Universidad de
SanAndrés

Maestría en Economía

Machine Learning para Economistas

Profesor: Walter Sosa Escudero
Asistente: Tomás Pacheco

Trabajo Práctico 3: Análisis Descriptivo y Predicción de Desocupación

De Boeck, Carolina
Hausvirth, Martina
Hayduk, Gaspar

Fecha de entrega: 17 de Noviembre de 2024

Parte I: Analizando la base

Ejercicio 1

En la Encuesta Permanente de Hogares (EPH) del INDEC, las personas desocupadas se identifican a partir de la variable ***ESTADO***, que clasifica la condición de actividad de cada individuo. Una persona se considera desocupada si, al momento de la encuesta, no tiene empleo pero está buscando activamente trabajo y está disponible para trabajar. Esta definición incluye tanto a quienes buscan trabajo por primera vez como a quienes quedaron sin empleo y están en busca de uno nuevo.

En la descripción de la condición de actividad, los encuestados se pueden clasificar entre ocupados, desocupados, inactivos y menores de 10 años. En este contexto, aquellos que se seleccionan como desocupados, son quienes no participan de ninguna actividad productiva, pero que manifiestan la voluntad de participar.

Ejercicio 2

Inciso a)

En este ejercicio, unificamos y estandarizamos los datos de 2004 y 2024, aplicando un mapeo de categorías y seleccionando solo las observaciones del aglomerado "Gran Tucumán - Tafí Viejo", para crear una base de datos consolidada y homogénea para facilitar el análisis.

Inciso b)

En primer lugar, realizamos un análisis descriptivo de las variables en nuestra base unificada para evaluar si existen datos atípicos o sin sentido que debamos descartar. Para esto, generamos un resumen que almacena los valores de cada variable y su frecuencia en una lista llamada *resultados*. Luego, convertimos esta lista en un DataFrame *resumendf*, lo que nos permite revisar los datos observados y compararlos con los valores esperados según el resumen de referencia del INDEC. Durante esta revisión, identificamos varios ceros en variables como $H15$, $CH10$ y $CH11$, los cuales no corresponden a valores válidos y los interpretamos como ausencias de observación.

A continuación, procedimos a filtrar los datos para eliminar valores que no tenían sentido (en este caso, que fueran negativos) en categorías como ingresos y edad. Para ello, definimos un filtro que conserva los valores no negativos o NaN tanto en la variable de edad (CH06) como en las columnas de ingresos (PP08D1, IPCF, PP08D4, PP08F1, PP08F2, PP08J1, PP08J2, PP08J3). Aplicamos este filtro al dataframe original, obteniendo una versión limpia y depurada de los datos denominada *datospanellimpio*.

Inciso c)

Para este ejercicio, realizamos un gráfico de barras mostrando la composición por sexo para 2004 y 2024.

