Trabajo práctico número 3

Carrera

Licenciatura en Ciencias del comportamiento

Asignatura

Ciencia de datos

Autoras

Luciana Crupnik

Josefina Maria Laborda Moreno

Justina Reinke

Profesor

Ignacio Spiousas

Maria Noelia Romero

La Encuesta Permanente de Hogares (EPH), realizada por el INDEC, es un programa nacional que recopila datos continuos sobre las características sociodemográficas y económicas de la población. Entre los indicadores clave que proporciona, destaca la tasa de desocupación en el mercado laboral.

Parte 1: Análisis exploratorio de la base

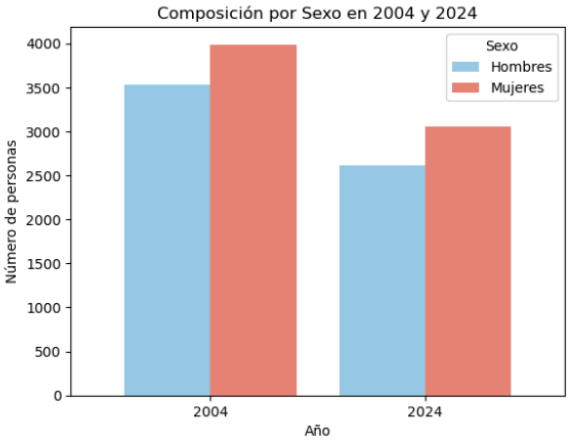
En esta sección, se utilizarán dos bases de datos pertenecientes a la EPH, que corresponden al primer trimestre del 2004 y al primer trimestre del 2024. El objetivo de esta sección es dar a conocer aquellas variables que son relevantes para el análisis y sus observaciones.

Para determinar a las personas desocupadas, se identificó la variable “ESTADO”, la cual indica las condiciones de actividad de las personas, ya sea si estaban ocupadas, desocupadas, inactivas, menores a 10 años o si no había respuesta al cuestionario individual. Para el primer trimestre de 2004, se identificaron 2717 desocupados, a diferencia del primer trimestre del 2024, con 1362 desocupados.

Posteriormente, se realizó una filtración de aquellas observaciones de la variable “AGLOMERADOS”, que solo pertenecían a Ciudad Autónoma de Buenos Aires y al Gran Buenos Aires. Luego, se unificaron las bases de datos para poder trabajar con un mismo conjunto de datos que obtengan tanto las observaciones de 2004 y del 2024.

Sin embargo, se identificaron que algunas de las variables contenían respuestas sin sentido. Por ejemplo, la variable “CH06” contiene datos que responden a la pregunta “¿Cuántos años cumplidos tiene?” o la variable “P47T” la cual hace referencia al monto total de ingreso individual percibido en el mes de referencia (sumatoria de ingresos laborales y no laborales). Para estos casos, se eliminaron todas aquellas respuestas con valores menores a 0. A su vez, en el proceso de depuración de los datos, se eliminaron columnas irrelevantes para el análisis, para simplificar el conjunto de datos y concentrar el análisis en variables clave. Entre ellas, pp09c, pp09c_esp, idecocur, pdecocur, idecindr, pdecindr, idecifr, pdecifr, ideccfr y pdeccfr.

Una vez realizada la limpieza, se implementó un gráfico de barras para comparar la composición de hombres y mujeres correspondientes a cada trimestre. Este gráfico permite visualizar las variaciones en la composición de la población por sexo entre ambos períodos. A pesar de la limpieza y unificación de las bases, la representación muestra cómo el porcentaje de mujeres continúa siendo superior en ambos años. En 2004, hay una mayor diferencia a favor de las mujeres, mientras que en 2024 se observa una leve disminución en la cantidad de hombres.



