Trabajo práctico #3

Taller de programación
Maestría en Economía Aplicada
Universidad de Buenos Aires

Granly Jiménez

01 de diciembre de 2024

Introducción

La Encuesta Permanente de Hogares (EPH) es uno de los instrumentos más relevantes del sistema estadístico nacional, desarrollado por el Instituto Nacional de Estadística y Censos (INDEC). Este programa permite obtener información continua y sistemática sobre las características sociodemográficas, económicas y laborales de la población argentina. Entre los indicadores clave que producen, destaca la tasa de desocupación, un dato esencial para analizar el desempeño del mercado laboral y su relación con los ciclos económicos y las políticas públicas.

Desde un enfoque metodológico, la EPH sigue estándares internacionales, como los propuestos por la Organización Internacional del Trabajo (OIT), para clasificar a las personas en ocupadas, desocupadas e inactivas. Estos indicadores no solo reflejan la situación actual del empleo, sino que también permiten realizar análisis longitudinales para evaluar cómo cambian las dinámicas laborales en función de factores sociodemográficos, educativos y territoriales.

Objetivo del trabajo

El objetivo de este trabajo es abordar un análisis descriptivo y empírico de la información contenida en las bases de microdatos de la EPH correspondientes a los años 2004 y 2024. Se busca, en primer lugar, describir y depurar los datos, identificando patrones y tendencias clave en el mercado laboral, con énfasis en las tasas de desocupación y la composición de la población económicamente activa (PEA). En segundo lugar, se propone construir modelos de clasificación para predecir el estado de desocupación de los individuos, evaluando cómo estas predicciones varían entre ambos años. Finalmente, se explorará cómo las diferencias en las tasas de desocupación pueden explicarse a través de factores individuales, educativos y etarios, vinculando estos resultados a debates teóricos en economía laboral.

