Comenzado el	jueves, 27 de febrero de 2025, 18:10
Estado	Finalizado
Finalizado en	martes, 11 de marzo de 2025, 10:58
Tiempo empleado	11 días 16 horas
Puntos	12,46/15,00
Calificación	**8,31** de 10,00 (**83**%)

Pregunta **1**

Parcialmente correcta

Se puntúa 0,80 sobre 1,00

Indique cuál de las siguientes afirmaciones forman parte de la gestión de riesgos (marque todas las opciones que considere necesarias)

Seleccione una o más de una:

- [x] a. Detección de operaciones fraudulentas con tarjetas de crédito ✔
- [x] b. Compra de opciones *call* sobre el tipo de cambio euro-dólar ✔
- [x] c. Detección de amenazas y vulnerabilidades que puedan suponer una pérdida ✔
- [x] d. Calificación de malos pagadores a traves de Bureaus de crédito **externos** (como *Experian* o *Asnef*) ✔
- [ ] e. Adoptar medidas de prevención de icendios **una vez que ya se ha sufrido un incendio que ha destruido parcialmente las intalaciones**
- [ ] f. Ninguna de las restantes opciones es cierta

En realidad todas las opciones mencionadas pueden considerarse como diferentes componentes de la gestión de (diferentes tipos de) riesgos.... incluida la adopción de medidas de prevención de inciendios después haber sufrido uno (aunque sea tarde, mejor prevenir nuevas amanezas en el futuro).

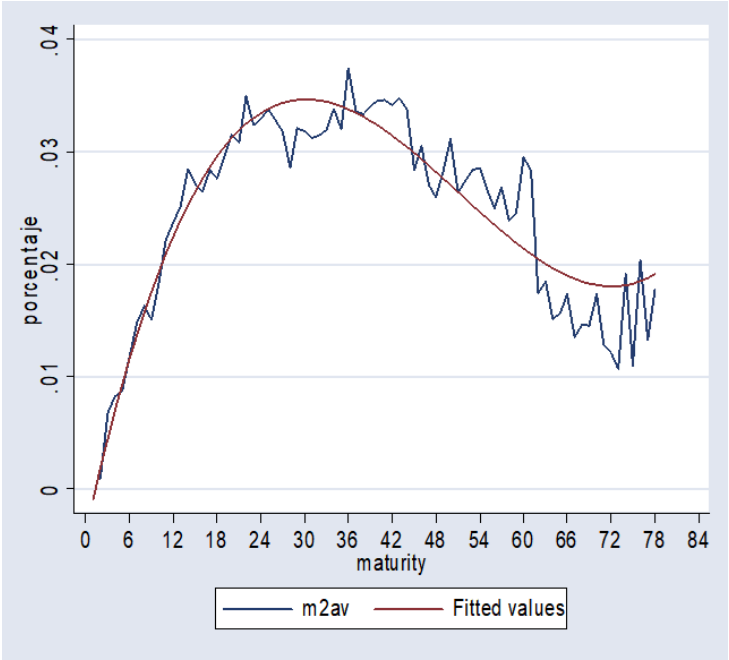
Las respuestas correctas son: Detección de operaciones fraudulentas con tarjetas de crédito, Compra de opciones *call* sobre el tipo de cambio euro-dólar, Detección de amenazas y vulnerabilidades que puedan suponer una pérdida, Calificación de malos pagadores a traves de Bureaus de crédito **externos** (como *Experian* o *Asnef*), Adoptar medidas de prevención de icendios **una vez que ya se ha sufrido un incendio que ha destruido parcialmente las intalaciones**

Pregunta **2**

Correcta

Se puntúa 1,00 sobre 1,00

Suponga que está realizando un modelo de Scoring de Riesgo de impago de créditos, y que dispone de la siguiente información sobre el porcentaje de créditos impagados (m2av) según el tiempo de vigencia del préstamo desde su concesión inicial ¿qué horizonte temporal utilizaría para definir la variable objetivo?. (marque todas las opciones que considere necesarias)



Seleccione una o más de una:

- a. Dos meses, ya que esa es la definición de mora del indicador m2av
- b. Debe considerarse una ventana completa durante toda la vigencia del crédito
- c. Entre más o menos los 24 y los 48 meses, ya que es el periodo para el que se observa el máximo de morosidad en la gráfica ✔
- d. A partir de aproximadamente los 48 meses, ya que a partir de entonces se reduce la tasa de morosidad según la gráfica
- e. La elección de la ventana es en realidad una cuestión de negocio que no sesga los resultados ni la interprtación del modelo de puntuación de riesgos (risk scorecard)
- f. Ninguna de las restantes opciones es cierta

No es necesario utilizar una ventana temporal que incluya todo el periodo de vigencia del crédito. De hecho una buena opción es elegir una ventana en la que se observe una mayor tasa de mora.

