TRABAJO PRÁCTICO Nº3

CIENCIA DE DATOS

Entrega TP3

Asignatura

Ciencia de Datos

Primavera 2024

Alumnas

Micaela Vollert

Sofía Dillon

Juana Matticoli

Profesores

María Noelia Romero

Ignacio Spiousas

Fecha de entrega

3/11/2024

PARTE I: Analizando la base

EJERCICIO 1

Identificamos a las personas desocupadas en la información disponible en la página del INDEC a través de la Encuesta Permanente de Hogares (EPH). Esta encuesta realiza preguntas enfocadas en aspectos como el empleo, el desempleo, la estructura de los hogares, los ingresos y las condiciones de vida. Se puede determinar la situación laboral de los individuos, y se considera que una persona es desocupada si cumple las siguientes condiciones:

- No se encuentra empleado (no tiene trabajo) al momento de responder la encuesta
- Busca activamente empleo: La persona actualmente ha enviado currículums, ha asistido a entrevistas, o se ha inscripto a agencias de empleo durante el último mes
- No ha realizado ningún trabajo ocasional durante ese período.

EJERCICIO 2

Después de descargar los archivos correspondientes de la página web del INDEC, eliminamos todas las observaciones que no corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires. Para ello, buscamos los códigos de aglomerados en el diccionario de variables “Diseño de registro y estructura para las bases preliminares" donde el número 32 correspondía a CABA y el 33 a partidos de GBA. Antes de proceder a unir las bases de datos del 2004 y del 2024 en una sola base, nos aseguramos de que las variables comunes existentes entre ambas tengan los mismos nombres y en el mismo formato. Por lo tanto, transformamos todos los nombres de las columnas en minúsculas. Luego, unimos ambas bases utilizando la función ‘concat’, obteniendo una nueva base de 14698 filas y 181 columnas.

A continuación, realizamos la limpieza de la base de datos. En primer lugar, eliminamos aquellas columnas que contenían una gran cantidad de valores vacíos o ‘NaN’ y que contenían datos que consideramos irrelevantes. También, eliminamos aquellas observaciones con valores que no tenían sentido: edades menor a 0 o mayor a 100, ingreso menor a 0, y sexo distinto a 1 o 2 (es decir, valores que no sean mujer o varón). Chequeamos que se hayan eliminado los valores inadecuados a modo de prueba, imprimiendo los valores mínimos y máximos de ingreso, y los valores mínimos y máximos de edad.

Procedimos a realizar un gráfico de barras mostrando la composición por sexo para 2004 y 2024 (Ver **Figura 1**). Observamos que en ambos años se observa que hay más cantidad de mujeres que de hombres. La diferencia entre los sexos parece ser similar en ambos períodos. Además, notamos que hay menos cantidad de observaciones, tanto en hombres y mujeres, en el año 2024.

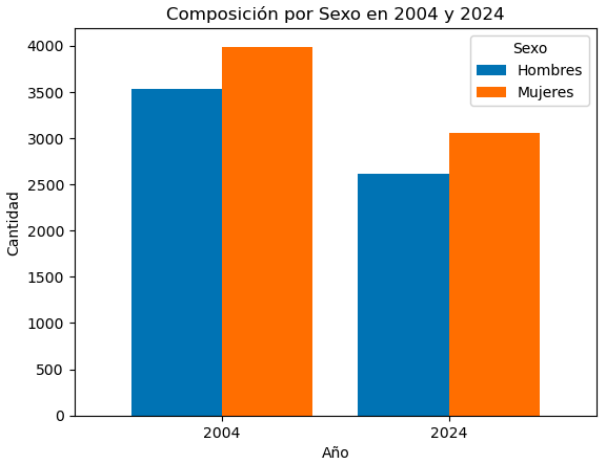

Figura 1. Composición por Sexo en 2004 y 2024.

Luego, realizamos una matriz de correlación para 2004 y 2024 con las variables: CH04, CH06, CH07, CH08, NIVEL ED, ESTADO, CAT_INAC, IPCF. Utilizamos el comando 'seaborn' para graficar las matrices con mapas de calor, y así ilustrar con colores el grado de las correlaciones. Observamos que las correlaciones de 2004 son muy similares a las correlaciones de 2024. Además, se evidencia en ambos períodos una fuerte correlación entre las variables 'cat inac' y 'estado', y una correlación débil entre 'ch06' y 'ch07'. Ver **Figura 2** y **Figura 3** a continuación.

