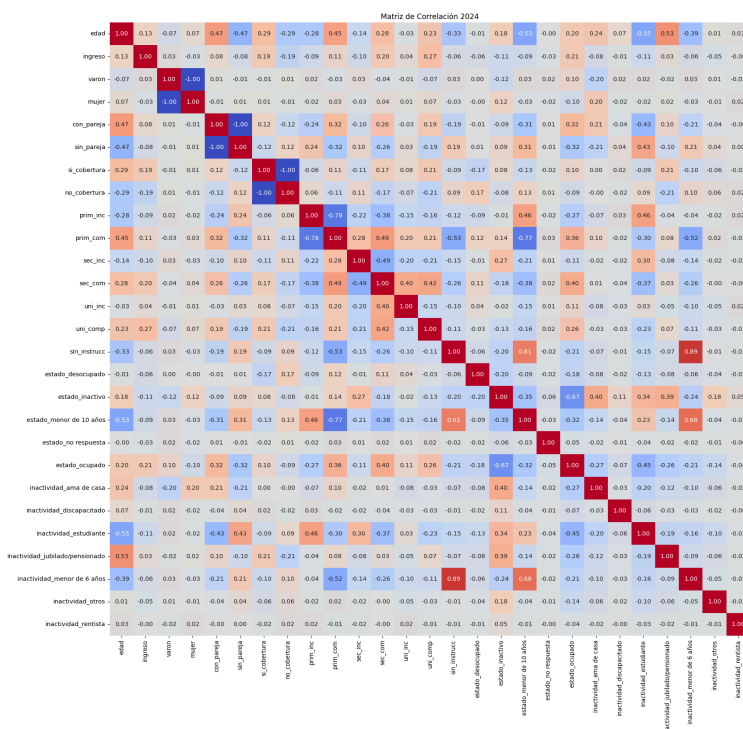
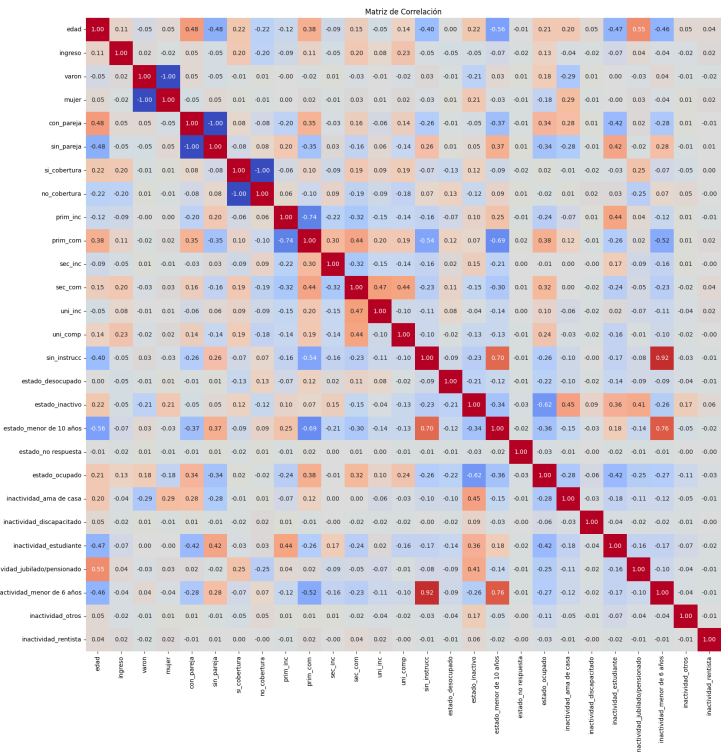


Gráfico 3: matriz de correlación 2024

Para ambos años, la correlación más alta fue entre persona sin instrucción (educación) y persona inactiva menor de 6 años o condición de actividad como menor a 10 años. También correlacionaron las variables persona sin instrucción y persona inactiva.