1.1 Gráfico de barras. Composición por sexo en 2024 y 2004.

Luego, la matriz de correlación permite analizar las relaciones entre diferentes variables sociodemográficas y económicas en los años 2004 y 2024. En 2004, observamos algunas correlaciones significativas. La edad (CH06) muestra una correlación leve negativa con el nivel educativo (-0.17), lo que sugiere que las personas mayores tienden a tener un nivel educativo más bajo. Asimismo, la relación entre estado civil (CH07) y edad es negativa y moderada (-0.55). Hay una correlación positiva entre estado de actividad (estado) y categoría de inactividad (cat_inac) de 0.85, lo cual indica una relación estructural entre ser inactivo y pertenecer a ciertas categorías, como jubilados o estudiantes. Por último, la relación entre estado civil y estado de actividad es moderadamente positiva (0.44), lo que podría reflejar una relación entre el estado civil y la condición laboral en esta muestra.

En 2024, la correlación entre edad y estado civil se mantiene moderadamente negativa (-0.55), lo que reafirma la tendencia observada en 2004 de que el estado civil de las personas varía significativamente según la edad. La relación entre estado de actividad y categoría de inactividad sigue siendo alta (0.81), lo que subraya una asociación persistente entre estar inactivo y ciertas categorías de inactividad. La correlación entre estado civil y estado de actividad también permanece moderadamente positiva (0.42), aunque ligeramente menor que en 2004.

En ambos años, la correlación entre el ingreso per cápita familiar (IPCF) y el estado de actividad es baja, lo cual indica que el estado ocupacional no está fuertemente relacionado con el ingreso familiar per cápita.

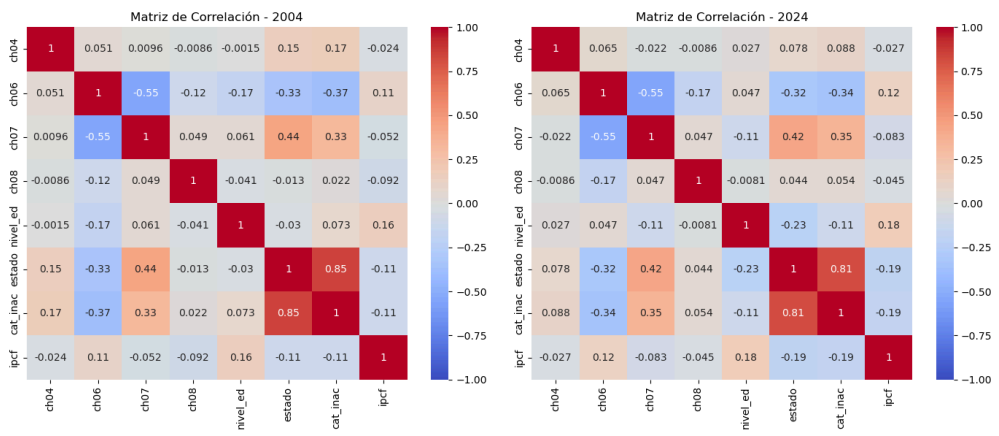


Figura 1.2 matriz de correlación para los años 2004 y 2024

En cuanto a la cantidad de desocupados presentes en la muestra, se identificaron un total de 528 para el año 2004 y 281 para el 2024. Por otro lado, 2800 personas se encontraban inactivas en el año 2004 y 2440 en el 2024. Se calculó la media para la variable “IPCF”, la cual indicaba el monto de ingreso per cápita familiar percibido en el mes de referencia, según si el estado de la persona, para cada año. En primer lugar, en el 2004, la media del ingreso para las personas con estado “ocupado”, la media fue de 476, para las personas con estado “desocupado” fue de 224, y para los “inactivos”, de 315. En segundo lugar, en el 2024, los valores de las medias de los ingresos aumentaron significativamente, cuyo incremento puede deberse a cambios en el valor de la moneda, a la inflación y otros cambios en la economía durante 20 años. Para aquellas personas con estado “ocupado”, se obtuvo una media de 307791, para los “desocupados” de 94095 y para los “inactivos” de 142072.

Se identificaron y separaron las observaciones en función de si respondieron o no la pregunta sobre su condición de actividad (variable ESTADO). En total, se contabilizó que 10 personas no respondieron sobre su condición de actividad.

