Universidad de
SanAndrés

Tercer Trabajo Práctico

Ciencias de Datos

2 de Octubre 2024

Profesor: Ignacio Spiousas

Zahira Chaia
Martina Lucas

Parte I: Analizando la base

Ejercicio 1

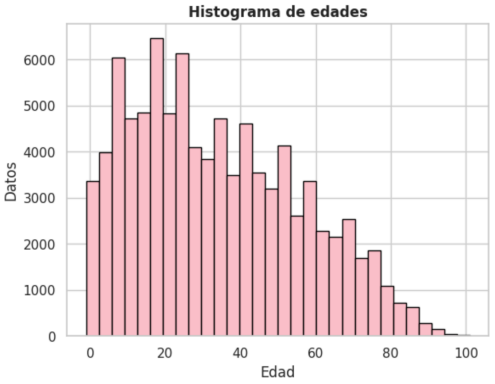
Según el INDEC, la población desocupada se refiere a las personas que desean trabajar, buscan empleo activamente, pero no logran conseguirlo. Este grupo no incluye a quienes enfrentan otras formas de precariedad laboral, como empleos transitorios, trabajos con pocas horas, quienes han dejado de buscar empleo por falta de oportunidades, empleos con salario mínimo o puestos por debajo de su nivel de calificación.

Ejercicio 2

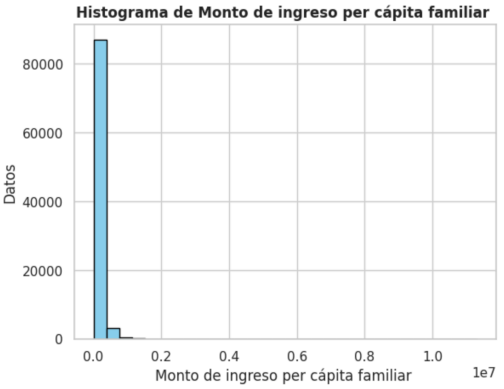
<u>Inciso b</u>

Comenzamos evaluando las variables del DataFrame utilizando la función describe(). Dado que el conjunto de datos contiene un gran número de categorías, identificamos aquellas que resultaban relevantes: ['anio', 'ch04', 'ch06', 'ch07', 'ch08', 'ch09', 'nivel_ed', 'estado', 'cat_inac', 'ipcf']. Con esta selección, construimos un nuevo dataset. Posteriormente, implementamos una función para evaluar el tipo de dato de cada variable, transformando todas ellas al mismo tipo, específicamente a cadena de texto (string), para facilitar su uso y análisis.

Procedimos a examinar todas las variables numéricas, comenzando con la variable correspondiente a la edad ("**ch06**"). Para ello, elaboramos un histograma que nos permitió visualizar la distribución de la población según su edad. Posteriormente, analizamos los datos negativos y encontramos que había 334 registros en esta categoría. Este número representaba aproximadamente el 0.37% del total de los datos. Por lo tanto, decidimos proceder a la eliminación de estos registros.



Posteriormente, examinamos la variable "**ipcf**" (Monto de ingreso per cápita familiar) y elaboramos un histograma. Al evaluar los datos, constatamos que no había registros negativos. A continuación, calculamos la media y el desvío estándar para identificar los datos extremos. Utilizando el criterio de aquellos datos que estaban por encima de tres desvíos estándar de la media (Diligent, 2024), eliminamos 1134 registros superaban este umbral, lo que representaba aproximadamente el 1.25% del total de los datos.

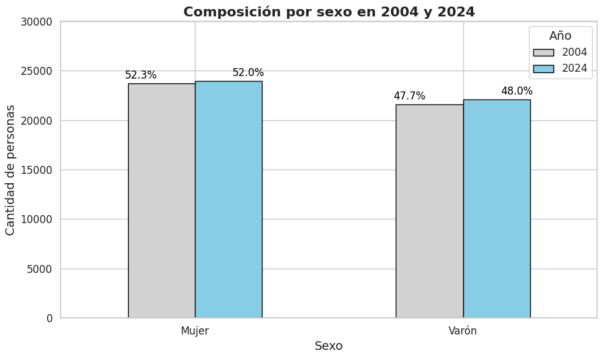


A continuación, mediante una función llamada evaluar_ns_nr (df) observamos que varias columnas tenían datos clasificados como Ns./Nr.