En relación con lo mencionado anteriormente, otra variable que mostró asociación fue la edad, tanto con el estado de inactividad (jubilado/menor de edad) como así con el nivel de instrucción (primario/ secundario/ terciario). El resto de las correlaciones se podrían clasificar de moderadas (.4 a .6) a bajas.

Posteriormente se creó un nuevo *dataframe* que incluía sólo la información de las personas que respondieron a su condición de actividad, es decir, si estaban empleadas, desempleadas, etc. y se añadió una columna binomial PEA (Población económicamente activa) que tomaba el valor uno si la persona clasificaba dentro de esa categoría. Se realizaron agrupaciones de PEA por año y se graficaron (Gráfico 4), obteniendo como resultado que en 2024 hubo un aumento de 4% aproximadamente de la población no económicamente activa en comparación con el año 2004. A partir de estos

datos, se realizó otro gráfico representativo de la proporción de población en edad de trabajo (PET) dentro de la PEA (Gráfico 5). Si bien la gran mayoría de la PEA se encuentra dentro de la PET, a simple vista hay bastante simetría en la comparación por años.

Para sintetizar estos

resultados, se llevó a cabo un gráfico de barras de PEA y PET en ambos años (Gráfico 6).

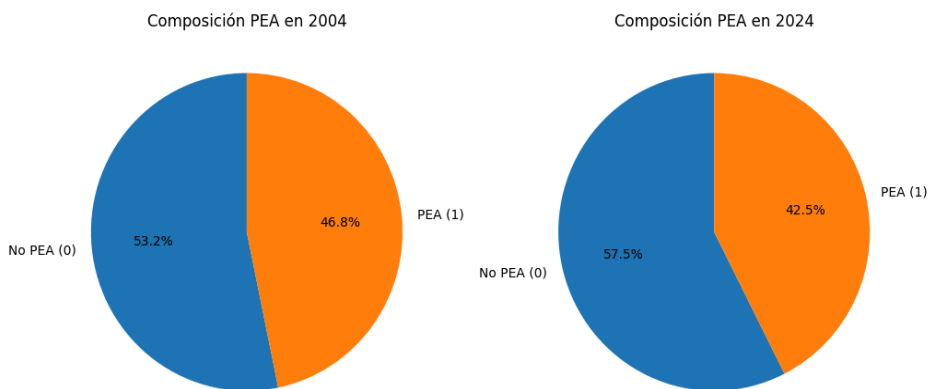


Gráfico 4: comparación entre PEA 2004 y 2024

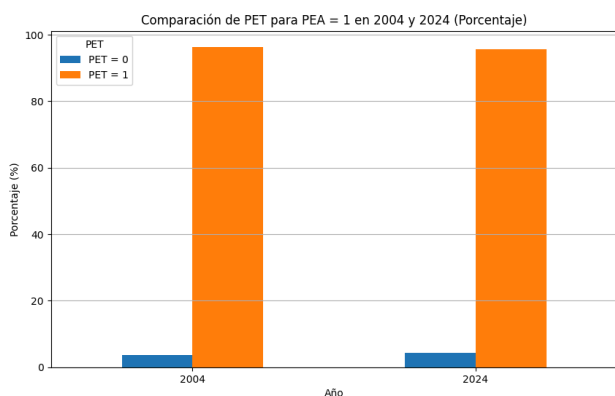


Gráfico 5: comparación de PET para PEA