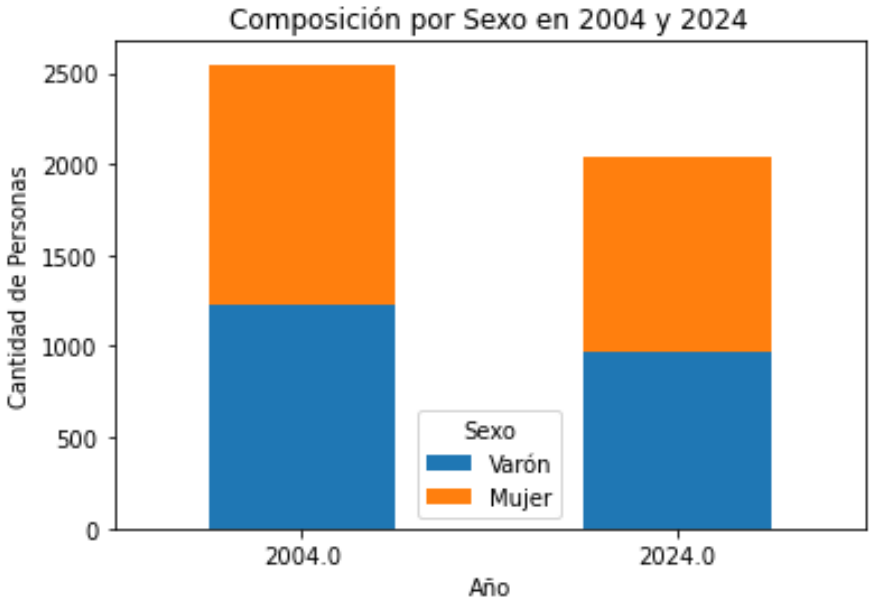

Figura 1: Composición por Sexo en 2004 y 2024

Lo que podemos observar de este gráfico, es que si bien la proporción de sexos se mantuvo medianamente constante en el tiempo, disminuyeron la cantidad de personas encuestadas para esta región entre el 2004 y el 2024.

Inciso d)

En este ejercicio, calculamos la matriz de correlación para los datos del año 2004 y 2024 utilizando un conjunto de variables de interés ($CH04$, $CH06$, $CH07$, $CH08$, $NIVELED$, $ESTADO$, $CATINAC$, $IPCF$).

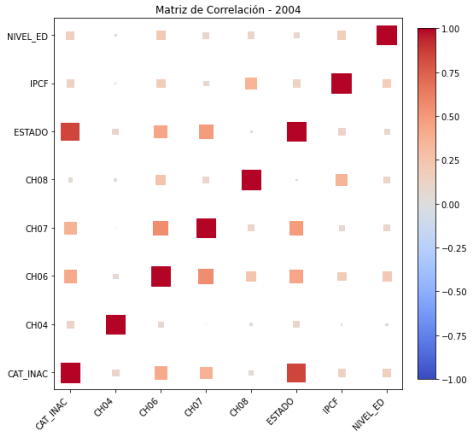

Figura 2: Matriz de Correlaciones 2004

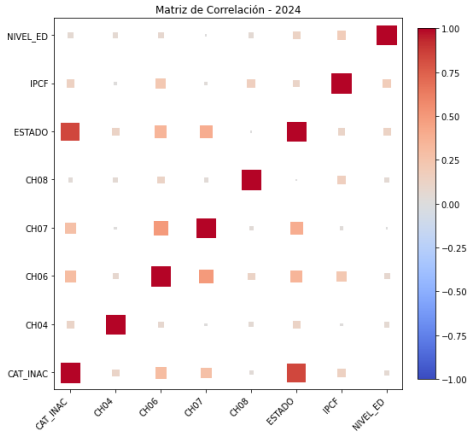

Figura 3: Matriz de Correlaciones 2024

En primer lugar calculamos las matrices de correlación para observar la relación entre pares de variables en cada año por separado.

Luego, para visualizar las matrices de correlación, utilizamos un gráfico tipo *heatmap*, inspirado en los comandos sugeridos en la consigna. Definimos una función llamada heatmap que convierte la matriz de correlación en formato largo utilizando pd.melt() y representa cada correlación como un

cuadrado con tamaño y color proporcionales a la magnitud de la correlación. Dentro de la función heatmap, se utiliza ax.scatter() para crear un gráfico de dispersión con marcadores cuadrados y una paleta de colores divergente (coolwarm), asignando el color rojo a las correlaciones positivas y el color azul a las negativas, mientras que los valores cercanos a cero son neutros (blancos), y se proyectan con menos intensidad. Podemos observar una correlación alta entre ESTADO y CAT_INAC; pues ESTADO es la variable que nos dice si el individuo es ocupado (=1), desocupado (=2), inactivo (=3) o menor de 10 años (=4), mientras que CAT_INAC nos dice si el individuo es jubilado, rentista, estudiante, ama de casa, menor de 10 años o discapacitado. Si sé que el individuo es jubilado, sé que seguro es inactivo (ESTADO=3).

Inciso e)

En nuestra muestra de datos, encontramos que hay en total hay 259 personas desocupadas, y 1,842 personas inactivas, lo cual resulta bastante llamativo, ya que este último número supera la cantidad de personas ocupadas que fueron encuestadas.

