Universidad de SanAndrés

Tercer Trabajo Práctico

Informe

Alumnas:

Angela Navajas McCormick

Legajo: 33542

Tiziana Sol Paciente

Legajo: 33377

Materia:

Ciencia de Datos

Lic. en Ciencias del Comportamiento

Docentes:

Maria Noelia Romero

Ignacio Spiousas

Fecha de entrega:

3 de Noviembre, 2024

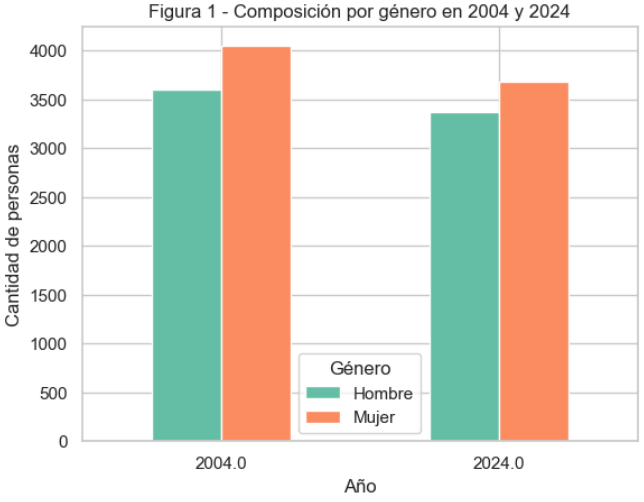
Parte 1: Analizando la base

En este trabajo se utilizarán los datos obtenidos a través de la Encuesta Permanente de Hogares (EPH), particularmente nos centraremos en la tasa de desocupación. El INDEC identifica a las personas desocupadas como individuos que, si bien no tienen ocupación, buscan activamente trabajar y se encuentran disponibles para hacerlo. Este indicador no incluye otras formas de precariedad laboral.

Como primera parte del trabajo, se unificaron las bases correspondientes al primer trimestre de 2004 y 2024 y se hicieron las modificaciones necesarias a los nombre de las variables y formatos de los datos para poder unificar correctamente las bases y continuar con el análisis. Después se eliminaron las observaciones que no corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires, obteniendo así una base con 14698 observaciones. Luego se hizo una pequeña revisión de los valores en las variables para verificar que tomen valores con sentido, se verificó que las edades e ingresos tengan valores positivos ya que no pueden tomar un valor negativos. Sumado a esto, se verificó que todos las variables tomen de valores que fueron propuestos como etiquetas en diccionario de variables del INDEC.

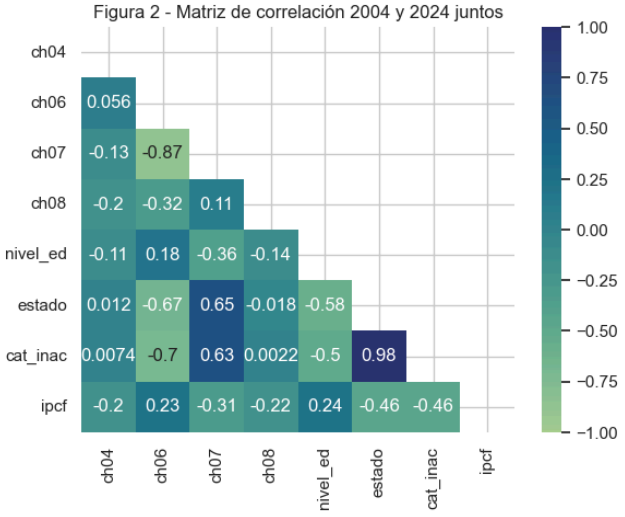
Para ver la composición por género para cada año, se realizó un gráfico de barras (Figura 1), en el cual podemos observar que en ambos años, la cantidad de mujeres que respondieron la encuesta fue mayor a la cantidad de hombres. Por otro lado, podemos ver que tanto para hombres como para mujeres, la cantidad de personas que respondieron la encuesta disminuyó en el año 2024.

Figura 1 - Composición por género en 2004 y 2024



Como siguiente paso, se elaboró una matriz de correlación considerando las siguientes variables: *ch04* (sexo), *ch06* (años), *ch07* (estado civil), *ch08* (cobertura médica), *nivel_ed* (nivel educativo), *estado* (condición de actividad), *cat_inac* (categoría de inactividad) e *ipcf* (monto de ingreso per cápita familiar percibido en el mes de referencia). Primero se realizó con los años 2004 y 2024 en conjunto para ver el comportamiento general entre variables, como se muestra en la Figura 2, la correlación positiva más fuerte (0.98) se encuentra entre las variables de *estado* e *inactividad*. La correlación negativa más fuerte es entre el *estado civil* y la *edad* (-0.87). Además, la *edad* se correlaciona negativamente con *estado* e *inactividad*, con valores de -0.67 y -0.7, respectivamente. Por otro lado, el *estado civil* presenta una correlación positiva con estas variables, con un valor de 0.68 para *estado* y 0.64 para la de *inactividad*.