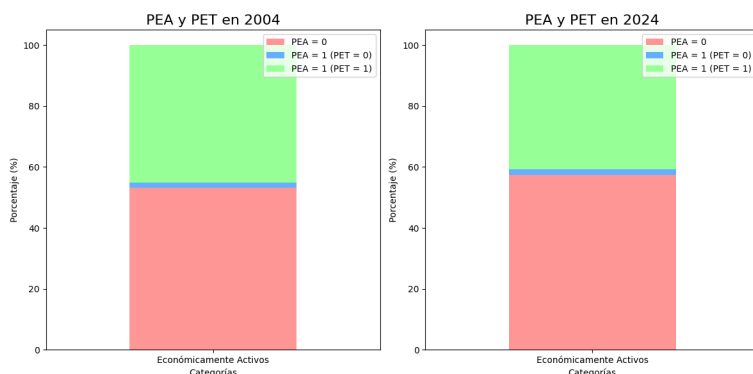


Gráfico 6: PEA y PET por año

Al *dataframe* “respondieron” se le agregó una columna binaria que tomó el valor 1 en aquellos casos en los que el estado de la persona fuera desocupado y 0 para el resto de las

opciones. De este modo se pudo obtener la proporción de desocupados según el nivel educativo alcanzado (Gráfico 7) y el rango etario (Gráfico 8). Los datos obtenidos demuestran que

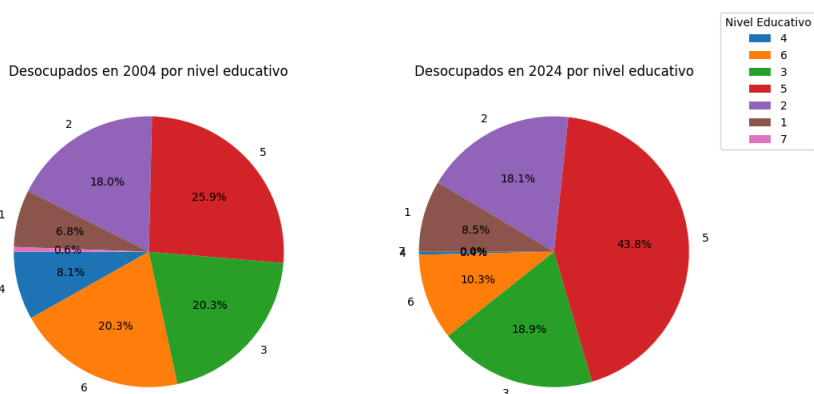


Gráfico 7: desocupados según nivel educativo en 2004 y 2024

en 2024 hay menor cantidad de desempleados (528 en 2004 vs. 281 en 2024), pero además se observa en el gráfico que aumentó la proporción de personas sin título universitario desempleadas (43%) en relación con el total. Esto puede deberse a cambios en la demanda del mercado laboral, ya que se exige más perfeccionamiento y al incremento de la competencia, que puede provocar que aquellos sin título queden excluidos. Sin embargo, en términos absolutos, los desocupados sin título universitario de 2024 (51) son menos que los de 2004 (95).

El gráfico 8 muestra que el rango etario con más desocupación es entre 20 y 29 años tanto en 2004 como así también en 2024. La tendencia en ambos años cambia levemente, evidenciando como en 2024 el segundo mayor rango etario con desempleo es de personas de entre 40-49 años, lo que se contrasta con la situación actual [del país](#).