(no sabe / no responde). En la variable "ch07", la cantidad de valores Ns./Nr es 1 y en la variable "ch09"es 3, lo que representa menos del 0.1% de los datos. En la variable "ch08", la cantidad de Ns./Nr es 134, equivalentes al 0.149% del total. Dado el bajo porcentaje con respecto al total decidimos proceder a eliminarlos, ya que no aportaban datos significativos para el análisis.

Inciso c

El gráfico muestra la composición por sexo de la población encuestada en los años 2004 y 2024, representando las cantidades de personas identificadas como "Mujer" y "Varón" en ambos períodos. Se observa que las cantidades entre mujeres y varones se mantienen bastante similares en ambos años, lo que sugiere una distribución equilibrada en términos de género en la muestra tomada. En el caso de las mujeres, hay una leve disminución en el número total en 2024 en comparación con 2004, mientras que para los varones también se nota una pequeña disminución, aunque menos pronunciada.



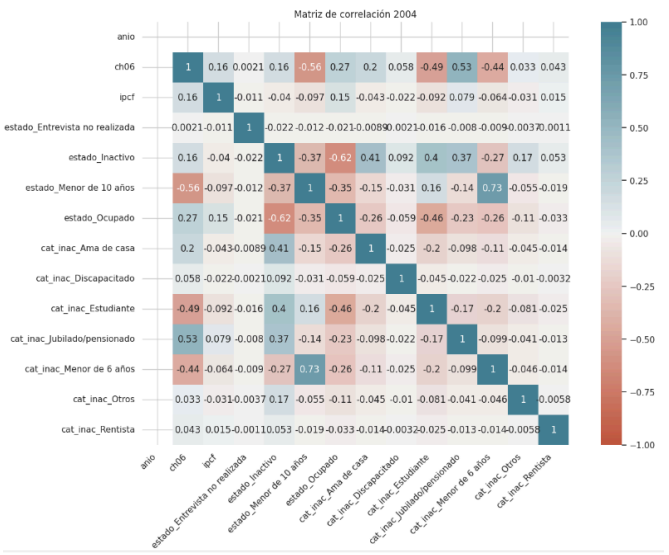
Inciso d

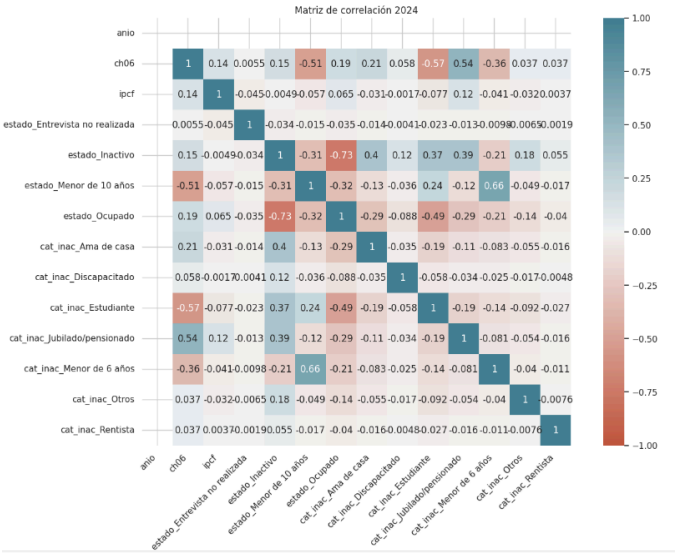
En primer lugar, seleccionamos las columnas de interés y creamos un nuevo data frame para la matriz de correlación. El primer paso fue reducir, si era posible, la cantidad de categorías de cada columna. De la columna **"ch08"** decidimos eliminar la distinción entre aquellos que tienen obra social y aquellos que además de obra social cuentan con otros tipos de cobertura. De esta forma redujumos las variables a 'Obra social', 'mutual/prepaga/servicio de emergencia', 'Planes y seguros públicos' y 'No paga ni le descuentan'. Realizamos un procedimiento similar con la variable "**nivel_edu"** , donde eliminamos la distinción entre aquellas personas que dejaron incompleto un nivel educativo y quienes no lo comenzaron (por ejemplo, fusionamos las personas con primaria incompleta y sin instrucción). Po último modificamos la variable **"ch08"** referida al estado civil, quedándonos con las categorías "soltero", "separado, divorciado o viudo" y "Casado o unido".