Figura 2 - Matriz de correlación 2004 y 2024 juntos



Sumado a este análisis, se observó cómo se relacionaban las variables en cada año por separado con la finalidad de comparar su comportamiento. En la Figura 2.1 se ven las correlaciones de las variables de 2004 y en la Figura 2.2 de las variables de 2024. A modo de comparación, se

muestra que las variables se comportan similar en la mayoría de las correlaciones entre años, sin embargo hay algunas diferencias en la correlación de *cobertura médica* y *años* que en 2004 toma un valor de -0.18 y en 2024 de -0.32, entre *nivel educativo* y *años* la relación pasa de ser negativa a positiva ya que en 2004 tiene un valor de -0.35 y en 2024 de 0.18, entre *nivel educativo* y *estado civil* que en 2004 la relación era de 0.13 y en 2024 la relación es negativa de -0.36, entre *estado* y *nivel educativo* la relación cambió de ser en 2004 de -0.033 a ser de -0.58 en 2024, por último entre las variables *categoría inactivo* y *nivel educativo* la relacion tambien cambio de tomar en 2004 un valor de 0.06 y en 2024 -0.5.

Figura 2.1 - Matríz de correlación 2004

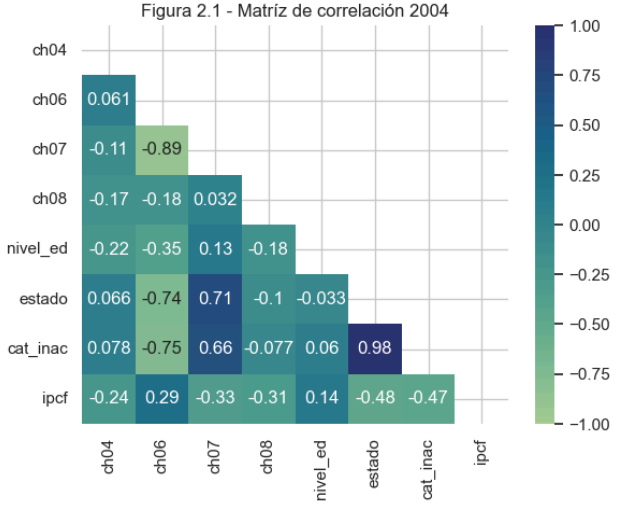
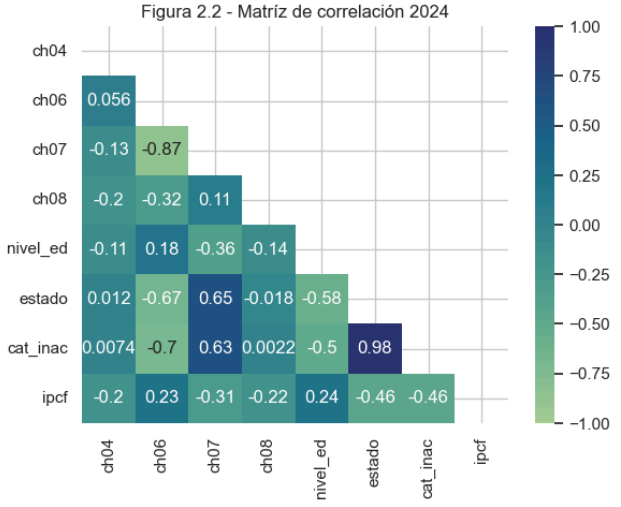