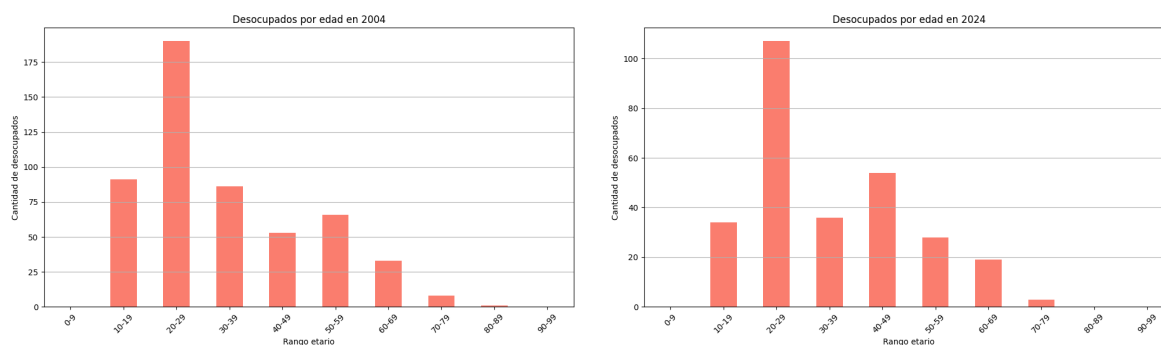


Gráfico 8: desocupados por rango etario en 2004 y 2024

PARTE 2: Predicciones

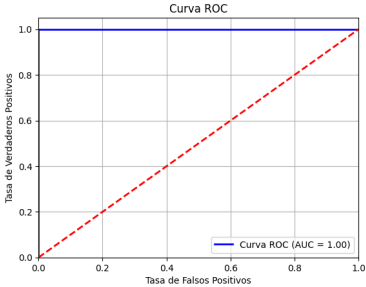
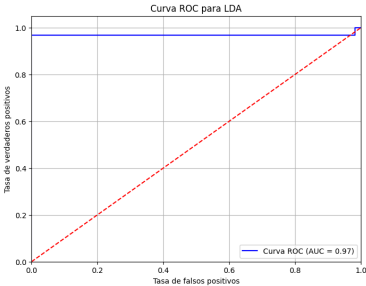
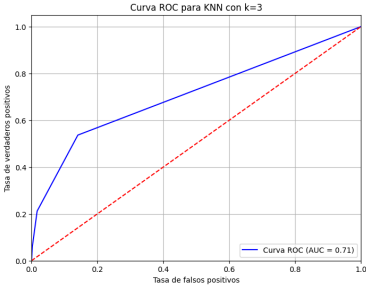
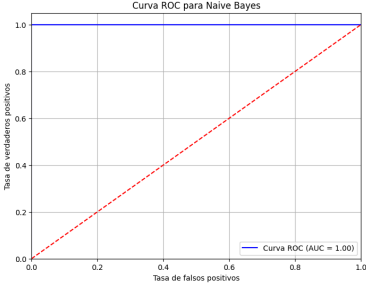
Luego del análisis de datos realizados previamente, se creó un modelo predictivo para prever si una persona está desocupada o no. Se utilizó el *dataframe* “respondieron” y las mismas variables empleadas para la construcción de las matrices (Gráfico 2 y 3), que fueron segmentadas en subvariables *dummies*. También se importaron los módulos *Numpy* y *scikit-learn* con algunas de sus funciones como *train_test_split*, *LogisticRegression*, entre otras. La muestra total fue dividida en datos de entrenamiento (70%) y datos de testeo (30%), seteando a la semilla igual a 101. Se estableció a *desocupado* como la variable dependiente (vector y) y al resto de las variables como independientes (matriz X). Se agregó la columna “*intercept*” para permitir que el modelo ajuste a esta, que es el valor de la variable dependiente cuando todas las variables independientes son cero.

Se implementaron cuatro modelos predictivos (regresión logística, LDA, KNN y Naive Bayes) y con cada uno de ellos se computó el área bajo la curva, la curva ROC, los valores de *accuracy* y la matriz de confusión. Esta representa: fila 1, columna 1= verdaderos negativos; fila 1, columna 2= falsos positivos; fila 2, columna 1= falsos negativos; fila 2, columna 2= verdaderos positivos. El verdadero negativo representa el número de casos en el que el modelo predijo correctamente que la persona no está desocupada (predicción 0) y efectivamente no lo está. El falso positivo representa los números de casos en los que el modelo predijo que la persona está desocupada pero en realidad no lo está. El falso negativo representa la cantidad de casos en los que el modelo predijo que la persona no está