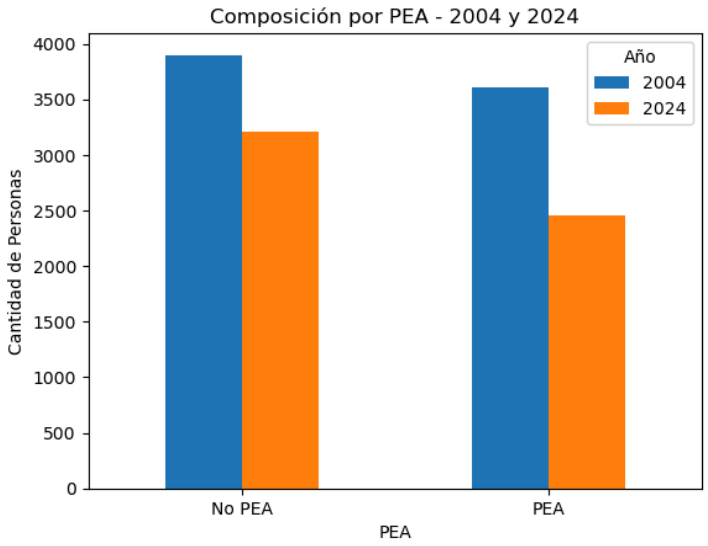


Figura 1.3 Composición por Población Económicamente activa

Luego se realizó un gráfico de barras (*Figura* 1.3) que muestra la composición de la Población Económicamente Activa (PEA) y No PEA en los años 2004 y 2024. En 2004, se observa una mayor cantidad de personas tanto en la PEA como en la No PEA en comparación con 2024. Esta disminución en 2024 sugiere una reducción en el número de personas que respondieron estar activas o inactivas en el mercado laboral, lo cual podría deberse a factores demográficos, como una menor población total, o a cambios en la estructura de la fuerza laboral que afectan ambas categorías. La disminución en el número

de personas tanto en la PEA como en la No PEA en 2024 en relación con 2004 podría indicar una reducción en la participación laboral de la población.

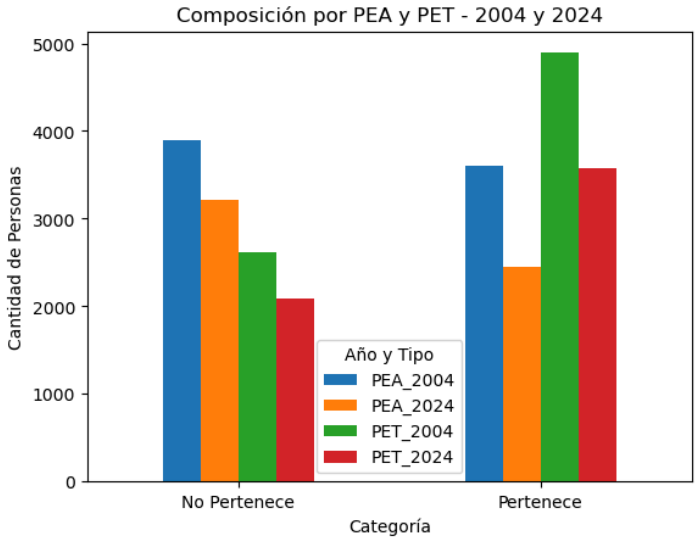


Figura 1.4 Gráfico de barras. Composición por PEA y PET

A continuación, se completó el análisis exploratorio con otro gráfico de barras que ilustra la composición de la Población Económicamente Activa (PEA) y la Población en Edad para Trabajar (PET) en los años 2004 y 2024. Se observa que en ambos años, la PET es mayor que la PEA, lo cual es esperado, ya que la PET incluye a todas las personas entre 15 y 65 años, independientemente de su participación activa en el mercado laboral.

En el año 2004, tanto la PEA como la PET tienen una cantidad considerable de personas que pertenecen a sus categorías respectivas, aunque la PET es notablemente mayor. Esto sugiere que no todas las personas en edad de trabajar están activas en el mercado laboral. En 2024, tanto la PEA como la PET presentan una disminución en el número de personas en comparación con 2004, lo cual podría estar relacionado con cambios demográficos o económicos que afectan la participación laboral y la estructura de la población en edad de trabajar. Es importante destacar que la cantidad de personas desocupadas en 2004 fue de 528 mientras que en 2024 fue de 281.

Parte 2: Clasificación

En esta segunda parte, se buscará prever si una persona se encuentra desocupada o no, utilizando diversas variables relacionadas con sus características individuales.