Figura 2.2 - Matríz de correlación 2024

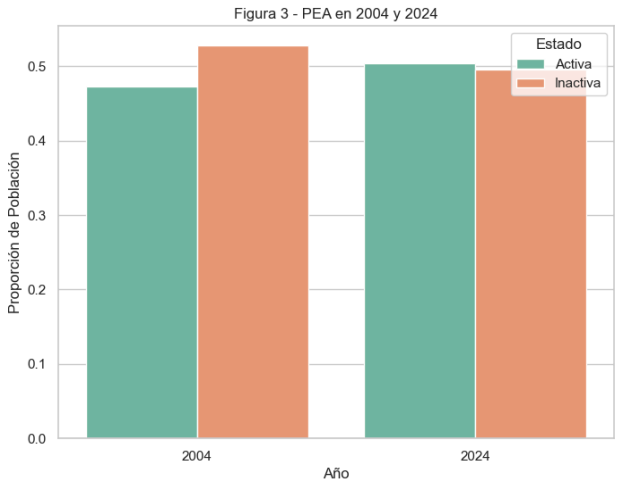


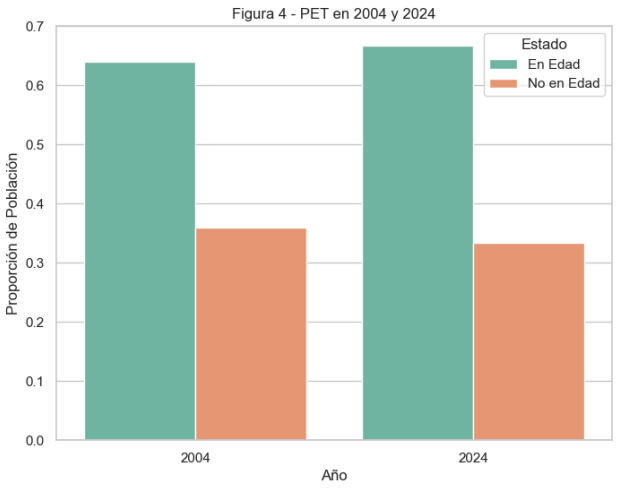
Se realizó el contabilizó la cantidad de desocupados e inactivos que poseía la muestra, esta está compuesta por 839 desocupados y 5,462 inactivos. Para facilitar la comparación e interpretación del cambio en los ingresos de cada año, se ajustó la media de los salarios utilizando una calculadora de inflación histórica de Argentina. En la Tabla 1, se observa que el ingreso ajustado de los desocupados en 2004 fue de $104,573 y en 2024 fue de $85,019, lo que representa una clara disminución en el IPCF de las personas desocupadas. De manera similar, los ocupados en 2004 tenían un ingreso ajustado de $220,351, superior al ingreso promedio de 2024, que fue de $207,644. Por último, el ingreso ajustado de los inactivos en 2004 fue de $147,056, siendo también mayor que los ingresos de 2024 que alcanzaron los $130,704.

Tabla 1 - Media de ingreso per capita familiar según estado por año:

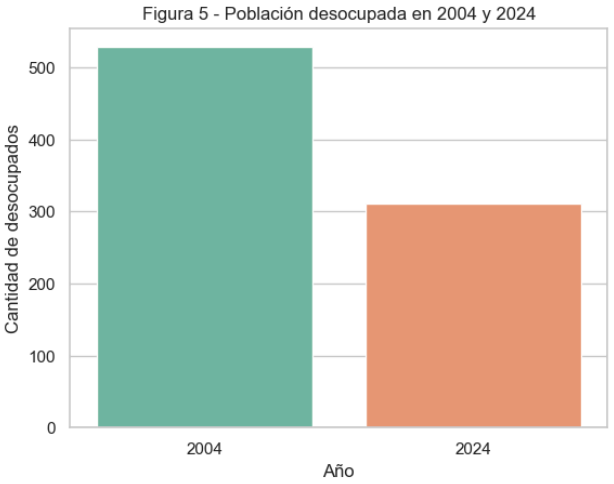
estado	ano4	
Desocupado	2004.0	224.23
	2024.0	85019.15
Entrevista no realizada	2004.0	52.53
	2024.0	0.00
Inactivo	2004.0	315.89
	2024.0	130704.60
Menor de 10 años	2004.0	246.26
	2024.0	104353.66
Ocupado	2004.0	476.06
	2024.0	207644.84

Para analizar la falta de respuestas del ítem sobre los ingresos de las personas, se dividió la base de datos en dos grupos: aquellos que lo respondieron y los que no respondieron. De esta manera, se obtuvieron dos conjuntos de datos: uno con 51 respuestas en "*no_resp*" y otro con 14,647 respuestas en "*respondieron*". Posteriormente, se añadieron dos columnas llamadas *PEA* y *PET*, que permiten identificar a la población económicamente activa y a la población en edad para trabajar, respectivamente. En la Figura 3 se observa un aumento de 0.3 puntos porcentuales en la población económicamente activa en 2024 en comparación con 2004. Del mismo modo, la Figura 4 muestra que la población en edad para trabajar también incrementó en 2024, con un aumento de 0.2 puntos porcentuales. Al comparar las categorías, se puede apreciar claramente que, en ambos años, no toda la población en edad para trabajar está económicamente activa.