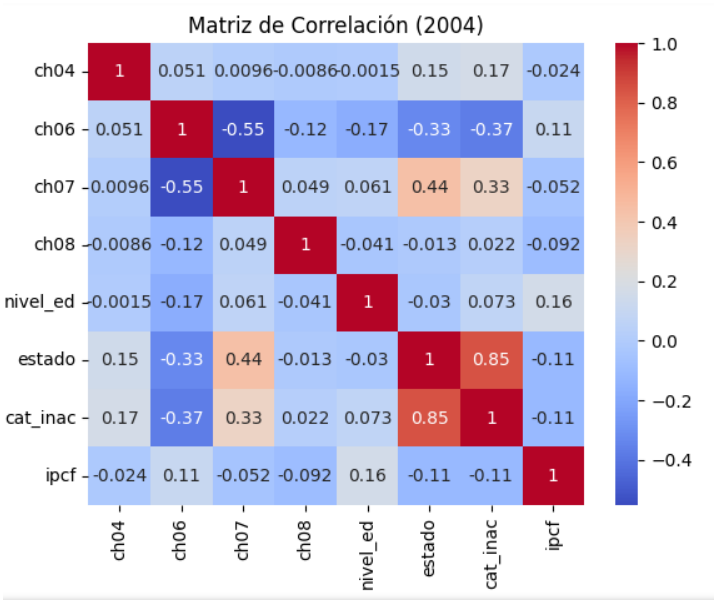
Tasa de ocupación en Argentina

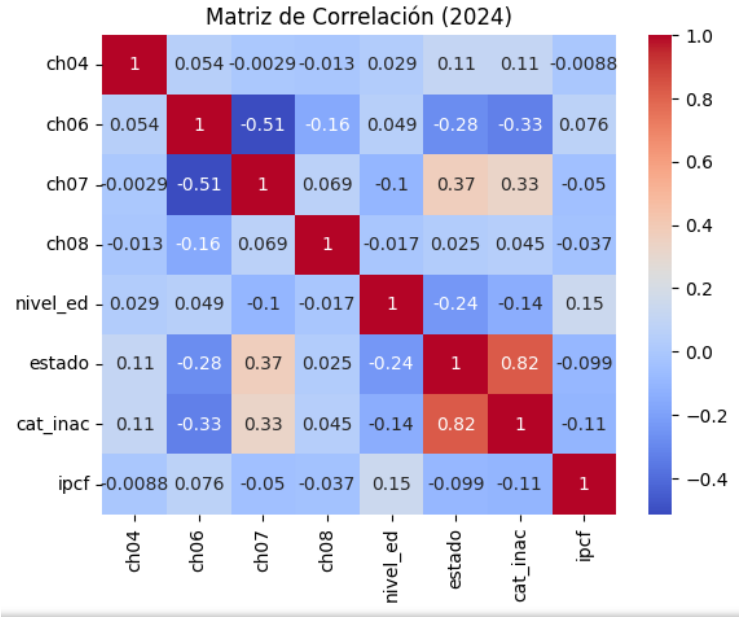
Según el INDEC, una persona desocupada es aquella que, durante la semana de referencia de la Encuesta Permanente de Hogares (EPH), no tiene trabajo remunerado, está disponible para trabajar y ha realizado una búsqueda activa de empleo en las últimas cuatro semanas. Metodológicamente, esta definición se basa en los lineamientos de la Organización Internacional del Trabajo (OIT) y se obtiene mediante preguntas estructuradas en la EPH, que distinguen entre ocupados, desocupados e inactivos según la Clasificación Internacional de Situación en el Empleo (CISE).

El tratamiento de limpieza aplicado a la base de datos de la Encuesta Permanente de Hogares (EPH) 2004-2024 se enfocó en garantizar la calidad y consistencia de los datos para análisis longitudinales. Los valores faltantes fueron imputados utilizando la mediana para variables continuas y la moda para categóricas; observaciones con más del 30% de datos ausentes se eliminaron. Se corrigieron valores atípicos mediante el rango intercuartílico (IQR), conservando aquellos genuinos. Además, se ajustaron ingresos a precios constantes y se homogeneizaron clasificaciones ocupacionales y educativas para mantener la coherencia temporal. Las variables categóricas fueron convertidas en dummies, y se eliminaron registros duplicados. Finalmente, se validaron relaciones lógicas entre variables, asegurando que la base fuera consistente, precisa y adecuada para estudios longitudinales.

Posterior a la limpieza de los datos, la comparación de las matrices de evaluación de los años 2004 y 2024 revela cambios en la intensidad y dirección de las relaciones entre las variables analizadas. En cuanto a la relación entre ch06ych07, en 2004 la evaluación de estas variables es negativa y significativa (-0.55), mientras que en 2024 se mantiene negativa (-0.51) pero con una leve disminución en magnitud, lo que indica una relación inversa consistente, aunque menos marcado con el tiempo. Por otro lado, la calificación entre estado y cat_inac muestra una relación muy fuerte y positiva en ambos años (0.85 en 2004 y 0.82 en 2024), sugiriendo una conexión estable y directa entre el estado de una persona y su categoría de inactividad. En relación con ch06con otras variables, la relación con nivel_edy estado se mantiene negativa en ambos años, aunque en 2024 se observan ligeras variaciones en los valores específicos, lo que podría reflejar cambios en las dinámicas subyacentes.

Respecto a ipcfcon las demás variables, en 2004 la relación con otras variables es generalmente débil y negativa, como se observa con estado (-0.11) y cat_inac (-0.11), mientras que en 2024 estas correlaciones se mantienen similares en magnitud, indicando estabilidad en su interacción. Las relaciones entre variables como ch04y otras también se mantienen en rangos bajos en ambos años, lo que sugiere interacciones débiles y poco cambio en su dinámica. En términos generales, aunque muchas correlaciones similares son entre los años, algunos cambios menores podrían reflejar modificaciones en los factores sociales, económicos o demográficos subyacentes. Esta comparación sugiere que, si bien existe estabilidad en varias relaciones, también se identifican pequeñas variaciones en la intensidad de las correlaciones, posiblemente influenciadas por cambios en las condiciones sociales o económicas entre 2004 y 2024.