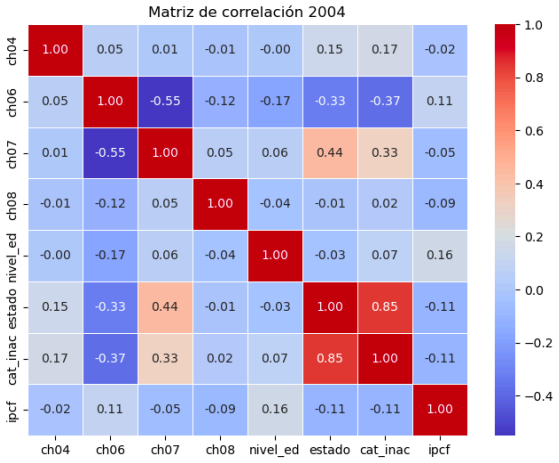

Figura 2. Matriz de correlación entre variables en 2004.

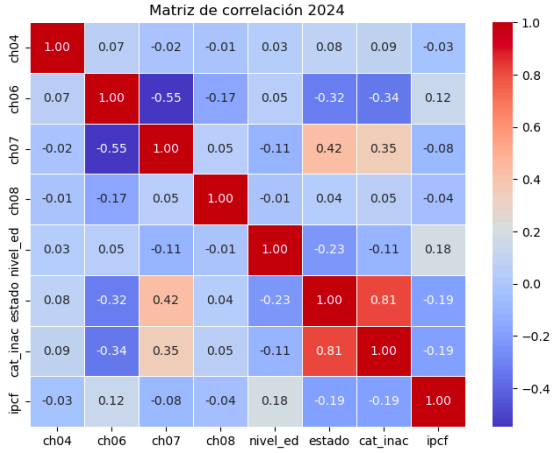

Figura 3. Matriz de correlación entre variables en 2024.

Para finalizar, contamos la cantidad de personas desocupadas e inactivas, según lo que señalaba su condición de actividad. La cantidad de desocupados en la muestra eran 809, mientras que la cantidad de inactivos en la muestra eran 5.249.

EJERCICIO 3

Con el fin de evaluar la cantidad de personas que no respondieron la pregunta sobre condición de actividad (es decir, donde el estado del participante es igual a 0) separamos la base de datos en 2

bases separadas: una base con los que sí respondieron y otra base con los que no respondieron. En la base ‘no respondieron’ obtuvimos un total de 10 participantes.

EJERCICIO 4

En la base de datos ‘respondieron’ creada en el ejercicio anterior, agregamos una columna llamada PEA que distingue la Población Económicamente Activa, tomando el valor 1 si el participante se encuentra en estado ‘ocupados’ o ‘desocupados’ y tomando el valor 0 si se encuentra en estado ‘inactivo’ o ‘menor de 10 años’. Luego filtramos las observaciones según el año. Creamos un gráfico de barras para comparar PEI (Población Económicamente Inactiva) y PEA (Población Económicamente Activa) en ambos períodos (ver **Figura 5**).

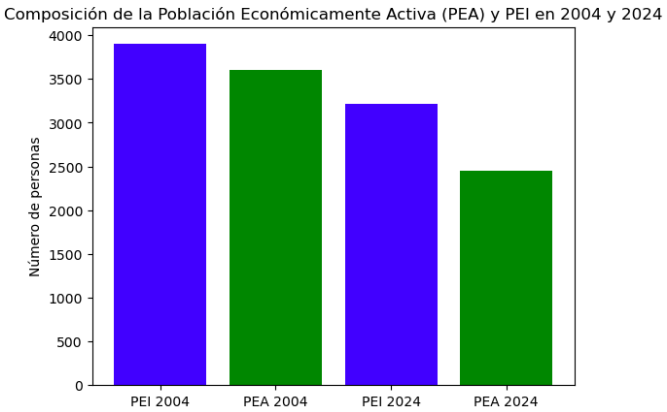

Figura 5. Gráfico de barras de Población Económicamente Activa e Inactiva en 2004 y en 2024.

Este gráfico de barras ilustra que en cada período la población económicamente inactiva supera a la población económicamente activa. Además, observamos que en 2004 la diferencia entre la población económicamente activa y la inactiva es menor que en 2024. Durante el 2024 se encuestaron una menor cantidad de sujetos.