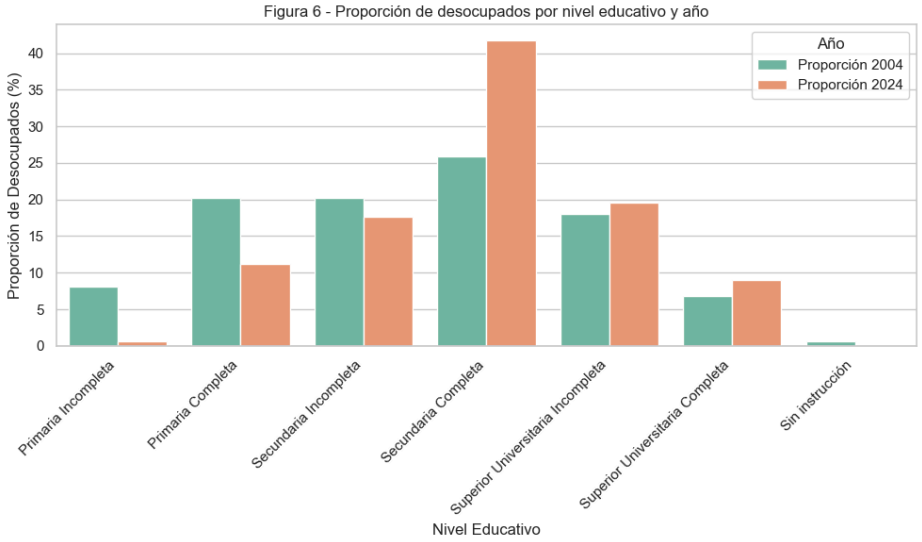


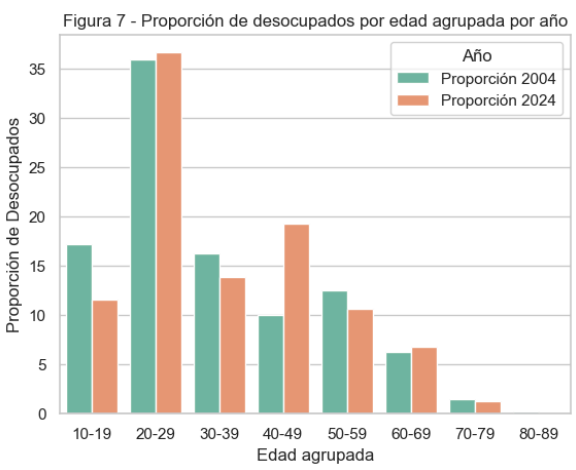
Se calculó el total de personas desocupadas en cada año. Como se puede ver en la Figura 5, el total de desocupados en 2004 fue de 528 mientras que en 2024 fue de 311.



Cuando dividimos la proporción de desocupados por nivel educativo (Figura 6) se puede ver que en 2024, las proporciones de desocupados con niveles de primaria completa y secundaria completa disminuyeron en comparación con 2004, lo mismo que la proporción de desocupados con secundaria incompleta. Sin embargo, el grupo con educación superior universitaria incompleta mostró un aumento en la proporción de desocupación, siendo esta una de las categorías con mayor cambio. Además, el porcentaje de desocupados sin instrucción formal se mantuvo bajo en ambos años, aunque con una ligera disminución en 2024.

Por último, se formaron grupos etarios de diez años para observar los cambios en la desocupación con respecto a la edad (Figura 7). En el grupo de 20 a 29 años, la proporción de desocupados aumentó en 2024 respecto a 2004, y sigue siendo el grupo con mayor desocupación en ambos años. En los grupos de mayor edad (40 a 49, 60 a 69 años), la proporción de desocupados aumentó en 2024 en comparación con 2004. En los grupos de 30 a 39 y 50 a 59 se observa una ligera disminución en 2024. De la





misma forma, los grupos de edad más avanzados, de 70 a 79, mantienen proporciones bajas en ambos años, aunque se nota un leve incremento en 2024.

Parte II: Clasificación

Como primer paso hacia el objetivo de predecir si una persona está desocupada o no, se generó una base llamada *'base_clasificacion'* que contiene una selección de variables que se consideraron relevantes para la predicción. Estas variables son: *'ano4'*, *'ch04'*, *'ch06'*, *'ch07'*, *'ch08'*, *'nivel_ed'*, *'estado'*, *'cat_inac'*, *'componente'*, *'h15'*, *'mas_500'*, *'aglomerado'* y *'ch03'*. Se prepararon los conjuntos de datos de entrenamiento y testeo para los años 2004 y 2024. Para facilitar este proceso se desarrolló una función llamada *'preparar_datos'*, que comienza filtrando las observaciones por el año correspondiente, divide los subconjuntos usando la funcion 'train_test_split', asignando un 70% de la muestra a la base entrenamiento y un 30% al testeo y un valor semilla de 101 con 'random_state'. Se estableció el valor "desocupado" dentro de la variable *estado* como la variable dependiente y el resto de las variables de *base_clasificacion* fueron tomadas como parte de la variable independiente. También se incorporó una columna de intercepto en ambos conjuntos de datos (X_test = X_test.assign(intercept=1; X_train = X_train.assign(intercept=1). Como resultado, se obtuvieron cuatro conjuntos de datos (`X_train`, `y_train`, `X_test` y `y_test`) para cada uno de los años analizados.