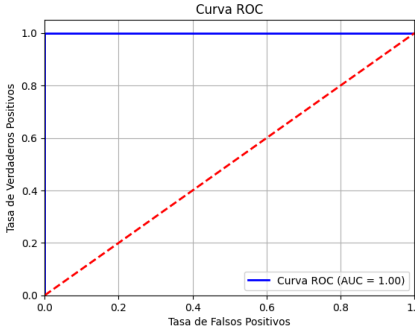
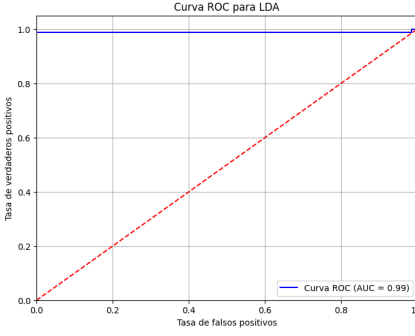
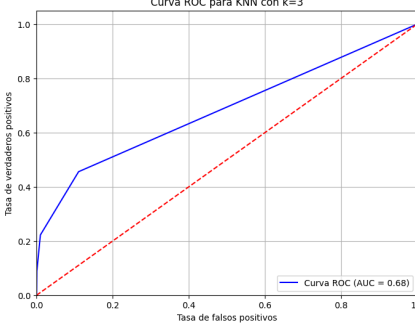
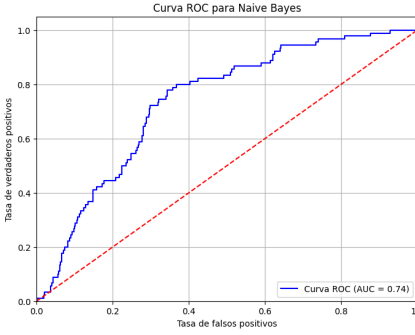
desocupada pero en realidad sí lo está. Por último, el verdadero positivo es la cantidad de casos en los que el modelo predijo que la persona no está desocupada y esto es así.

Además, en el caso del modelo LDA para 2024, se utilizó la técnica de *cross-validation* para probar el modelo en diferentes conjuntos de datos y generalizar los resultados. Todos los análisis se realizaron por separado para cada uno de los años y los resultados obtenidos fueron los siguientes:

Año 2004

Regresión logística		<ul style="list-style-type: none"> • Matriz de confusión: [[2120 0] [0 151]] • Precisión del modelo de regresión logística: 1.0 • Área bajo la curva (AUC): 1.0
Análisis discriminante lineal		<ul style="list-style-type: none"> • Matriz de confusión: [[2120 0] [6 145]] • Precisión del modelo LDA: 0.9973579920739762 • Área bajo la curva (AUC): 0.9675121829314007
KNN con K=3		<ul style="list-style-type: none"> • Matriz de confusión para KNN: [[2084 36] [119 32]] • Precisión del modelo KNN: 0.931748128577719 • Área bajo la curva (AUC) para KNN: 0.7085249281519429
Naive Bayes		<ul style="list-style-type: none"> • Matriz de confusión para Naive Bayes: [[2120 0] [0 151]] • Precisión del modelo Naive Bayes: 1.0 • Área bajo la curva (AUC) para Naive Bayes: 1.0