EJERCICIO 5

A continuación, en la base de datos ‘respondieron’ agregamos una columna llamada PET (Población en Edad para Trabajar), que tomaría el valor 1 si la persona tiene entre 15 y 65 años cumplidos, y de lo contrario, tome el valor 0. Filtramos los datos según el año en el que fue tomada la observación. Realizamos un gráfico de barras mostrando la composición de PET durante 2004 y durante 2024. Nombramos ‘SET’ a la población Sin Edad para Trabajar (es decir, cuando la edad del participante era menor a 15 o mayor a 65). El gráfico (ver **Figura 6**) evidencia que en ambos períodos hay menos personas sin la edad correspondiente para trabajar, ya que las barras rojas que indican SET son significativamente más bajas que las barras amarillas. Tanto en 2004 como en 2024, las diferencias entre PET y SET parecen ser similares, sugiriendo que la proporción entre ambas

condiciones se mantuvo relativamente estable. Nuevamente se evidencia que en la base correspondiente al año 2024 hay menos datos.

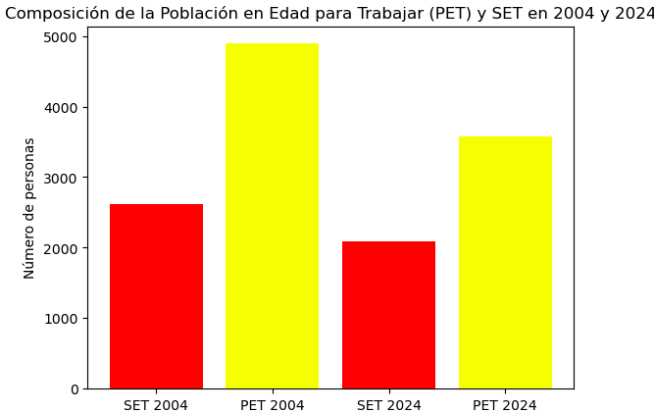

Figura 6. Gráfico de barras que ilustra la composición de PET en 2004 y en 2024.

EJERCICIO 6

Agregamos una columna llamada ‘desocupado’ basadas en la variable ‘estado’, que toma el valor 1 si el participante está desocupado, y el valor 0 de lo contrario. Obtuvimos 2 resultados muy distintos en ambos períodos. Para el año 2004, había una cantidad de 528 personas desocupadas. Para el año 2024, había una cantidad de 281 personas desocupadas. Es decir, la cantidad de personas desocupadas se redujo casi a la mitad en el segundo período.

PARTE II: Clasificación

EJERCICIO 1

En primer lugar, eliminamos variables insignificantes y aquellas que estaban en formato ‘string’. Además, reemplazamos los valores ‘NaN’ restantes por la mediana de cada variable. Una vez realizada esta limpieza, procedimos a manipular las nuevas bases.

En este ejercicio realizamos una división de los datos en conjuntos de entrenamiento (*train*) y prueba (*test*) haciendo uso del comando ‘train_test_split’ de la librería ‘sklearn’. Trabajamos con los datos correspondientes a los años 2004 y 2024, filtrando la base ‘respondieron’ para cada año en particular. Denominamos ‘desocupado’ como la variable dependiente (Y), mientras que las variables independientes fueron todas las demás, a las que se les añadió una columna de unos. Añadimos la columna de unos debido a que en ciertos modelos estadísticos es importante para representar el intercepto o un término constante, es decir, el valor que toma la variable dependiente cuando todas las variables independientes son cero. Posteriormente, dividimos los datos: un 70% de ellos fueron utilizados para *train* y un 30% para *test*, utilizando una semilla fija de aleatorización

('random_state=101'). Finalmente, imprimimos los tamaños de los conjuntos de datos, verificando la cantidad de observaciones en cada uno. Ver **Figura 7** y **Figura 8** a continuación.

```
Año 2004:
Tamaño de X_train: (5254, 140)
Tamaño de X_test: (2252, 140)
Tamaño de y_train: (5254,)
Tamaño de y_test: (2252,)
```

Figura 7. Datos sobre la cantidad de observaciones que hay en las bases *train* y *test* en 2004.

```
Año 2024:
Tamaño de X_train: (3966, 140)
Tamaño de X_test: (1701, 140)
Tamaño de y_train: (3966,)
Tamaño de y_test: (1701,)
```

Figura 8. Datos sobre la cantidad de observaciones que hay en las bases *train* y *test* en 2024.