Se implementaron 4 modelos para cada año para así evaluar cuál era el mejor predictor. Comenzando por 2004, se realizó una regresión logística que permite modelar la probabilidad de que Y pertenezca a la categoría desocupado. La matriz de confusión indica que el modelo predijo correctamente 2128 casos negativos y un caso positivo, por otro lado cometió 5 errores al clasificar negativos como positivos y 159 al clasificar positivos como negativos. El modelo entonces, demuestra una alta precisión para clasificar casos negativos y más dificultades para reportar los casos positivos. Se obtuvo un AUC de 0.8692, teniendo en cuenta que el máximo es 1, este resultado indicaría un buen desempeño en la clasificación de las observaciones y por lo tanto, la probabilidad de que el modelo clasifique correctamente a los desocupados es de razonablemente alta. Este resultado, se puede visualizar en la curva de ROC graficada en la Figura 8. La accuracy del modelo fue de 92.88%.

Se implementó un análisis discriminante lineal (Figura 9). Este modelo obtuvo un AUC de 0.8534 lo cual demostraría una capacidad razonable para discriminar entre clases, sin embargo, de forma similar que con la regresión logística, la matriz de confusión muestra 2129 verdaderos negativos y 1 verdadero positivo, junto con 4 falsos positivos y 158 falsos negativos. Esto demuestra que el modelo tiene un desempeño sólido en la identificación de la clase negativa,

Figura 8 - Curva ROC de LOGIT - 2004

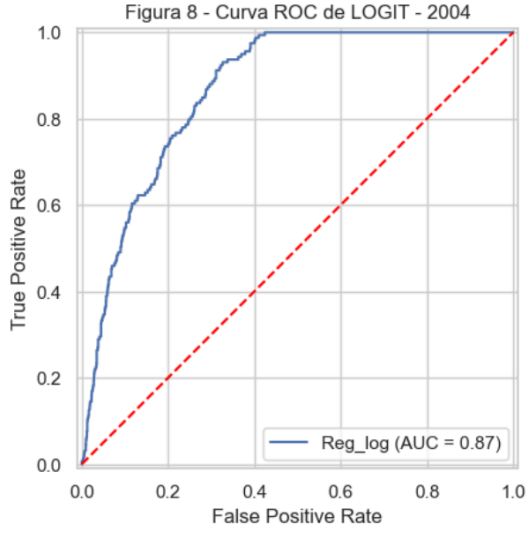
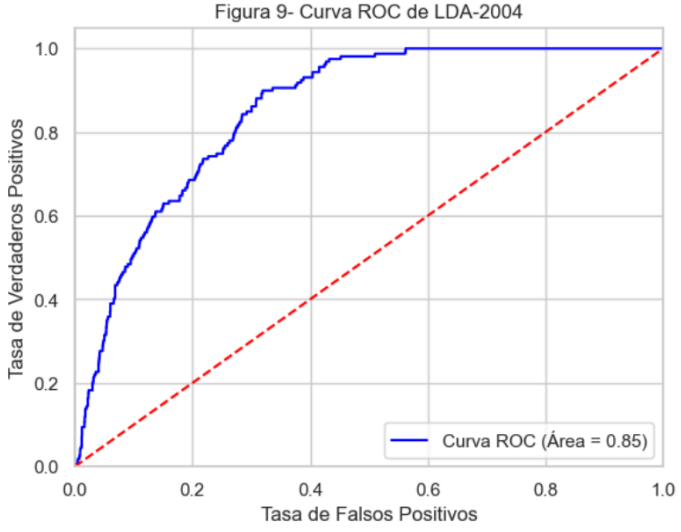


Figura 9- Curva ROC de LDA-2004



aunque enfrenta desafíos al clasificar los casos positivos. La exactitud global del modelo fue de 93%, reflejando su precisión en la clasificación general.