Para comenzar, se dividieron los datos de la base “respondieron” para cada año, en una base de prueba, con el 30% de los datos, y una base de entrenamiento con el 70% de los datos. Las dimensiones de los conjuntos se definieron de la siguiente manera. Para el año 2004, el conjunto de entrenamiento cuenta con 5,254 observaciones y 174 variables,

mientras que el conjunto de prueba contiene 2,252 observaciones y las mismas 174 variables. Para el año 2024, el conjunto de entrenamiento incluye 3,966 observaciones con 174 variables, y el conjunto de prueba tiene 1,701 observaciones y 174 variables también. La variable “desocupado” se estableció como dependiente y el resto de las variables como independientes.

Posteriormente, se implementaron los siguientes métodos de clasificación.

En primer lugar, la regresión logística es un modelo estadístico que permite predecir la probabilidad de pertenecer a un grupo. La precisión de este modelo es muy alta, la cual indica que el modelo hace predicciones correctas el 99.5% de los casos. A su vez, el AUC, con un valor de 0.9987, indica una excelente capacidad para discriminar entre grupos. Por último, se implementó una matriz de confusión, en la cual se identificaron 2096 casos negativos y 145 casos positivos. Hubo un solo caso de falso positivo, en donde se predijo positivo cuando era negativo, y 10 casos de falso negativo, es decir que se predijo negativo cuando era positivo. En la curva ROC, cuanto más cerca esté del eje superior izquierdo, mejor capacidad de clasificación. En este caso, observamos que la capacidad es muy alta.

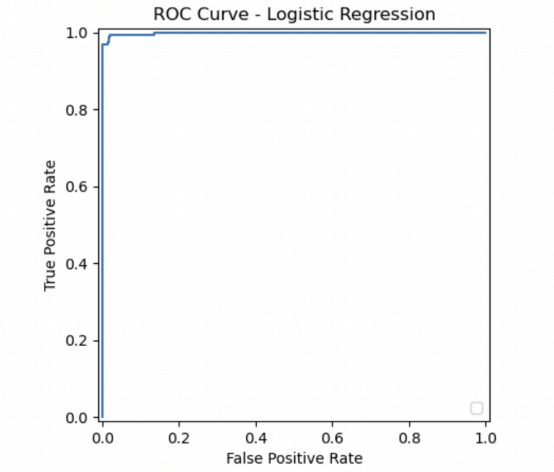


Figura 1.5 Curva ROC de Regresión Logística

En segundo lugar, se implementó un análisis discriminante lineal. El modelo en este caso, presentó un 92.8% de precisión; apesar de que sea un valor muy alto, es menor en comparación con la regresión logística. Lo mismo sucede con el AUC, con un valor de 0.9753, no supera la capacidad de discriminación delo modelo anterior. Por último, la matriz de confusión identificó correctamente 2084 casos negativos y solo 5 positivos. Sin embargo, hay 13 falsos positivos y 150 falsos negativos. Apesar de que la curva ROC en este caso se encuentre cerca del eje superior izquierda, está mínimamente más alejado que la curva de la regresión, lo que nos indica que este modelo no supera al anterior.

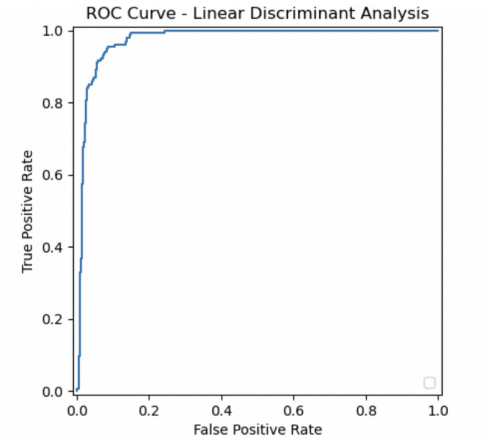


Figura 1.6 Curva Roc de Análisis Discriminante Lineal

En tercer lugar, el método KNN con k=3, indica que el modelo busca los 3 vecinos más cercanos en el espacio de características y asigna la clase más común entre ellos. La presión de este modelo es similar al de LDA, con un valor de 92.4%. Sin embargo, el AUC es considerablemente más bajo que los anteriores, con un valor de 0.6895. La matriz de confusión identificó correctamente 2047 casos negativos y 33 positivos, pero presentó 50 falsos positivos y 122 falsos negativos. La curva ROC se encuentra notablemente más alejada del eje superior izquierdo.