El valor '140' correspondiente a la variable X tanto en *train* como en *test* indica la cantidad de variables regresoras/predictoras. Este indicador se ausenta para la variable estudiada ya que es únicamente una, la variable dependiente.

EJERCICIO 2

En primer lugar, implementamos una Regresión Logística haciendo uso nuevamente de la librería 'sklearn'. Planteamos el modelo y lo entrenamos, para luego hacer las predicciones en base a los datos de prueba (*test*) y reportar la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy para cada período. Los datos de salida del año 2004 se ilustran en la **Figura 9**. A través de la matriz de confusión, visualizamos el rendimiento de un algoritmo de clasificación. Reportamos 2.097 verdaderos positivos (predicciones correctas de clase "no desocupado") y 153 verdaderos negativos (predicciones correctas de la clase "desocupado"). No hay ningún caso incorrectamente clasificado como "desocupado", y hay solamente 2 falsos negativos (casos que fueron incorrectamente clasificados como "no desocupado"). Por lo tanto, casi todos los datos fueron clasificados correctamente, a excepción de 2. En cuanto al valor de *accuracy*, reportamos un 99.91%, lo que significa que el modelo está clasificando correctamente el 99.91% de las observaciones. Esto se alinea con lo evidenciado en la matriz de confusión, ya que el valor de *accuracy* es el cociente entre los datos correctamente predichos y la totalidad de los datos. En cuanto al AUC (Area Under the Curve), reportamos un valor de 0.9965, lo cual es muy bueno ya que queremos que sea lo más cercano a 1 posible. Esto indica que el modelo es casi perfecto en distinguir entre las clases "desocupado" y "no desocupado". Por último, en la curva ROC podemos observar la relación entre la Tasa de Verdaderos Positivos y la Tasa de Falsos Positivos. Al ver que la curva casi toca los puntos de la esquina superior izquierda concluimos que es casi perfecta, dado que esto implica una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.

En lo que respecta al método de Regresión Logística para el año 2024, observamos los siguientes indicadores expuestos en la **Figura 10**. En este caso la cantidad de verdaderos positivos son 1.603 y la cantidad de verdaderos negativos, 13. Las predicciones incorrectas para este período son más que para el 2004: 4 falsos positivos y 81 falsos negativos. Consecuentemente, el porcentaje de

accuracy es menor, de un 95%. No obstante, el indicador demuestra que el modelo está haciendo predicciones muy precisas. Por otro lado, el valor AUC reportado es de 0,94 aproximadamente, ilustrando que el modelo es muy bueno ya que un valor de 1 representaría una perfecta capacidad de discriminación. La curva ROC también enfatiza que el modelo tiene un buen rendimiento ya que ilustra su capacidad de mantener una alta tasa de verdaderos positivos sin incrementar mucho los falsos positivos.

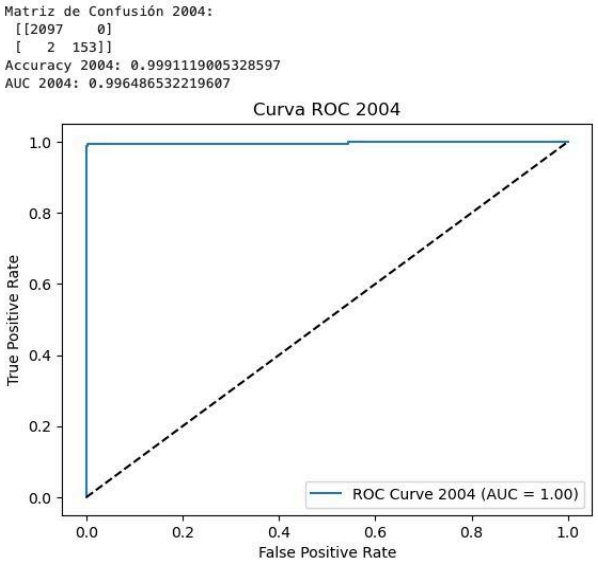

Figura 9. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2004.