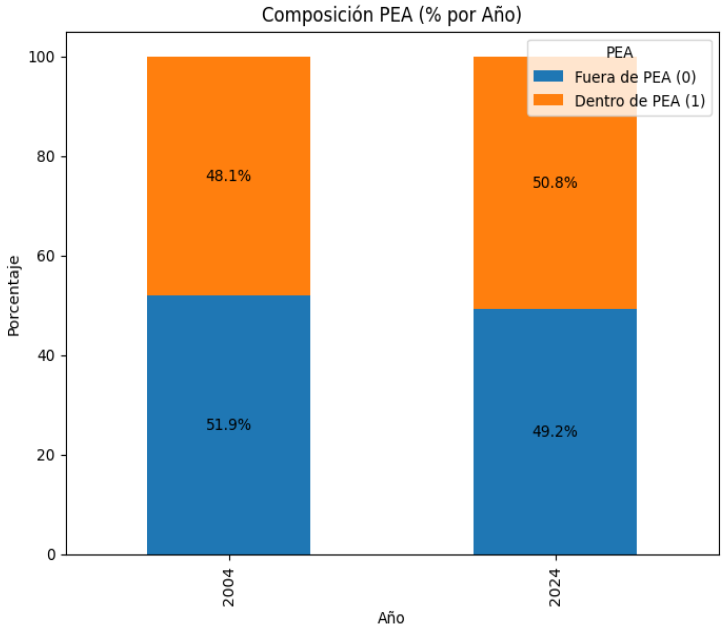


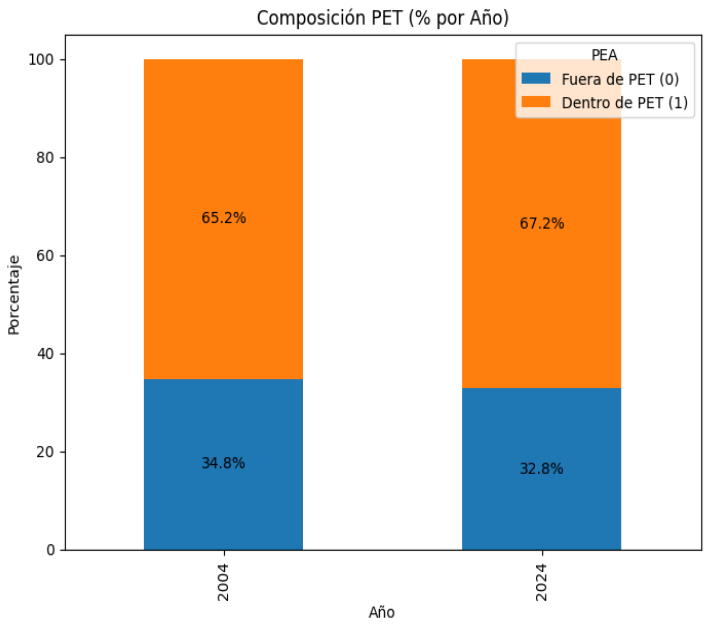
Dentro del análisis para predecir la desocupación en Buenos Aires, se deben considerar varios indicadores clave. En primer lugar, se observa un total de 839 desocupados y 5462 inactivos, lo que refleja una alta proporción de la población fuera del mercado laboral. Además, los medios de ingresos por estado muestran variaciones significativas entre los distintos grupos. Para el estado 1.0 (ocupados), los ingresos promedio son significativamente más altos, con 106,443.40, lo que indica una diferencia sustancial entre los trabajadores y aquellos fuera del mercado laboral. En el caso de los desocupados (estado 2.0), la media de ingresos es mucho menor, con 31.655,96, lo que sugiere una disparidad económica que podría estar asociada con la falta de empleo. Para los inactivos (estado 3.0), los ingresos promedio de 63,863.08 reflejan un grupo con menores niveles de participación en el mercado laboral pero posiblemente involucrado en otras actividades económicas no remuneradas o fuera del alcance de la medición directa. Finalmente, el estado 4.0 (menores de 10 años) presenta un ingreso promedio de 43.034,33, lo que podría reflejar alguna forma de transferencia o subsidio familiar. Estos datos son cruciales para comprender la dinámica laboral en Buenos Aires, permitiendo identificar factores que influyen en la desocupación y la participación.