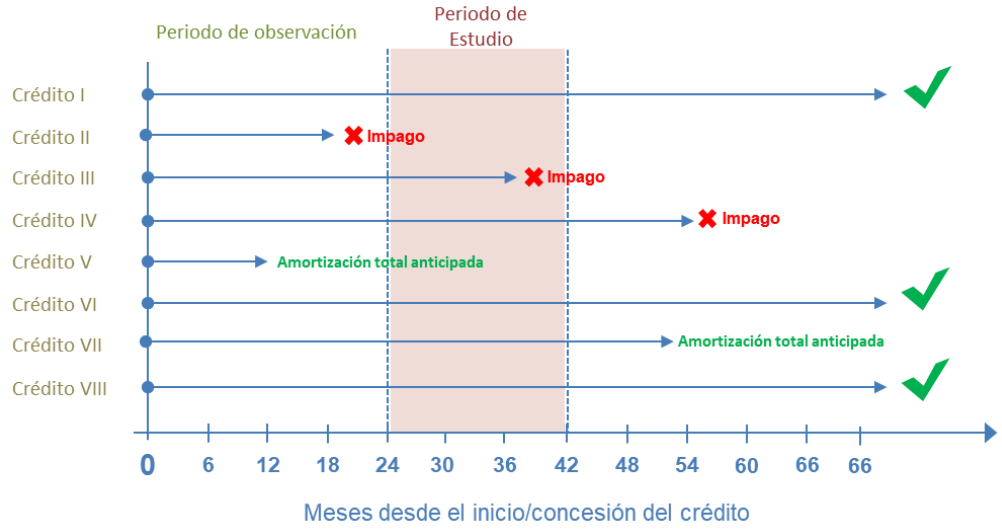
La respuesta correcta es: Entre más o menos los 24 y los 48 meses, ya que es el periodo para el que se observa el máximo de morosidad en la gráfica

Pregunta **3**

Correcta

Se puntúa 1,00 sobre 1,00

Suponga que tiene información de los siguientes siete créditos. Teniendo en cuenta la ventana o periodo de estudio considerado en el gráfico y el comportamiento de cada uno de los créditos en dicha ventana de observación, ¿cuál de esos siete gráficos **descartaría** para realizar un análisis de riesgo de impago? (marque todas las opciones que considere necesarias)



Seleccione una o más de una:

- [ ] a. Crédito I
- [x] b. Crédito II ✔
- [ ] c. Crédito III
- [ ] d. Crédito IV
- [x] e. Crédito V ✔
- [ ] f. Crédito VI
- [ ] g. Crédito VII
- [ ] h. Crédito VIII

El objetivo es detectar default en la ventana de estudio

Las respuestas correctas son: Crédito II, Crédito V

Pregunta **4**

Parcialmente correcta

Se puntúa 0,75 sobre 1,00

Una de las fases más interesantes en la construcción de modelos de puntuación de riesgos es la que corresponde con la tramificación de variables numéricas, la agrupación de niveles o categorías de las variables categóricas y la trasformación WOE de las variables predictoras. Esta fase es anterior a la estimación del modelo de clasificación o modelo de probabilidad propiamente dicho.

Por favor, de las siguintes afirmaciones, marque **todas** las opciones que considere que **son correctas**

Seleccione una o más de una:

- [ ] a. En realidad la tramificación de variables numéricas nunca es necesaria, siempre es mejor utilizar las variables numéricas (rentas de los solicitantes, edad, importe del crédito solicitado, etc) tal cual o con alguna trasformación monótona (transformación logarítmica por ejemplo) para no perder información.
- [x] b. La tramificación de variables numéricas sirve para poder aplicar los estadísticos de concentración (Valor de Información y/o Gini) utilizados para ver la importancia que pueden tener las potenciales variables predictoras antes de incorporarlas al modelo de clasificación propiamente dicho ✔
- [x] c. La tramificación de variables numéricas permite captar no-linealidades, aunque esto no sería necesario si se utilizasen modelos predictivos que no fuesen esencialmente lineales. Por ejemplo, en lugar de utilizar la regresión logística utilizar modelos de redes neuronales ✔
- [ ] d. Con la transformación WOE se pierde capacidad del modelo de regresión logística para captar no-linealidades. En este sentido, es preferible convertir las variables categóricas en variables dicotómicas (variables dummy) para cada uno de los niveles o categorías de las variables independientes.
- [x] e. La tramificación de variables numéricas y agrupación de categorías permite construir tarjetas de puntuación más fáciles de interpretar y de aplicar en la práctica con una simple calculadora de mano. Además permite aplicar estadísticos de concentración como el WOE, con los que es posible hacer una selección inicial de variables potencialmente predictoras. Un criterio aplicable es seleccionar aquellas variables que tengan un WOE mayor a 0.02, aunque este valor tampoco es una regla fija. ✖
- [ ] f. Ninguna de las restantes opciones es cierta