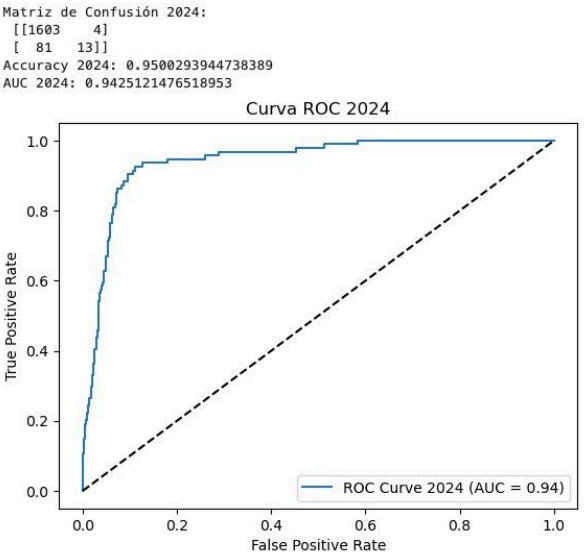

Figura 10. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2024.

En segundo lugar, implementamos el Análisis Discriminante Lineal (LDA), una técnica de clasificación supervisada que se utiliza para encontrar una combinación lineal de características que mejor separen dos o más clases. La **Figura 11** muestra los resultados de un modelo de Análisis Discriminante Lineal (LDA) del año 2004. La matriz de confusión revela una clasificación perfecta, con 2097 verdaderos negativos y 155 verdaderos positivos, sin ningún error (ni falsos positivos ni falsos negativos). El modelo tiene una exactitud (accuracy) del 100%, lo que significa que todas las predicciones fueron correctas. Además, la curva ROC indica un área bajo la curva (AUC) de 1.0, lo que refleja un rendimiento ideal al separar las clases, ya que el modelo distingue perfectamente entre las clases positivas y negativas sin cometer errores de clasificación.

Por otro lado, en la **Figura 12** se observan los resultados obtenidos del modelo LDA del año 2024. Los datos demuestran que la matriz de confusión de este modelo tiene una clasificación precisa, con 1607 verdaderos negativos y 93 verdaderos positivos, y solo un error con un falso positivo. Esto sugiere un rendimiento casi perfecto del modelo. El accuracy es cercano al 99.94%, lo que indica una excelente capacidad de predicción. La curva ROC muestra un área bajo la curva (AUC) de 0.99, lo que confirma la elevada precisión del modelo al diferenciar entre clases. Este resultado es muy positivo y sugiere un desempeño óptimo del modelo LDA para este conjunto de datos.

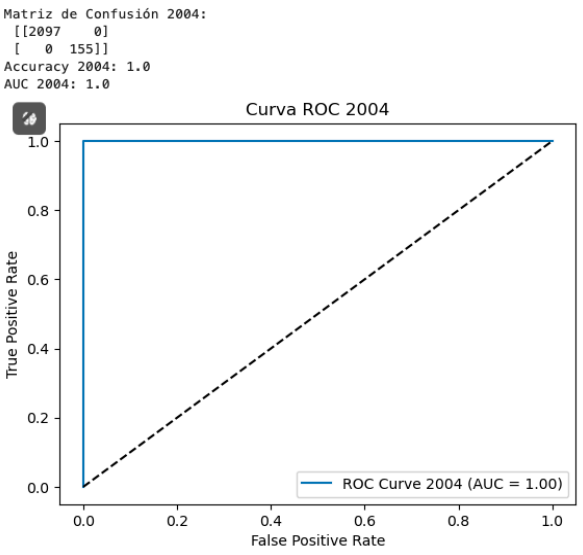

Figura 11. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2004.

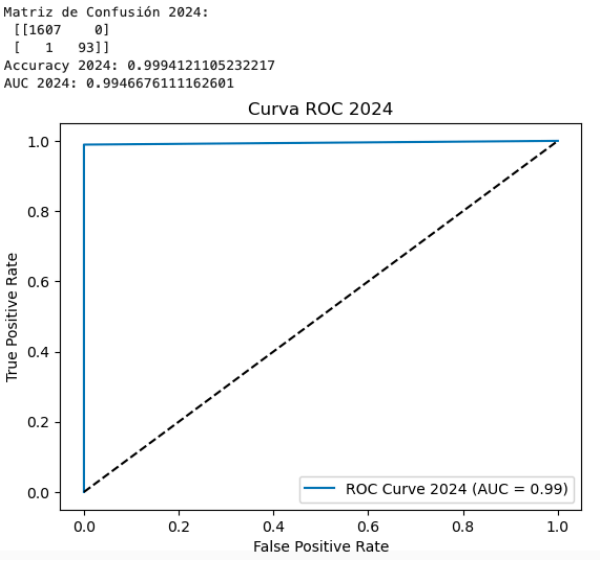

Figura 12. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2024.