En tanto al análisis realizado con K-vecinos cercanos (KNN). Este modelo demostró un AUC de 0.7067, demostrando un buen desempeño en la clasificación de las observaciones en clases; en la matriz de confusión, se observaron 2067 verdaderos positivos, 66 falsos positivos, 128 falsos negativos y 31 verdaderos negativos, demostrando que identifica correctamente los positivos, pero es problemático al identificar los negativos. La exactitud (accuracy) del método es de 0.915, por lo que se refleja una buena precisión en la clasificación general. Este resultado, se puede visualizar en la curva de ROC graficada en la Figura 10.

Con el método de Naive Bayes, se obtuvo un valor de AUC de 0.79 demostrando que el desempeño de la clasificación es bueno. En tanto a la matriz de confusión, se observaron 1224 verdaderos positivos, 909 falsos positivos, 0 falsos negativos y 159 verdaderos negativos, esto hace notar la dificultad del modelo en identificar correctamente a los positivos y no demuestra problemas en identificar correctamente los negativos. El valor de accuracy que toma es de 0.603, evidenciando que la clasificación que genera este método no es la más exacta.

Figura 10 - Curva ROC de KNN - 2004

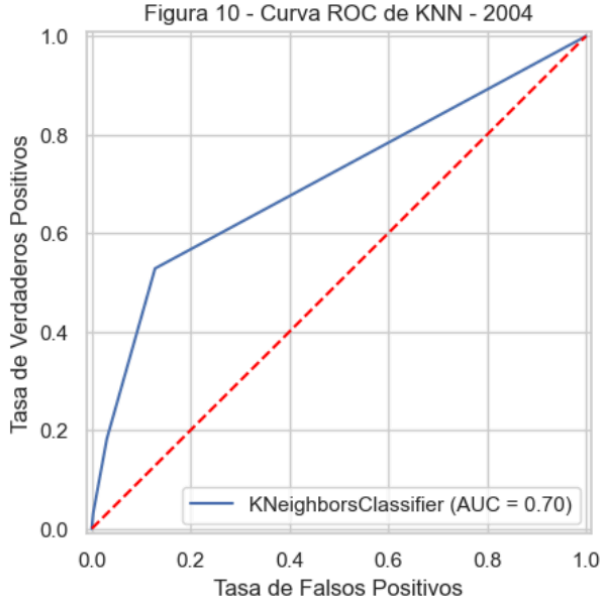
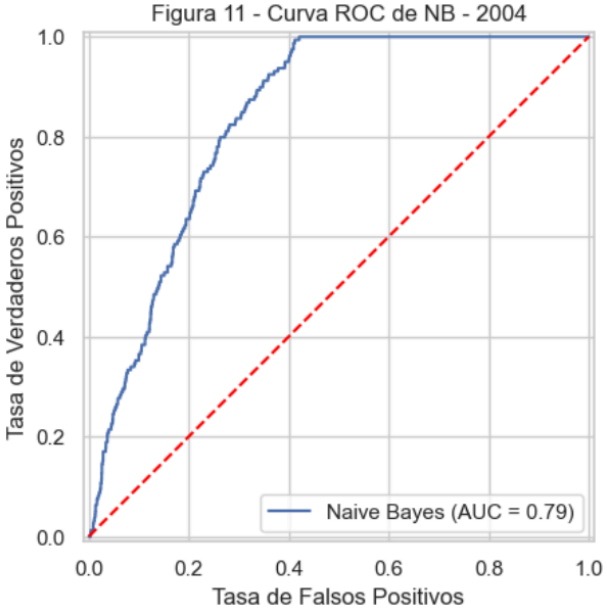


Figura 11 - Curva ROC de NB - 2004



En el análisis de 2024, se realizó una regresión logística (Figura 12) con la que se obtuvo un AUC de 0.8638 que visibiliza la buena manera de la herramienta en clasificar las observaciones en casos. En la matriz de confusión, se obtuvieron 1995 verdaderos positivos, 1 falso positivo, 107 falsos negativos y 0 verdadero negativo, demostrando un buen desempeño en identificar positivos, pero mal rendimiento en identificar correctamente los negativos. Por último, el nivel de exactitud del modelo es de 0.9486, por lo que se puede determinar que este modelo tiene buena precisión en la clasificación general.

Seguido de esto, en el análisis discriminante lineal (Figura 13) se obtuvo un AUC de 0.8375, demostrando tener un buen desempeño en la clasificación de observaciones, en la matriz de confusión se identificaron 1994 verdaderos positivos, 2 falsos positivos, 107 falsos negativos y 0 verdaderos negativos, lo que nuevamente demuestra que el modelo es bueno identificando correctamente las observaciones positivas y no las negativas. Por último, el nivel de exactitud de clasificación general del modelo es de 0.95, siendo este un número muy cercano a 0 y por ende, haciendo que el modelo sea un buen clasificador.