En cuanto al ingreso per cápita familiar, la siguiente tabla nos muestra el IPCF según el estado de actividad. Debido a la nominalidad que carga IPCF y a la inflación entre 2004 y 2024 nos pareció apropiado agrupar por año también.

Descripción	Año	IPCF
Entrevista no realizada	2004	212.39
	2024	0.00
Ocupado	2004	254.97
	2024	162267.64
Desocupado	2004	142.83
	2024	101478.12
Inactivo	2004	204.79
	2024	127529.07
Menor de 10 años	2004	151.23
	2024	98507.63

Cuadro 1: Media de ingreso per cápita familiar (IPCF) por estado de actividad

Ejercicio 3

La no respuesta en encuestas de hogares es un desafío recurrente, especialmente en preguntas relacionadas con ingresos. Dado que los individuos no se postulan voluntariamente para responder, no todos están dispuestos a proporcionar respuestas a cada una de las preguntas formuladas.

En nuestro análisis de la base de datos, observamos que hay un total de 10 personas que no respondieron sobre su condición de actividad. Estas observaciones se han separado en un nuevo DataFrame denominado *norespondieron*, mientras que las observaciones que sí respondieron están agrupadas en un DataFrame llamado *respondieron*.

Ejercicio 4

La siguiente figura muestra la composición de la Población Económicamente Activa en 2004 y 2024. Podemos observar que los desempleados ocupan una menor proporción de la muestra total en 2024 que en 2004.

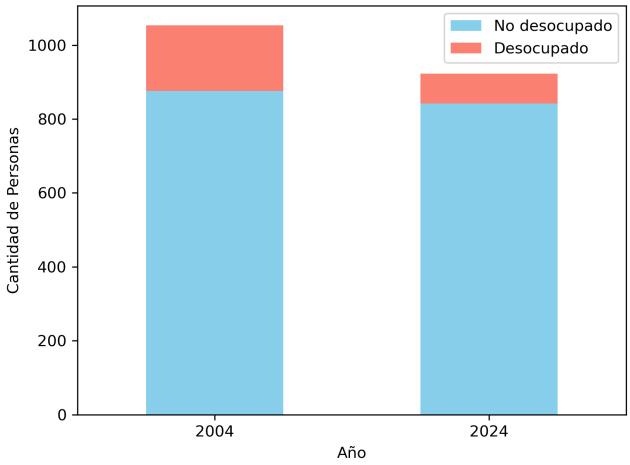

Figura 4: Composición de la Población Económicamente Activa en 2004 y 2024

Ejercicio 5

La siguiente figura muestra la composición de la Población en Edad de Trabajar en 2004 y 2024. A ojo, podemos observar que las personas en edad de trabajar ocupan una mayor proporción de la muestra total en 2024 que en 2004.

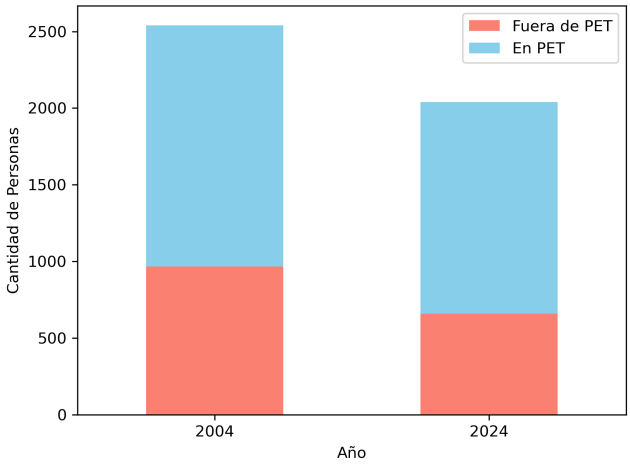

Figura 5: Composición de la Población en Edad de Trabajar en 2004 y 2024

Ejercicio 6

Año	Cantidad de Desocupados
2004	178
2024	81

Cuadro 2: Desempleados por año

Inciso a)

La siguiente tabla muestra la proporción de desocupados por nivel educativo para 2004 y 2024. Podemos observar que el desempleo cayó en todos los niveles educativos comparando 2024 contra 2004. El nivel educativo que menos disminuyó fue el nivel Superior Universitario Completo, esto probablemente se deba a que es el nivel educativo que menos pude verse afectado en términos de empleo por la recesión entre 1998 y 2002.