El siguiente paso fue transformar las variables en numéricas, diferenciando entre binomiales, categóricas a las que se les puede atribuir un orden y categóricas a las que no. Renombramos la variable binomial sexo estableciendo que 0 = varón y 1 = mujer. A continuación, ordenamos la variable "nivel educativo" según el nivel de instrucción alcanzado (0 = "sin instrucción", 1 = primaria completa, etc.). Por último, le atribuimos un orden a la columna correspondiente al estado civil como: 0 = "soltero", 1 = "separado,

divorciado o viudo" y 3 = "Casado o unido". Hicimos algo similar con la variable **"ch08"**, ordenando según el nivel de mayor cobertura estableciendo el siguiente orden: 0 = 'No paga ni le descuentan', 1 = planes y seguros públicos, 2 = 'Mutual/Prepaga/Servicio de emergencia', 3 = 'Obra social '.

Por último, para lidiar con las variables a las que no se les puede atribuir un orden, utilizamos One-Hot encoding. Aunque consideramos otras técnicas como el *frequency encoding* (Neural Ninja, 2023) para evitar aumentar excesivamente el número de variables, concluimos que la frecuencia en la que se presentan las categorías (como por ejemplo "ama de casa" o "menor de 10 años") no eran relevantes en este análisis. Por esta razón, decidimos utilizar One-Hot encoding (Kozina et al., 2024), técnica que permite convertir variables categóricas en una serie de variables binarias independientes. Esto permite calcular la correlación entre una categoría específica (por ejemplo, "Desocupado") y otra variable, en lugar de calcular una correlación entre números que podrían no tener un significado intrínseco.





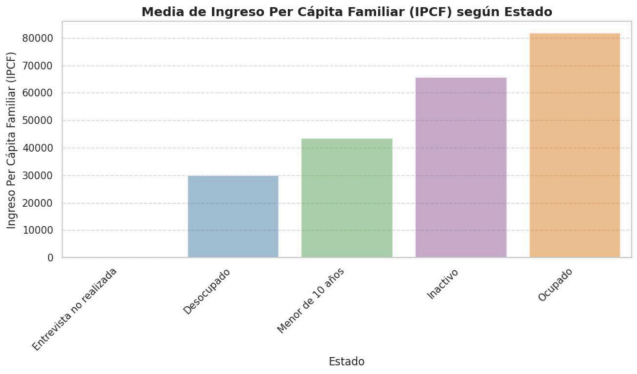
En las figuras anteriores podemos observar la matriz de correlación del año 2004 y el año 2024. A partir de ellas identificamos las siguientes correlaciones significativas, que osn muy similares en ambos casos. Existe una correlación negativa notable entre estado_Ocupado y estado_Inactivo (-0.62 y -0.73), algo esperable ya que son categorías mutuamente excluyentes. Además, estado_Inactivo muestra una correlación positiva moderada con cat_inac_Ama de casa (cerca de 0.40 en ambas), indicando que las personas inactivas son en gran parte amas de casa. Por otro lado, lógicamente, hay correlaciones entre distintas variables relacionadas con la edad. La variable estado_Menor de 10 años se correlaciona con otras variables relacionadas con la edad como ch06 y cat_inac_Menor de 6 años. Además, 'cat_inac_Jubilado/pensionado' tiene correlaciones positivas con

estado_Inactivo (cerca de 0.40) y ch06 (aproximadamente 0.50), sugieriendo que los jubilados o pensionados tienden a estar inactivos y tienen mayor edad.

Inciso e

Mediante la función **shape()** contamos la cantidad de desocupados en la columna "estado" del data frame completo, sumando 4061 que representa un 4.53% del total. Luego realizamos el mismo procedimiento para calcular la cantidad de personas inactivas, siendo 36096, un 40.23% del total. Por último calculamos la media de ingreso per cápita familiar (IPCF) dependiendo del "estado" y obtuvimos los siguientes resultados.

	estado	ipcf
1	Entrevista no realizada	34.492928
0	Desocupado	30007.465008
3	Menor de 10 años	43570.328151
2	Inactivo	65629.846407
4	Ocupado	81853.163418



El gráfico muestra una clara tendencia donde el grupo "Ocupado" alcanza el mayor IPCF, seguido por los "Inactivos". Los "Desocupados" y aquellos cuya "Entrevista no fue realizada" tienen los ingresos más bajos, lo que subraya la relación entre el estatus laboral y el IPCF.