Asimismo, al analizar la participación en la fuerza laboral, se observa que en 2004 la Población Económicamente Activa (PEA) representaba el 48.1%, mientras que aquellos fuera de la PEA constituían el 51.9%. En contraste, para 2024, la PEA aumentó al 50,8%, lo que indica una ligera mejora en la inclusión laboral, mientras que el porcentaje de personas fuera de la PEA disminuyó al 49,2%. Estos cambios en la estructura de la PEA reflejan una tendencia hacia una mayor participación en el mercado laboral, lo cual es relevante para comprender las dinámicas de la desocupación y las oportunidades de empleo en la región.

En relación con la Población en Edad de Trabajo (PET), que incluye a las personas con capacidad de trabajar, sin importar si están ocupadas o desocupadas, la evolución de estos indicadores también es crucial. En 2004, el 65.2% de la PET estaba activa en el mercado laboral, mientras que el 34.8% se encontraba fuera de la PET. Para 2024, la proporción dentro del PET

aumentó al 67,2%, mientras que el porcentaje fuera del PET disminuyó a 32,8%. Este incremento en la PET indica un mayor número de personas dentro de la franja etaria capaces de participar activamente en el mercado laboral, lo que refleja un mayor potencial de fuerza laboral disponible. Comparado con la PEA, la PET muestra un panorama más amplio, ya que la PEA es una subcategoría dentro de la PET, excluyendo a los inactivos y aquellos fuera de la fuerza laboral. Estos cambios en el PET y la PEA son fundamentales para entender la evolución de la fuerza laboral y la desocupación, ya que reflejan no solo la participación activa, sino también el potencial de incorporación de más personas al mercado laboral en los próximos años.