	Primario incompleto	Primario completo	Secundario incompleto	Secundario completo	Superior universitario incompleto	Superior universitario completo	Sin Instrucción
2004	2.10 %	7.62 %	6.72 %	14.23 %	17.23 %	5.37 %	0.32 %
2024	0.70 %	3.57 %	4.05 %	6.72 %	5.12 %	4.32 %	0

Cuadro 3: Proporción de Desocupados por nivel educativo

Inciso b)

La siguiente tabla muestra la proporción de desocupados por edad agrupada para 2004 y 2024. Podemos observar que la proporción de desocupados cayó en todos los grupos de edad entre 2004 y 2024.

Grupo de Edad	[0, 10)	[10, 20)	[20, 30)	[30, 40)	[40, 50)	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)
2004	0	5.67 %	18.54 %	10.00 %	7.54 %	4.55 %	2.74 %	0.94 %	0	0
2024	0	2.45 %	10.00 %	5.75 %	4.23 %	2.90 %	2.27 %	0	0	0

Cuadro 4: Proporción de Desocupados por grupo de edad

Parte II: Clasificación

El objetivo de esta sección es predecir si una persona está desocupada o no utilizando distintas variables de características individuales. Como predictores decidimos utilizar las siguientes variables:

- CH04: Sexo
- CH06: Edad
- CH07: Estado Civil (unido, casado, separado/a o divorciado/a, viudo/a, soltero/a)
- CH08: Cobertura Medica (Obra social (incluye PAMI), Mutual / prepaga / servicio de emergencia, Planes y seguros públicos, No paga ni le descuentan, Obra social y mutual / prepaga / servicio de emergencia, Obra social y planes y seguros públicos, Mutual / prepaga / servicio de emergencia / Planes y seguros públicos, Obra social, mutual / prepaga / servicio de emergencia y planes y seguros públicos)
- NIVEL_EDU: Nivel educativo (Primario incompleto (incluye educación especial), Primario completo, Secundario incompleto, Secundario completo, Superior universitario incompleto, Superior universitario completo, Sin instrucción)
- IPCF: Ingreso per capita familiar
- CAT_INAC: Categoría de inactividad (Jubilado / Pensionado, Rentista, Estudiante, Ama de casa, Menor de 6 años, Discapacitado)

A excepción de la edad y el ingreso familiar per capita, todas estas variables toman valores donde cada valor corresponde a una categoría. Para lidiar con esto, realizamos One Hot Encoding, creando una dummy por categoría.

Ejercicio 2

Regresión Logística (RL)

A continuación se reporta el AUC, la curva ROC, la matriz de confusión y el Accuracy para el método Regresión Logística para los años 2004 y 2024.

2004 (Accuracy: 0.932)

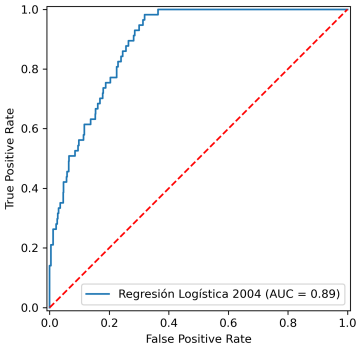

Figura 6: Curva ROC de la RL para 2004. El AUC es de 0.89

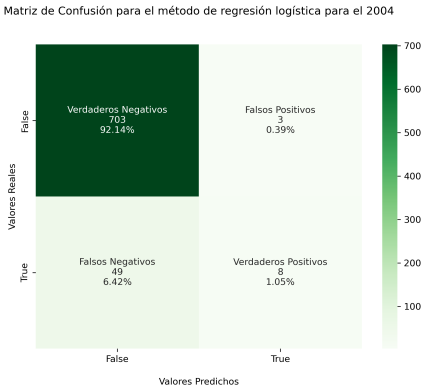

Figura 7: Matriz de Confusión de la RL para 2004

2024 (Accuracy: 0.954)