Las respuestas correctas son: La tramificación de variables numéricas sirve para poder aplicar los estadísticos de concentración (Valor de Información y/o Gini) utilizados para ver la importancia que pueden tener las potenciales variables predictoras antes de incorporarlas al modelo de clasificación propiamente dicho, La tramificación de variables numéricas permite captar no-linealidades, aunque esto no sería necesario si se utilizasen modelos predictivos que no fuesen esencialmente lineales. Por ejemplo, en lugar de utilizar la regresión logística utilizar modelos de redes neuronales

Pregunta **5**

Correcta

Se puntúa 1,00 sobre 1,00

Una de las fases más interesantes en la construcción de modelos de puntuación de riesgos es la que corresponde con la tramificación de variables numéricas y la agrupación de niveles o categorías de las variables categóricas. Esta fase es anterior a la estimación del modelo de clasificación o modelo de probabilidad propiamente dicho.

Según la práctica de riesgos que vimos en clase la librería `optbinning` permite realizar este primer análisis.

Considere el conjunto de datos de créditos alemanes del fichero `germancredit.csv` utilizado en clase que preparamos que dividimos entre muestra de entrenamiento y de test siguiendo las siguientes instrucciones:

```
# Importamos la librería pandas y optbinning
>>> import pandas as pd
>>> from sklearn.model_selection import train_test_split
>>> from optbinning import Scorecard, BinningProcess, OptimalBinning

# Imnportamos el conjunto de daros "germacredit"
>>> dt=pd.read_csv('germancredit.csv')

#Recodifico esta variable creditability (variable objetivo) para que sea binaria
>>> dt["y"]=0
>>> dt.loc[dt["creditability"]=="good",["y"]]=0
>>> dt.loc[dt["creditability"]=="bad", ["y"]]=1
>>> dt.drop(labels='creditability',inplace=True, axis=1)

# Creo la muestra de entrenamiento y de test
>>> dt_train, dt_test = train_test_split(dt, stratify= dt["y"], test_size=.25, random_state=1234)
```

A continuación realizamos la tramificación óptima de la **variable edad (`age,in,years`)**

```
# Realizamos la trimificación optima de age.in.years
>>> variable="age.in.years"
>>> X=dt_train[variable].values
>>> Y=dt_train['y'].values
>>> optb = OptimalBinning(name=variable, dtype="numerical", solver="cp")
>>> optb.fit(X, Y)
>>> optb.splits
>>> binning_table = optb.binning_table
>>> binning_table.build()
```

Según los resultados obtenidos aplicando este código, marque **todas** las afirmaciones que considere que **son correctas**

Seleccione una o más de una:

- a. Se han creado siete tramos de edad (sin considerar ni los `Missing` ni `Special`) con puntos de corte en [25.5, 29.5, 33.5, 35.5, 38.5, 44.5]
- b. Se han creado siete tramos de edad (sin considerar ni los `Missing` ni `Special`) con puntos de corte en [25.5, 29.5, 34.5, 36.5, 38.5, 41.5] ✔
- c. Según los datos de la tabla, los solicitantes más jóvenes tienen una mayor propensión a ser *buenos clientes* que los mayores
- d. La variable Edad es una buena candidata para formar parte del modelo, porque el Valor de la información es 0.227806
- e. Hay 146 individuos entre 25.50 y 29.50 años (que suponen aproximadamente el 19,47% de la muestra de entrenamiento). En esta franja de edad hay 100 buenos clientes y 46 malos que hicieron impago por los que la tasa de malos clientes es del 31.51%, muy próximo al 30% de malos clientes en la muestra total de entrenamiento, por lo que el WOE de esta categoría debe estar próximo a cero (-0.070769) ✔
- f. Ninguna de las restantes opciones es cierta

Las respuestas correctas son: Se han creado siete tramos de edad (sin considerar ni los `Missing` ni `Special`) con puntos de corte en [25.5, 29.5, 34.5, 36.5, 38.5, 41.5], Hay 146 individuos entre 25.50 y 29.50 años (que suponen aproximadamente el 19,47% de la

muestra de entrenamiento). En esta franja de edad hay 100 buenos clientes y 46 malos que hicieron impago por los que la tasa de malos clientes es del 31.51%, muy próximo al 30% de malos clientes en la muestra total de entrenamiento, por lo que el WOE de esta categoría debe estar próximo a cero (-0.070769)