Ejercicio 3

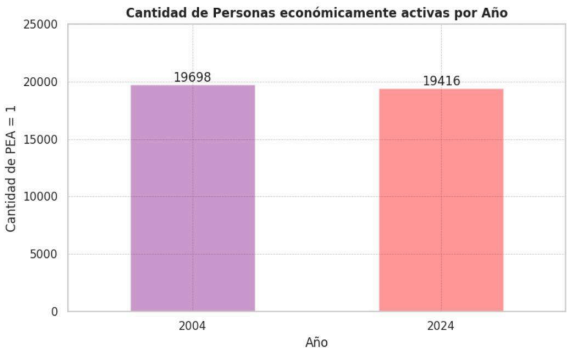
Como se señala en el informe, uno de los principales problemas de la Encuesta Permanente de Hogares (EPH) es el aumento en el número de hogares que no reportan sus ingresos. En este contexto, hemos calculado la cantidad de sujetos que no proporcionaron información sobre su condición de actividad, utilizando la variable "Entrevista individual no realizada", que registra dicho dato y cuyo total fue de 101. Con esta información, creamos un dataframe denominado "norespondieron". Posteriormente, generamos otro dataframe para aquellos que sí respondieron, el cual incluye las categorías "Inactivo", "Ocupado", "Desocupado" y "Menor de 10 años", todas correspondientes a la variable "estado".

Ejercicio 4

A continuación, se agregó una columna denominada "PEA" (Población Económicamente Activa) al DataFrame "respondieron". Esta nueva columna asignó un valor

de 1 si el estado del individuo era "Ocupado" o "Desocupado", y 0 en caso contrario. Posteriormente, se elaboró un gráfico de barras que mostró la composición por PEA para los años 2004 y 2024.

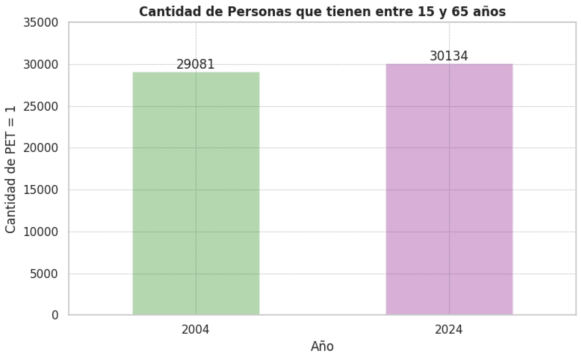
Se observó una disminución aproximada del 1.43% en la cantidad de personas económicamente activas entre 2004 y 2024. En 2004, se registraron 19,698 individuos en esta categoría, mientras que en 2024 la cifra descendió a 19,416. Este decrecimiento puede atribuirse a diversos factores. Sin embargo, no contamos con la información suficiente para extraer conclusiones definitivas al respecto.



Ejercicio 5

En primer lugar, se realizó un análisis de la variable "ch06", que contiene información sobre las edades de los sujetos, utilizando el método unique() para examinar su contenido. Se reemplazó la categoría "Menos de 1 año" por 1 y "98 y más años" por 98, ya que estos valores estaban en formato de texto. A continuación, se garantizó la consistencia de los datos al convertir la columna a tipo numérico. Se identificaron 0 valores negativos de edad, y finalmente se creó la columna "PET", que asignó el valor de 1 a las edades comprendidas entre 15 y 65 años. Se contabilizó un total de 59,215 personas en este rango etario. Posteriormente, se elaboró un gráfico de barras para mostrar la composición de la variable "PET" en los años 2004 y 2024. En 2004, se registraron 29,081 individuos, mientras que en 2024 la cifra ascendió a 30,134. Este incremento del 3.63% indica un aumento en la cantidad de personas en edad de trabajar (entre 15 y 65 años).

A pesar del aumento del 3.63% en la población en edad de trabajar, la PEA mostró una reducción del 1.43%. Esto podría sugerir que, aunque hay más personas en el rango de edad para trabajar, no todas están activamente participando en la fuerza laboral.



Ejercicio 6

Agregamos a la base de datos una columna llamada "desocupado" que tomó el valor de 1 si la persona estaba desocupada. La cantidad de personas desocupadas en 2004 fue de 2,713, mientras que en 2024 se registraron 1,342 personas desocupadas.