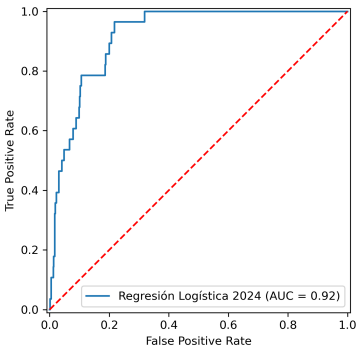

Figura 8: Curva ROC de la RL para 2024. El AUC es de 0.92

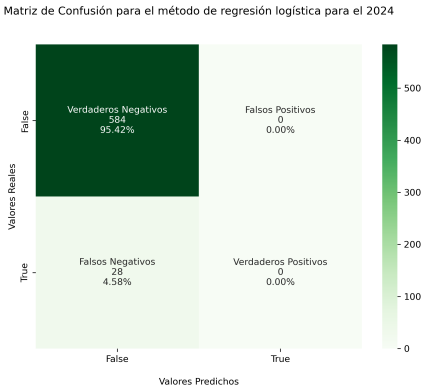

Figura 9: Matriz de Confusión de la RL para 2024

Análisis Discriminante Lineal (LDA)

A continuación se reporta el AUC, la curva ROC, la matriz de confusión y el Accuracy para el método de Análisis discriminante lineal para los años 2004 y 2024.

2004 (Accuracy: 0.927)

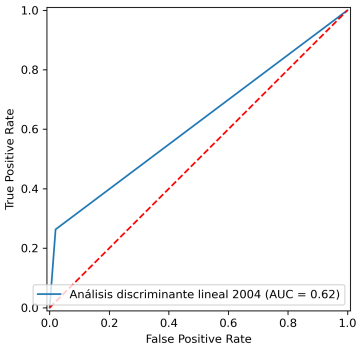

Figura 10: Curva ROC del LDA para 2004. El AUC es de 0.62

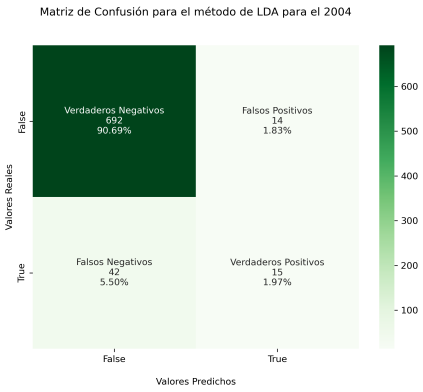

Figura 11: Matriz de Confusión del LDA para 2004

2024 (Accuracy: 0.954)

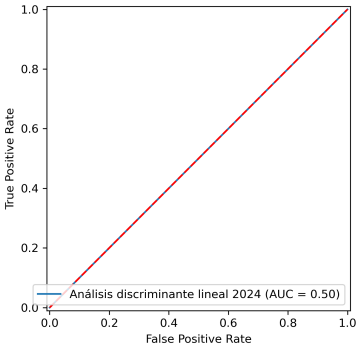

Figura 12: Curva ROC del LDA para 2024. El AUC es de 0.5

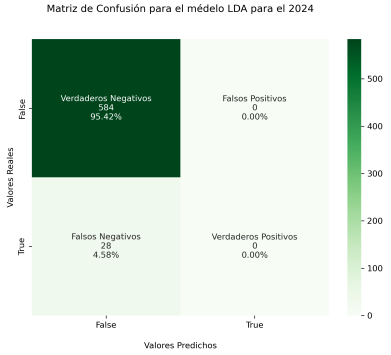

Figura 13: Matriz de Confusión del LDA para 2024

KNN con k=3

A continuación se reporta el AUC, la curva ROC, la matriz de confusión y el Accuracy para el método de KNN con k=3 para los años 2004 y 2024.