Año 2024

Regresión logística	 <p>Curva ROC</p> <p>Tasa de Verdaderos Positivos</p> <p>Tasa de Falsos Positivos</p> <p>Curva ROC (AUC = 1.00)</p>	<ul style="list-style-type: none"> Matriz de confusión: [[1614 0] [0 90]] Precisión del modelo de regresión logística: 1.0 Área bajo la curva (AUC): 1.0
Análisis discriminante lineal	 <p>Curva ROC para LDA</p> <p>Tasa de verdaderos positivos</p> <p>Tasa de falsos positivos</p> <p>Curva ROC (AUC = 0.99)</p>	<ul style="list-style-type: none"> Matriz de confusión: [[1614 0] [11 79]] Precisión del modelo LDA: 0.9935446009389671 Área bajo la curva (AUC): 0.9890541098719539
KNN con K=3	 <p>Curva ROC para KNN con k=3</p> <p>Tasa de verdaderos positivos</p> <p>Tasa de falsos positivos</p> <p>Curva ROC (AUC = 0.68)</p>	<ul style="list-style-type: none"> Matriz de confusión para KNN: [[1598 16] [70 20]] Precisión del modelo KNN: 0.9495305164319249 Área bajo la curva (AUC) para KNN: 0.6830028913672037
Naive Bayes	 <p>Curva ROC para Naive Bayes</p> <p>Tasa de verdaderos positivos</p> <p>Tasa de falsos positivos</p> <p>Curva ROC (AUC = 0.74)</p>	<ul style="list-style-type: none"> Matriz de confusión para Naive Bayes: [[1614 0] [90 0]] Precisión del modelo Naive Bayes: 0.9471830985915493 Área bajo la curva (AUC) para Naive Bayes: 0.7355431639818257

A nivel general, la mayoría de los modelos resultaron ser muy buenos prediciendo la variable desempleo, por lo que podríamos decir que la matriz X utilizada sí explica a la variable respuesta. Para el año 2004 tanto la regresión logística como Naive Bayes tienen un nivel de precisión igual a 1, esto quiere decir que en el 100% de los casos que el modelo pronostica un caso como verdadero o positivo, en la realidad es así. Sin embargo los otros modelos como

LDA y KNN también tienen un alto nivel de predicción. Por otro lado, para el año 2024 los modelos más exactos son la regresión logística y el análisis discriminante lineal (99%) ya que además de que el valor de precisión es muy alto, el AUC es cercano a 1. De hecho en todos los casos, excepto para KNN de 2024, el área bajo la curva es mayor a 0.7, lo que indicaría que los modelos tienen un buen rendimiento.

Por último se utilizaron los datos de las personas que no respondieron su estado de actividad y se entrenó un modelo usando el método de regresión logística en ambos años para posteriormente poder identificar si cada dato se corresponde con una persona desocupada o no. Se obtuvo una proporción de 0% de precisión de personas desocupadas, una posible explicación es que el modelo fue entrenado con muy pocos datos de este tipo y el desbalance produjo que se prediga a la clase mayoritaria. En el caso del año 2024 se agregó el modelo Naive Bayes que había alcanzado la misma precisión que el de regresión.

PARTE 3: Análisis de la región Tucumán

Se creó otro *dataframe* llamado “tucuman” a partir de la selección de la columna “región” de cada una de las bases (2004 y 2024) cuyos datos correspondieran a Tucumán. Se realizó el mismo proceso de limpieza de datos utilizado en la Parte 1 y también se añadieron las columnas PET, PEA y desocupado. Luego se calcularon dos tasas de desocupación: según el INDEC y una alternativa. Para calcular la tasa de desocupación según el INDEC para la región Tucumán se consideró el número de desocupados sobre la cantidad de población económicamente activa. Para el año 2004 se obtuvo que el 16% de la PEA estaba desocupada, mientras que para 2024 este valor disminuyó 8,6%. Por otro lado, se calculó la tasa de desocupación alternativa mediante la cantidad de desocupados sobre la cantidad de población en edad de trabajar. En este caso hubo una diferencia de 1,23% entre 2004 y 2024. Las tasas obtenidas pueden observarse en la tabla 1.

Año	TD INDEC	TD Alternativa
2004	0.168880	0.173998
2024	0.082822	0.186387

Tabla 1: tasas de desocupación para Tucumán

BIBLIOGRAFÍA

- Mankiw, N. G. (2008). *Macroeconomics* (6.ª ed.). Worth Publishers.

Mensaje adicional: fuimos a un evento para despejar un poco la cabeza y no pudimos dejar de pensar en ustedes y en esta materia :)