Por otro lado, realizamos el modelo K-Nearest Neighbors (KNN) para predecir si una persona está desocupada o no, usando variables de características individuales. El modelo se entrena con un conjunto de datos de entrenamiento (X_train y Y_train) con un parámetro k=3, es decir, se consideran los 3 vecinos más cercanos para la predicción. Luego de entrenar el modelo, se realizan predicciones con los datos de prueba sobre el conjunto de prueba (X_test). Se generan tanto las predicciones de clase (y_pred_knn) como las probabilidades de que una observación pertenezca a la clase positiva (desocupado), lo que se guarda en 'y_prob_knn'. Luego, el código evalúa el desempeño del modelo generando una matriz de confusión para medir la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

La **Figura 13** presenta todos los datos del modelo en 2004. Expone que la matriz de confusión predijo correctamente 2048 personas que no están desocupadas (verdaderos negativos) y 15 personas que están desocupadas (verdaderos positivos). Sin embargo, cometió 49 falsos positivos (personas clasificadas como desocupadas pero que no lo están) y 140 falsos negativos (personas clasificadas como no desocupadas, pero que sí lo están). Los datos de accuracy indican que el 91.61% de las predicciones realizadas por el modelo fueron correctas, lo que refleja un buen rendimiento general. Sin embargo, AUC (Área Bajo la Curva ROC) fue de 0,64 lo que indica que el modelo tiene una capacidad de discriminación entre clases moderada, lo cual es mejor que el azar, pero todavía puede mejorar. En cuanto al gráfico de la Curva ROC que es una representación visual del desempeño del modelo. En el gráfico la línea discontinua representa una predicción aleatoria (AUC=0.5) y la curva azul muestra el desempeño real del modelo. Aunque la azul está por encima de la línea de predicción aleatoria, esta no alcanza una discriminación óptima, lo que coincide con el AUC de 0.64.

A su vez, realizamos el mismo modelo de KNN con k = 3 pero para el año 2024. La **Figura 14** muestra en la matriz de confusión que el modelo predijo correctamente 1592 personas no desocupadas (verdaderos negativos) y 111 personas desocupadas (verdaderos positivos), con 15 falsos positivos y 83 falsos negativos. El accuracy es de 94.24%, lo que indica que el modelo clasifica correctamente la mayoría de los casos. Sin embargo, el AUC es de 0.6657, lo que refleja una capacidad moderada del modelo para distinguir entre desocupados y no desocupados. En La curva ROC, la curva del modelo está por encima de la línea de predicción aleatoria, pero aún faltaría una mejora en la identificación de personas desocupadas. Aunque el rendimiento general es bueno, con un accuracy alto, se observa que el modelo tiene dificultades para identificar correctamente a todas las personas desocupadas, como lo indica el AUC y los falsos negativos en la matriz de confusión.

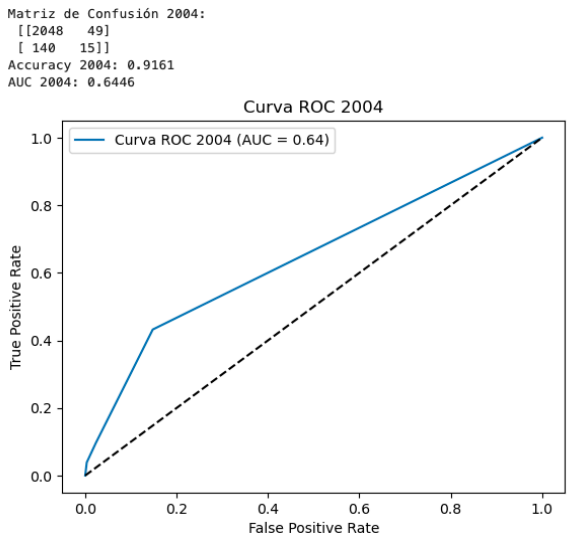

Figura 13. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2004.