2004 (Accuracy: 0.915)

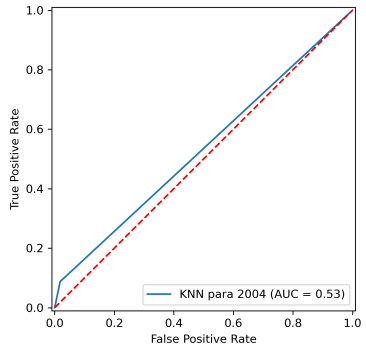

Figura 14: Curva ROC de KNN para 2004. El AUC es de 0.53

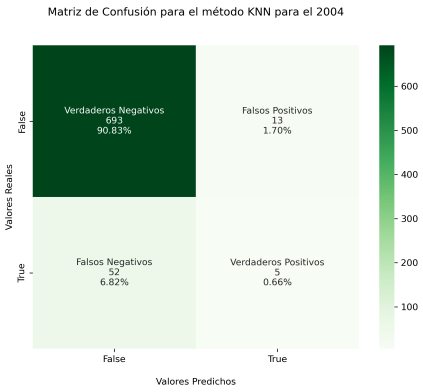

Figura 15: Matriz de Confusión de KNN para 2004

2024 (Accuracy: 0.935)

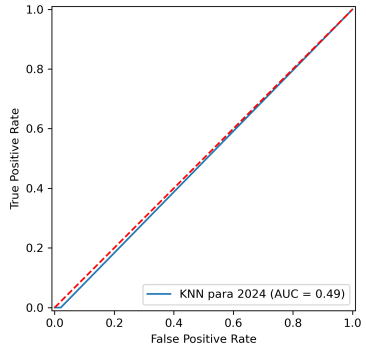

Figura 16: Curva ROC de KNN para 2024. El AUC es de 0.49

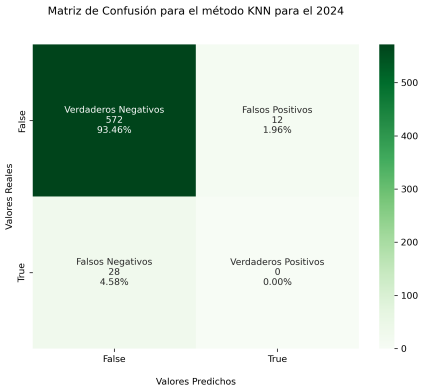

Figura 17: Matriz de Confusión de KNN para 2024

Ejercicio 3

Antes de decidir qué método predice mejor para cada año es necesario detallar qué métrica es la adecuada para seleccionar el método. El Accuracy mide el porcentaje de predicciones correctas sobre el total de predicciones, pero no toma en cuenta la distribución de clases; en un conjunto de datos desbalanceado, donde una clase es mucho más frecuente, el modelo puede lograr una alta precisión simplemente prediciendo siempre la clase mayoritaria. En cambio, el AUC mide la capacidad del modelo para distinguir entre las dos clases, independientemente de la distribución de clases. AUC se calcula en función de la tasa de verdaderos positivos y la tasa de falsos positivos a diferentes umbrales de decisión; captura mejor la habilidad del modelo para clasificar correctamente las instancias minoritarias (la clase menos frecuente) junto con la mayoría. El AUC refleja qué tan bien el modelo separa las clases, sin importar su proporción en el conjunto de datos. En nuestro caso, el desempleo (clase a predecir) está desbalanceada, hay muchos más no desempleados que desempleados. Por tanto, nos parece adecuado mirar el AUC antes que el Accuracy para elegir qué método predice mejor. Comparando el AUC para cada método para cada año, concluimos que el método de Regresión Logística es el método que mejor predice tanta para 2004 como para 2024.

Ejercicio 4

En nuestro caso para el aglomerado "Gran Tucumánn - Tafí Viejo" tenemos que solo 10 personas no respondieron su condición de actividad (ESTADO=0), 6 corresponden a 2004 y 4 a 2024. A las personas que no respondieron en 2004 les aplicamos el modelo de regresión logística estimado para 2004 y a las personas que no respondieron en 2024 les aplicamos el modelo de regresión logística estimado para 2024. Realizando la predicción no encontramos desempleados para ninguno de los dos años.