Figura 12 - Curva ROC de LOGIT - 2024

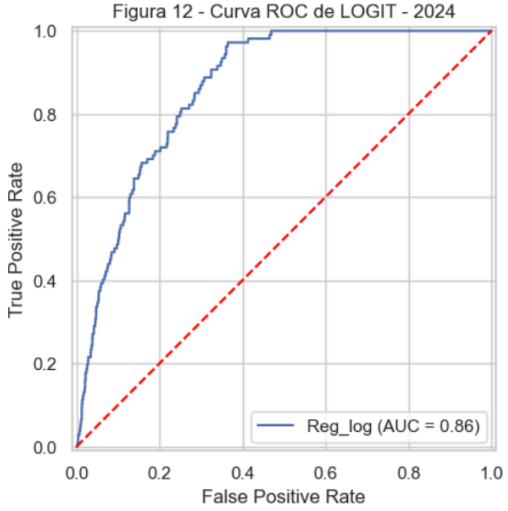
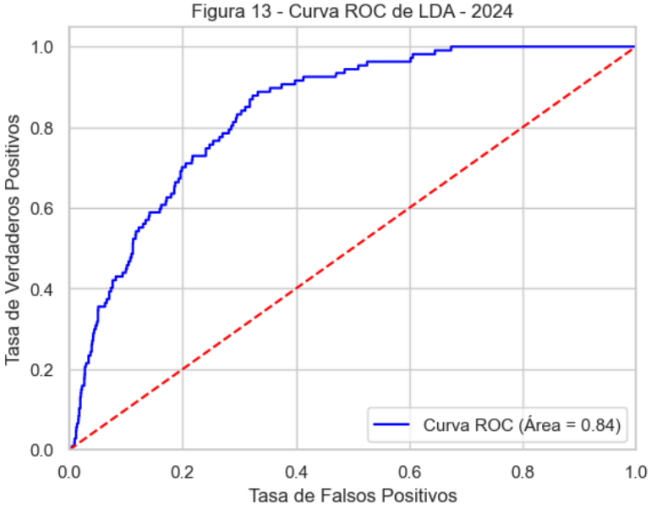


Figura 13 - Curva ROC de LDA - 2024



En el método de KNN (Figura 14) el AUC tomó el valor de 0.6630, demostrando ser un clasificador moderado para distinguir entre las clases positivas y negativas; en relación con la matriz de confusión, se identificaron 1969 verdaderos positivos, 27 falsos positivos, 98 falsos negativos y 9 verdaderos negativos, demostrando que identifica mejor los positivos que los negativos, ya que tiene mayor proporción de negativos mal clasificados. En tanto al nivel de exactitud, el modelo demostró tener un 94,1% de exactitud en la clasificación general.

Como último método utilizado, en Naive Bayes (Figura 15) se obtuvo un AUC de 0.76, demostrando que tiene un buen rendimiento en la clasificación de clases. En la matriz de confusión se identificacion 1039 verdaderos positivos, 957 falsos positivos, 0 falsos negativos y 107 verdaderos negativos, manifestando que el modelo identifica correctamente los negativos, pero no los positivos. Por último, el nivel de exactitud de clasificación general del modelo es de 0.545%, demostrando que no tiene un alto nivel de exactitud al clasificar.

Figura 14 - Curva ROC de KNN - 2024

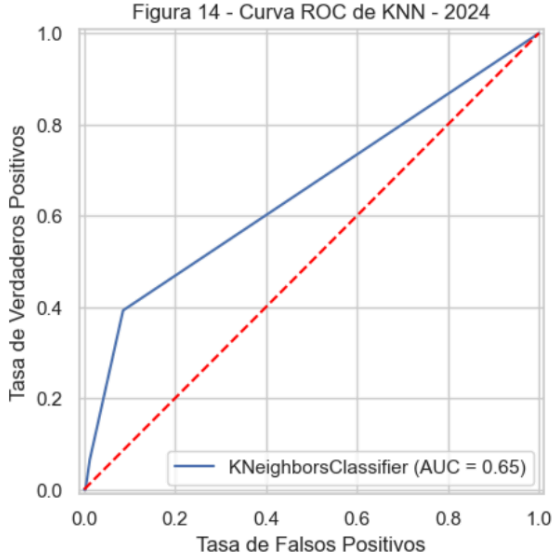
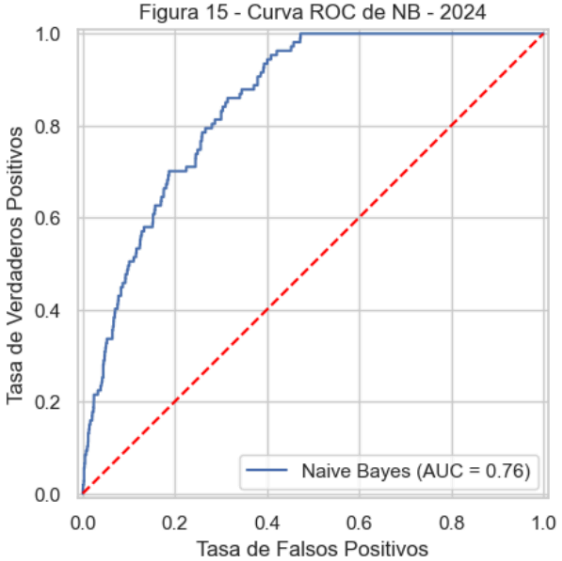