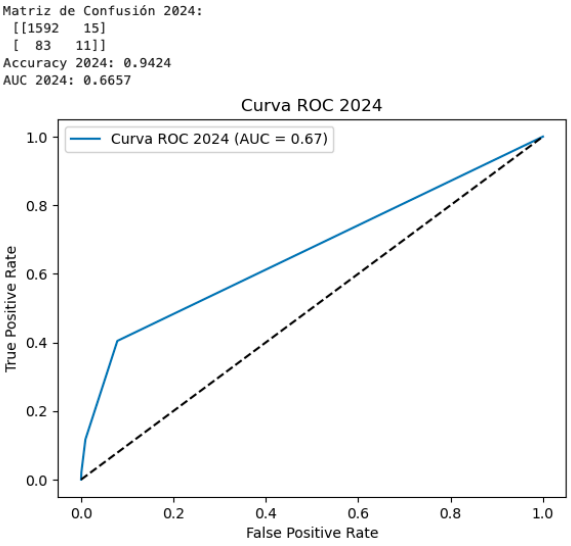

Figura 14. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2024.

Posteriormente, se implementó un modelo Naive Bayes Gaussiano para predecir si una persona está desocupada o no en los diferentes años. El modelo se entrena utilizando un conjunto de datos de entrenamiento. El modelo se entrena con un conjunto de datos de entrenamiento y luego realiza predicciones sobre los datos de prueba. Durante esta fase, el clasificador aprende las probabilidades condicionales de cada característica dada la clase (ocupado/desocupado). Es decir, calcula las probabilidades de que una persona esté desocupada o no, dada una serie de variables predictoras.

En la **Figura 15** se muestran los datos resultados de este modelo en el año 2004. La matriz de confusión muestra que el modelo predijo correctamente 2097 personas no desocupadas (verdaderos negativos) y 154 personas desocupadas (verdaderos positivos), con solo 1 falso positivo y 0 falsos negativos, lo que indica un excelente rendimiento. El accuracy del modelo es de 99.96%, lo que implica que casi todas las predicciones fueron correctas. Además, el AUC es de 0.997, lo que sugiere una excelente capacidad del modelo para distinguir entre personas desocupadas y no desocupadas. La

curva ROC muestra una línea casi perfecta, con una tasa de verdaderos positivos muy alta y una tasa de falsos positivos muy baja, lo que confirma el excelente desempeño del modelo en la clasificación de estos casos.

En la **Figura 16** se muestran los resultados del modelo de Naive Bayes con los datos del año 2024. La matriz de confusión muestra una clasificación de 725 verdaderos negativos y 93 verdaderos positivos, junto con un solo falso positivo. La accuracy del modelo es de aproximadamente 0.48, lo que indica que el modelo clasifica correctamente casi el 48% de las veces. En cuanto a la curva ROC, el valor del AUC es de 0.73, lo que refleja una capacidad razonable para diferenciar entre clases positivas y negativas.

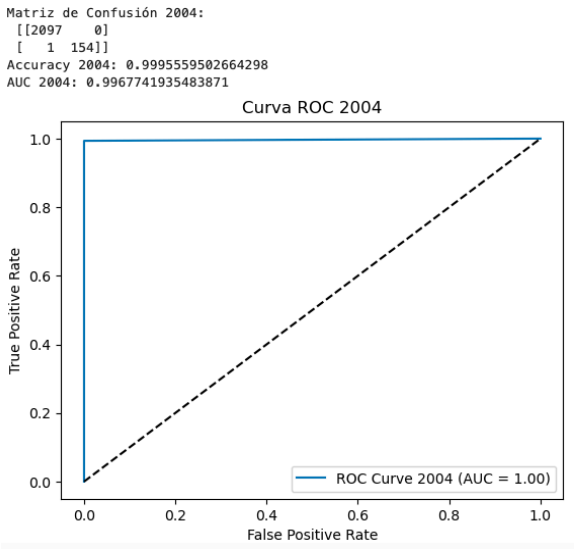

Figura 15. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2004.