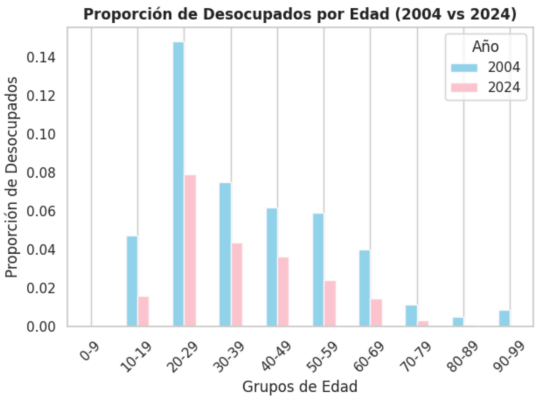
Inciso a

La cantidad de desocupados por nivel educativo ha experimentado cambios significativos entre 2004 y 2024. En 2004, el número de desocupados era de 567 en "Primaria Completa" (6.65%), 202 en "Primaria Incompleta" (2.16%), 609 en "Secundaria Completa" (10.35%), 594 en "Secundaria Incompleta" (7.28%), 13 en "Sin instrucción" (0.27%), 181 en "Superior Universitaria Completa" (5.14%) y 547 en "Superior Universitaria Incompleta" (11.15%). Para 2024, estas cifras mostraron una notable reducción: 120 desocupados en "Primaria Completa" (2.34%), 41 en "Primaria Incompleta" (0.65%), 497 en "Secundaria Completa" (4.98%), 319 en "Secundaria Incompleta" (3.34%), 0 en "Sin instrucción" (0.00%), 138 en "Superior Universitaria Completa" (2.25%) y 238 en "Superior Universitaria Incompleta" (4.20%). Estos datos indican una disminución general en la desocupación en todos los niveles educativos, lo que sugiere una mejor situación laboral en comparación con 2004.

Inciso b

Creamos una variable categórica para agrupar la edad en intervalos de 10 años (de 0 a 100) utilizando la columna "ch06". Luego, generamos una nueva columna llamada "edad_cada10" en el DataFrame respondieron, que clasificó a cada persona en estos grupos de edad. A continuación, calculamos la cantidad de desocupados por año y grupo de edad, y también obtuvimos la cantidad total de desocupados para calcular la proporción de desocupados en cada grupo.

A partir de este gráfico observamos las siguientes cosas relevantes. En primer lugar, en casi todos los grupos de edad, se observa una disminución de la desocupación en 2024 en comparación con 2004, especialmente en los grupos de 10-19, 20-29 y 30-39 años. Además, el grupo etario de entre 20 a 29 años tiene la proporción de desocupación más alta en ambos años, con un valor notablemente mayor en 2004. Esto tiene sentido si consideramos que este grupo está recién comenzando a insertarse en el mercado laboral. Por otro lado, la proporción de desocupados en los grupos de edades avanzadas es baja en ambos años, y apenas se observan cambios significativos. Es importante recordar la definición del INDEC de desocupación, que excluye de este grupo a quienes no buscan trabajo activamente, como suele ocurrir en este grupo etario.



Parte II: Clasificación

Ejercicio 1

En primer lugar, utilizamos el comando train_test_split, asignando el 70% de los datos al conjunto de entrenamiento y el 30% al conjunto de prueba, con una semilla de 101 para garantizar la reproducibilidad de los resultados (Nguyen, 2020). La separación de los datos en conjuntos de entrenamiento y prueba es crucial en el desarrollo de modelos de machine learning. El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba sirve para evaluar su rendimiento en datos no vistos. Esto permite detectar problemas como el sobreajuste, en el cual un modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos (Bishop, 2006). Posteriormente, establecemos la variable "desocupado" como la variable dependiente en el conjunto de entrenamiento (vector y), mientras que el resto de las variables serán consideradas como las variables independientes (matriz X).

Ejercicios 2 y 3

A partir de la división de la base en datos de entrenamiento y testeo implementamos cuatro métodos para cada año: regresión logística, Análisis discriminante lineal, KNN y naive bayes. A partir de la comparación de los resultados de los modelos de clasificación entre los años 2004 y 2024 se reportan los siguientes resultados.

Con la regresión logística para el año 2004 la matriz de confusión muestra 12,615 verdaderos negativos, 110 falsos positivos, 294 falsos negativos y 535 verdaderos positivos. La exactitud es del 97.01%, lo que indica una clasificación correcta en la mayoría de los casos. El área bajo la curva (AUC) de 0.989 sugiere una alta capacidad para distinguir entre clases positivas y negativas. La curva ROC refleja una alta tasa de verdaderos positivos en comparación con la tasa de falsos positivos , lo cual confirma un buen rendimiento del modelo. Por otro lado, para el mismo año, el modelo LDA presenta una exactitud del 94%, pero muestra serias limitaciones en la detección de casos positivos, con solo 4 verdaderos positivos y 825 falsos negativos. Aunque el AUC sugiere un buen rendimiento en general, la baja sensibilidad evidencia que no es efectivo para detectar la clase positiva. El modelo KNN para 2004 tiene una exactitud similar (94%), pero su AUC es menor (0.74), lo que indica un peor rendimiento en comparación con la regresión logística. Además, tiene una alta cantidad de falsos negativos (608), lo que sugiere que no es el mejor para identificar correctamente los casos positivos.