Figura 15 - Curva ROC de NB - 2024



Para definir cuál de los modelos predice mejor comparamos los AUC y el porcentaje de Accuracy de cada uno de los años (Tabla 2.1 y 2.2). Para ambos años, el modelo de LDA, tiene valores de Accuracy mayores que los otros métodos, demostrando un buen

Tabla 2.1 - Resultados para el año 2004:

Modelo	AUC	Accuracy
Regresión Logística	0.8692	0.9288
LDA	0.8534	0.93
KNN	0.7067	0.915
Naive Bayes	0.79	0.603

Tabla 2.2 - Resultados para el año 2024:

Modelo	AUC	Accuracy
Regresión Logística	0.8638	0.948645
LDA	0.8375	0.95
KNN	0.663	0.941
Naive Bayes	0.76	0.545

rendimiento. Toma también el segundo valor más alto de AUC, resultado que también se puede ver en las curvas ROC (Figuras 16 y 17) por lo que podemos asumir que tiene un buen rendimiento en la clasificación por clases, por lo que se cree que es el mejor método de predicción en este caso.

Figura 16 - Curvas ROC - Comparación de Modelos (2004)

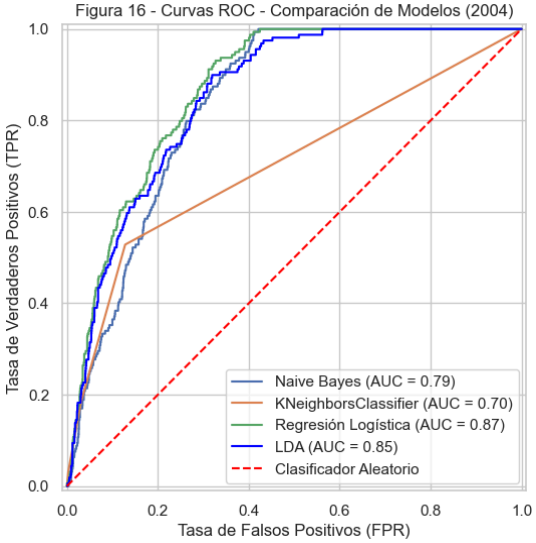
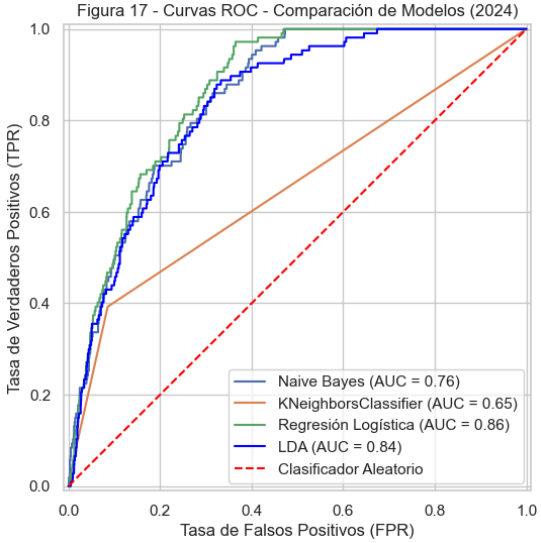


Figura 17 - Curvas ROC - Comparación de Modelos (2024)



Para finalizar con el trabajo, se utilizó el análisis discriminante lineal para entrenar un modelo con las personas que respondieron a la encuesta en ambos años y así predecir cuántas personas dentro dentro de la base 'no_resp' entraron en la categoría de desocupadas. De la misma forma que para evaluar los diferentes métodos anteriormente, se preparó la base de datos separando entre el conjunto de entrenamiento y de testeo. El modelo implementado obtuvo un puntaje de 96% de Accuracy e identificó a un desocupado dentro de las 51 personas que no respondieron la encuesta. Esto entonces refleja el 1.96% de las personas dentro de la base.