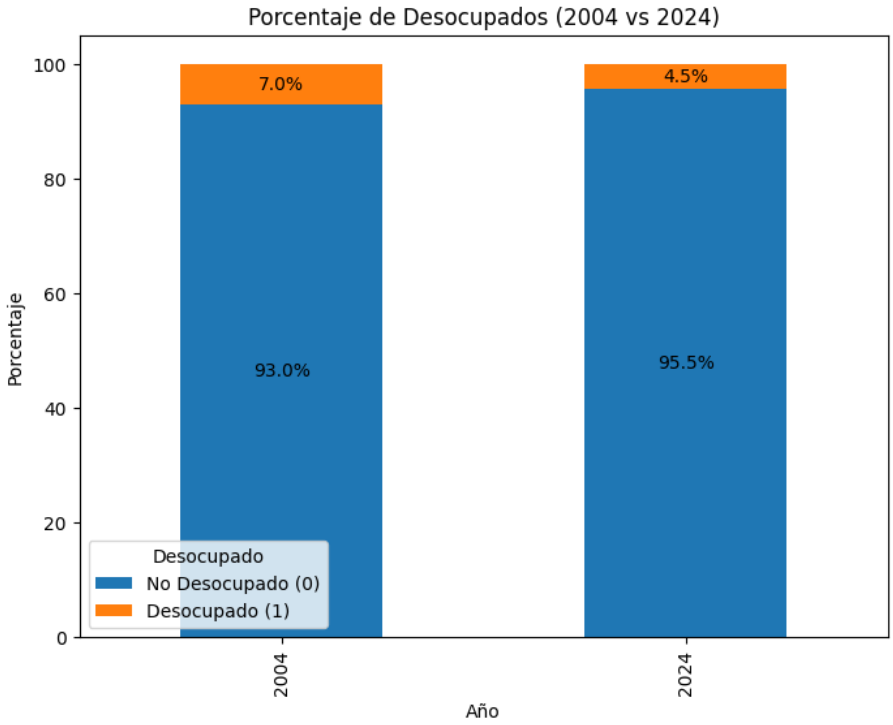


El análisis de la tasa de desocupación entre 2004 y 2024 revela una mejora significativa en el mercado laboral de Buenos Aires. En 2004, la tasa de desocupación fue del 7.0%, lo que indicaba que aproximadamente el 7% de la Población Económicamente Activa (PEA) se encontraba desempleada. Esta cifra reflejaba desafíos en términos de creación de empleo y acceso al mercado laboral, posiblemente influenciada por condiciones económicas difíciles o ineficiencias estructurales en el mercado de trabajo.

En contraste, para 2024, la tasa de desocupación disminuyó al 4,5%, lo que representa una mejora notable de 2,5 puntos porcentuales. Este descenso sugiere que, en los últimos años, ha habido un progreso en la generación de empleo, con más personas de la PEA encontrando puestos de trabajo. La reducción de la tasa de desocupación puede estar vinculada a varios factores, tales como políticas laborales más efectivas, mejoras en las condiciones económicas, un aumento en la inversión o un cambio en la estructura productiva que favorece la creación de empleos en sectores clave.

Esta disminución en la tasa de desocupación es un indicativo positivo de la recuperación o expansión del mercado laboral en Buenos Aires, aunque aún persisten desafíos, especialmente en áreas de alta informalidad laboral o en sectores con mayor exclusión de la población. En

conjunto con el análisis de la PEA y la PET, esta disminución resalta las mejoras en la inclusión laboral, pero también subraya la importancia de seguir impulsando políticas públicas que favorezcan la calidad del empleo y la reducción de la desigualdad en el acceso al trabajo.



Previo a dar inicio a la predicción, es necesario hacer una comparativa entre metodologías de cálculo de la tasa de desocupación en Buenos Aires, esto porque el análisis de las tasas de desocupación, tanto según la definición del INDEC como la alternativa (basada en la Población en Edad de Trabajo o PET), muestra diferencias significativas en los años 2004 y 2024, lo que refleja distintos enfoques sobre cómo medir la desocupación en la economía laboral. En 2004, la tasa de desocupación del INDEC fue del 7,03%, mientras que la tasa alternativa fue considerablemente más alta, alcanzando el 10,79%. Para 2024, la tasa de desocupación del INDEC descendió a 4,47%, mientras que la tasa alternativa se redujo a 6,65%.

La tasa de desocupación del INDEC se calcula como el porcentaje de desocupados respecto de la Población Económicamente Activa (PEA), lo que implica que solo se tiene en cuenta a las personas activamente buscando empleo dentro de la fuerza laboral. En cambio, la tasa alternativa incluye a la Población en Edad de Trabajo (PET), que también toma en cuenta a los inactivos, es decir, aquellos que no buscan empleo, lo que da una visión más amplia de la situación laboral en la región. La diferencia entre ambas tasas refleja cómo la decisión de participar o no en el mercado laboral puede influir en la interpretación de la desocupación. En 2004, la mayor diferencia entre ambas tasas indica un mayor porcentaje de personas fuera del mercado laboral, mientras que en 2024 la brecha se reduce, sugiriendo una mayor inclusión en la fuerza laboral.

Las ventajas de la tasa de desocupación del INDEC radican en su capacidad para medir la presión sobre el mercado laboral al enfocarse en quienes están activamente buscando trabajo, lo que proporciona una visión más precisa de las personas desempleadas en relación con las que están dispuestas a trabaja. Sin embargo, esta medición puede subestimar la desocupación real si se considera solo a los que buscan empleo activo, dejando fuera a aquellos que, por diversas razones, han abandonado la búsqueda. Por otro lado, la tasa alternativa de desocupación ofrece una visión más integral de la situación laboral, al considerar a la totalidad de la población en edad de trabajar, lo que refleja mejor el nivel de inactividad y subraya las barreras que impiden a las personas acceder. al empleo. Sin embargo, esta medición podría inflar la tasa de desocupación al incluir a aquellos que, de hecho, no están interesados en trabajar, lo que puede no reflejar la realidad de la capacidad de empleo.