En el año 2024, la regresión logística sigue siendo el modelo con la mejor exactitud (97.30%), pero su AUC cae dramáticamente a 0.64, lo que indica una disminución en su capacidad para discriminar entre clases. Aunque mantiene una baja cantidad de falsos positivos (13), el aumento de falsos negativos (346) sugiere una pérdida de efectividad en la detección de casos positivos. A pesar de su alta exactitud, el bajo AUC demuestra que este

modelo es menos efectivo para clasificar correctamente cuando las clases están menos diferenciadas. El modelo LDA en 2024 también muestra problemas similares, con solo 2 verdaderos positivos y 421 falsos negativos. KNN, con una exactitud también del 97%, pero un AUC de 0.66, muestra un rendimiento general bajo, aunque ligeramente mejor que el de LDA, pero aún con dificultades para discriminar las clases de manera efectiva.

En cuando al modelo de naive bayes en 2004, tiene una precisión perfecta (1.00) y un AUC de 1.00, lo que indica que clasifica correctamente todas las instancias. Este comportamiento podría indicar un sesgo de ajuste (overfitting) ya que es improbable que un modelo generalice tan bien en la práctica, lo que puede afectar su desempeño en datos nuevos. En contraste, en 2024, el modelo tiene una precisión de 0.97 y un AUC de 0.79, mostrando un comportamiento más realista. En este caso el problema principal es la precisión, que aunque alta, puede estar ocultando deficiencias en la detección de ciertas clases.

En resumen, para el año 2004, el mejor modelo es la regresión logística, que ofrece una alta exactitud y una destacada capacidad discriminativa, detectando la mayoría de los casos positivos con pocos errores. En cambio, para el año 2024, a pesar de que la regresión logística sigue siendo el modelo con mayor exactitud, su capacidad para diferenciar entre clases (AUC) disminuye significativamente, lo que sugiere que ningún modelo predice tan bien en 2024 como lo hacía en 2004. Esto indica que los datos de 2024 podrían requerir un enfoque diferente o un ajuste de los modelos para mejorar su capacidad discriminativa.

Ejercicio 4

A partir de la regresión lineal implementada en el punto anterior, procedimos a predecir la cantidad de personas desocupadas dentro de la base norespondieron. Antes de hacerlo fue necesario renombrar las variables de norespondieron siguiendo el mismo procedimiento que con la base respondieron. Además, ajustamos el número de columnas para que coincidiera con los datos con los que el modelo fue entrenado.

En primer lugar, implementamos el modelo creado a partir de los datos de 2004, donde se predijo que el número de desocupados es 75, lo que representa una proporción de aproximadamente 0.0.743 del total. Luego continuamos con la predicción de la cantidad de personas desocupadas dentro de la base norespondieron usando el modelo de 2024. Se predijo que el número de desocupados es de 50, lo que representa una proporción extremadamente baja, de aproximadamente 0.495 del total de la muestra.

Bibliografía

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Diligent. (n.d.). *Método outliers() en HCL*. En *Centro de ayuda de la plataforma Diligent One*. Recuperado de https://help.highbond.com/helpdocs/highbond/es/Content/robots/scripting/hcl/hcl_outliers.htm

IBM. (n.d.). *¿Qué es el análisis exploratorio de datos?* Recuperado el October 1, 2024, de https://www.ibm.com/topics/exploratory-data-analysis

Kozina, A., Nadolny, M., Hernes, M., Walaszczyk, E., & Rot, A. (2024). One Hot Encoding and Hashing Trick Transformation - Performance Comparison. 2024 14th International Conference on Advanced Computer Information Technologies (ACIT), Ceske Budejovice, Czech Republic, 699–704. https://doi.org/10.1109/ACIT62333.2024.10712459

Neural Ninja. (2023, 12 de junio). *Frequency encoding: counting categories for representation*. Recuperado el October 1, 2024, de https://letsdatascience.com/frequency-encoding/

Nguyen, M. (2020). *A guide on data analysis*. Bookdown. https://bookdown.org/mike/data_analysis/