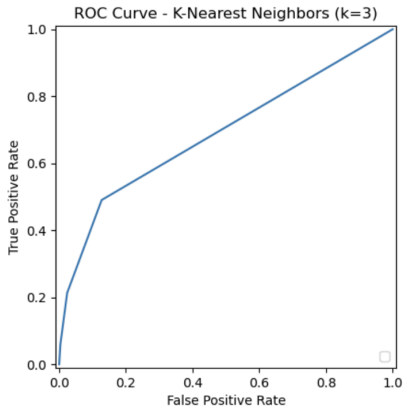


Figura 1.7 Curva ROC de KNN

En cuarto y último lugar, el modelo Naive Bayes es un método probabilístico basado en el teorema de Bayes. La precisión de este modelo supera todas las anteriores, con un valor 99.69%. El AUC es de 1.0, lo que indica un rendimiento perfecto para diferentenciar entre clases. La matriz de confusión identificó 2097 positivos y 148 negativos, 0 falsos positivos y 7 falsos negativos. La curva ROC se ajsuta perfectamente al punto (0,1). Por lo

tanto, se puede concluir que este modelo tiene un excelente rendimiento a comparación del los tres anteriores.

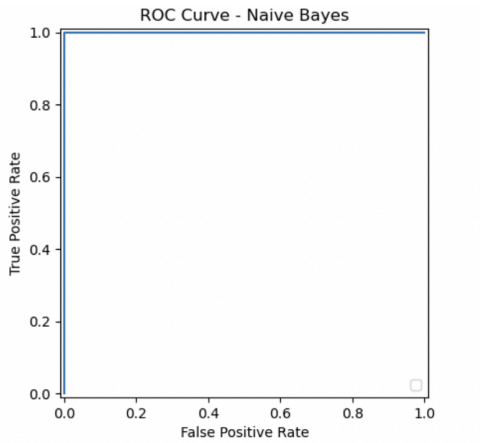


Figura 1.8 Curva ROC de Naive Bayes

Luego, se realizó un el análisis comparativo de los modelos de predicción entre 2004 y 2024, se encontró que el método Naive Bayes fue el mejor predictor en 2004, con un AUC (área debajo de la curva Roc que evalúa la capacidad para distinguir entre dos clases) de 1.0, lo cual indica una clasificación perfecta, además de un accuracy muy alto (0.9969) y el MSE para los datos de testeo más bajo (0.0031) entre todos los métodos. Esto sugiere que Naive Bayes logra diferenciar claramente entre las clases en ese año, siendo el modelo con el mejor rendimiento general.

En cambio, para el año 2024, el análisis discriminante lineal resultó ser el modelo más adecuado. Aunque el MSE de análisis discriminante lineal (0.0564) fue ligeramente superior al de Naive Bayes, su AUC de 0.9746 destacó entre los modelos, lo que sugiere una mejor capacidad de discriminación para clasificar entre las clases en este año. Además, el accuracy de análisis discriminante lineal (0.9436) fue comparable al de los otros modelos, consolidándolo como el método de clasificación para 2024.

Utilizando el método seleccionado (Naive Bayes para 2004 y Análisis Discriminante Lineal para 2024), se realizaron predicciones sobre la base *norespondieron* para identificar qué personas podrían ser desocupadas. Es decir, se utilizó los modelos con mejor performance para predecir los desocupados de la base de datos de aquellos que no respondieron su estado. Los resultados mostraron que la proporción de personas identificadas como desocupadas fue 0.0 tanto en 2004 como en 2024, lo cual significa que el modelo no clasificó a ninguna persona de esta base como desocupada. Este resultado sugiere que el modelo no encontró patrones suficientes en la base *norespondieron* para identificar desocupados, posiblemente debido a que esa base de datos solo contaba con 10 datos. Es decir, al desbalance de clases en el conjunto de entrenamiento.