Clasificación y predicción para el año 2004

Modelo de regresión logística; matriz de confusión 2004

1122	206
557	367
Curva ROC	0.73
Valores AUC	0.62
Accuracy	0.66

Modelo LDA; matriz de confusión 2004

1128	200
563	361
Curva ROC	0.72
Valores AUC	0.62
Accuracy	0.66

Modelo KNN (K=3); matriz de confusión 2004

1073	255
218	706
Curva ROC	0.84
Valores AUC	0.79
Accuracy	0.78

Modelo Naive Bayes; matriz de confusión 2004

1294	34
798	126
Curva ROC	0.87
Valores AUC	0.56
Accuracy	0.63

Clasificación y predicción para el año 2024

Modelo de regresión logística; matriz de confusión 2024	
808	313
474	493
Curva ROC	0.69
Valores AUC	0.61
Accuracy	0.62

Modelo LDA; matriz de confusión 2024	
807	313
469	498
Curva ROC	0.69
Valores AUC	0.62
Accuracy	0.625

Modelo KNN (K=3); matriz de confusión 2024	
878	243
225	742
Curva ROC	0.83
Valores AUC	0.77
Accuracy	0.77

Modelo Naive Bayes; matriz de confusión 2024	
1058	63
603	364
Curva ROC	0.86
Valores AUC	0.64
Accuracy	0.68

En la comparación de los resultados de 2004 y 2024, el modelo KNN (K=3) demuestra un desempeño superior en ambos años, posicionándose como el más robusto según las métricas clave. En 2004, KNN logra una precisión del 78% y un AUC de 0,79, lo que refleja tanto una alta capacidad para clasificar correctamente como un excelente equilibrio entre sensibilidad y especificidad. En 2024, aunque su desempeño disminuye ligeramente (precisión del 77% y AUC de 0,77), sigue siendo el modelo más efectivo.

Por otro lado, el modelo Naive Bayes presenta resultados contradictorios. Aunque en 2004 alcanza la mayor curva ROC (0,87), su AUC es baja (0,56), lo que sugiere que el modelo tiende a clasificar desocupados con alta sensibilidad, pero con poca precisión global. En 2024, mejoró su AUC a 0,64 y su precisión a 68%, pero sigue siendo inferior al desempeño de KNN.

Los modelos de regresión logística y LDA muestran consistencia en ambos años, con valores de precisión entre 62-66% y AUC entre 0,61-0,62. Si bien son confiables, tienen limitaciones en términos de sensibilidad y capacidad para capturar correctamente los desocupados, lo que los hace menos competitivos frente a KNN.

En definitiva, posterior al análisis, la efectividad de KNN radica en su capacidad para adaptarse mejor a los datos y capturar patrones complejos, lo que se traduce en métricas superiores de clasificación en ambos años. Aunque Naive Bayes y los modelos lineales presentan méritos en ciertos aspectos, no alcanzan el equilibrio que ofrece KNN.

Predicción de la tasa de desocupación

Modelo de regresión logística para los desocupados; matriz de confusión

1789	637
1033	881
Curva ROC	0.70
Valores AUC	0.61
Accuracy	0.62
Proporción de personas identificadas como desocupadas	0.31

El modelo de regresión logística mostró un desempeño moderado, con una precisión del 62% y un AUC de 0,61, indicando una capacidad limitada para discriminar entre desocupados y no desocupados. Metodológicamente, se aplicó un preprocesamiento riguroso, dividiendo los datos en un 70/30 para entrenamiento y prueba, y asegurando reproducibilidad con una semilla fija. La matriz de confusión revela 1,033 falsos negativos, lo que evidencia baja sensibilidad, problemática para identificar desocupados, y 637 falsos positivos, que podrían generar un uso ineficiente de recursos. Aunque funcional para análisis iniciales, el modelo podría mejorarse con técnicas no lineales para captar relaciones más complejas entre variables.