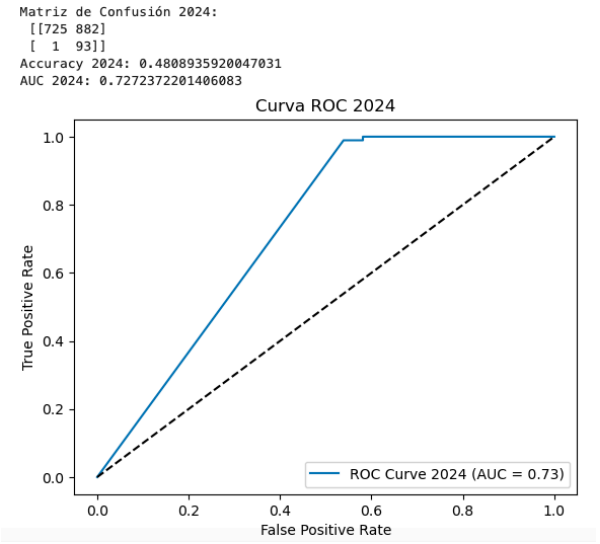

Figura 16. Matriz de confusión, valores de *accuracy* y AUC y curva ROC para el año 2024.

EJERCICIO 3

En el análisis de los modelos para los años 2004 y 2024, se observa que en 2004 (**Figura 17**) el modelo de Análisis Discriminante Lineal (LDA) se destaca como el mejor, alcanzando una precisión (accuracy) y una capacidad de discriminación (AUC) perfectas, con valores de 1.000 en ambas métricas. Esto indica que los datos de 2004 están bien representados en el espacio lineal que LDA utiliza, permitiendo clasificar con precisión todos los casos. En comparación, la Regresión Logística también muestra un alto rendimiento en 2004, con una precisión de 0.9991 y un AUC de 0.9965, aunque ligeramente inferior al de LDA. El modelo de Naive Bayes tiene un desempeño casi perfecto con una precisión de 0.9996 y un AUC de 0.9968, mientras que K-Nearest Neighbors (KNN) con k=3 muestra un rendimiento considerablemente menor, con una precisión de 0.9161 y un AUC de 0.6446, lo que indica que KNN no captura bien la estructura de los datos en ese año.

Por otro lado, en el análisis de los resultados para el año 2024 (**Figura 18**), LDA nuevamente se destaca, alcanzando una precisión de 0.9994 y un AUC de 0.9947, lo que lo convierte en el mejor

modelo también para este año, aunque su rendimiento es ligeramente inferior al del 2004. La Regresión Logística sigue siendo bastante precisa en 2024, con una precisión de 0.9500 y un AUC de 0.9425, aunque es superada por LDA en ambas métricas. Por su parte, KNN mejora levemente en términos de accuracy con 0.9424 y de AUC, alcanzando 0.6657, pero sigue siendo el de menor rendimiento en comparación con los demás métodos. Finalmente, el modelo de Naive Bayes muestra una disminución significativa en su precisión en 2024, con un valor de 0.4809, aunque su AUC es relativamente aceptable con 0.7272.

En conclusión, LDA es el modelo que mejor predice en ambos años, demostrando la mejor combinación de precisión y capacidad de discriminación en cada conjunto de datos. Los cambios en el rendimiento de Naive Bayes y KNN entre 2004 y 2024 podrían reflejar diferencias en la estructura de los datos, lo que afecta su capacidad de adaptación y precisión en cada año.

'Resultados 2004:'

	Modelo	Accuracy	AUC
0	Logistic Regression	0.999112	0.996487
1	LDA	1.000000	1.000000
2	KNN (k=3)	0.916075	0.644608
3	Naive Bayes	0.999556	0.996774

Figura 17: Performance modelos 2004

'Resultados 2024:'

	Modelo	Accuracy	AUC
0	Logistic Regression	0.950029	0.942512
1	LDA	0.999412	0.994668
2	KNN (k=3)	0.942387	0.665688
3	Naive Bayes	0.480894	0.727237

Figura 18: Performance modelos 2024

EJERCICIO 4

En este ejercicio, el objetivo fue identificar qué personas en el conjunto de datos denominado "norespondieron" son desocupadas, utilizando el Análisis Discriminante Lineal (LDA) previamente entrenado. Seleccionamos este modelo (lda_2004), ya que se destacó como el mejor en las instancias previas.

Al calcular la proporción de personas desocupadas en este grupo, el resultado fue igual a cero, con un array de predicciones que consistía únicamente en ceros. Esto indica que el modelo no identificó a ningún individuo como desocupado dentro del conjunto analizado. Consideramos que esta es una limitación significativa del análisis, la cual atribuimos a la escasa cantidad de observaciones en la base "norespondieron", que contiene solo 10 registros en total. Este bajo número de observaciones probablemente impidió que el modelo pudiera capturar la variabilidad necesaria para realizar predicciones significativas.