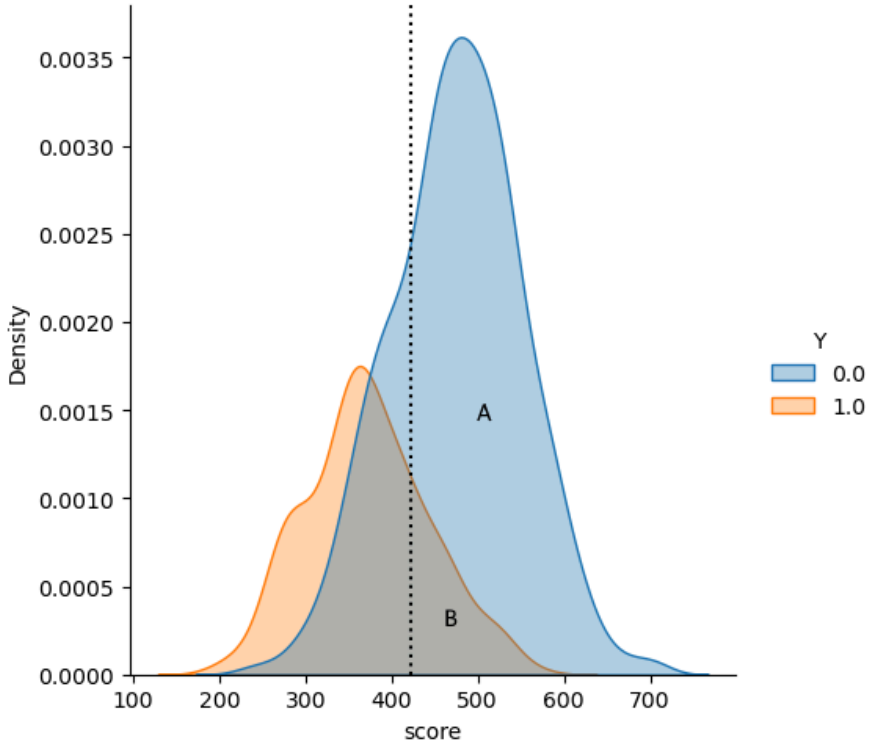
Pregunta **6**

Parcialmente correcta

Se puntúa 0,50 sobre 1,00

Suponga que una vez seleccionadas las variables predictoras que resultan más importantes ($IV > 0.02$), se ha estimado un modelo de regresión logística y a partir de las probabilidades estimadas para un conjunto de clientes de la muestra de test (algunos que hicieron Y=1, y otros que no, Y=0) se ha estimado su puntuación crediticia (*score*). Cuanto mayor es este *score* mejor es la calidad crediticia.

A continuación mostramos la distribución de frecuencias del grupo de malos clientes (Y=1) en naranja, y la distribución de buenos clientes (Y=0) en azul, según el *score* obtenido para cada uno. La linea vertical de puntos indica el punto de corte seleccionado para conceder o no los préstamos (para hacer los pronósticos de impago), que en este caso resulta ser de **423.14**.



Marque por favor **todas** las opciones que considere **correctas**

Seleccione una o más de una:

- a. El área **naranja** marcada con una **B** en el gráfico corresponde con los malos clientes a los que el modelo pronostica que serán buenos clientes, por lo que suponen un error de pronóstico del modelo (Falsos buenos clientes) ✔
- b. El cociente A/(A+B) es una medida de la **sensibilidad** (o sensibility o recall en inglés) del modelo para captar a los malos clientes ✖
- c. Todos los clientes que hayan obtenido un *score* menor que **423.14** tendrán una buena calidad crediticia, con una probabilidad relativamente baja de cometer impago, y por tanto todos esos clientes deberían ser pronosticados como buenos clientes (y concederles el crédito).
- d. El área **azul** marcada con una **A** en el gráfico corresponde con los buenos clientes a los que el modelo pronostica que efectivamente serán buenos clientes ✔
- e. El cociente A/(A+B) es una medida de la **precisión** del modelo para captar a los buenos clientes
- f. El cociente A/(A+B) es una medida de la **especificidad** del modelo para captar a los buenos clientes
- g. Ninguna de las restantes opciones es cierta
- h. El modelo no consigue separar totalmente a los buenos de los malos clientes, pero al menos a los buenos clientes sí que consigue asignarles, en media, mayor *score* que o los malos clientes ✔

Las respuestas correctas son: El área **naranja** marcada con una **B** en el gráfico corresponde con los malos clientes a los que el modelo pronostica que serán buenos clientes, por lo que suponen un error de pronóstico del modelo (Falsos buenos clientes), El área **azul** marcada con una **A** en el gráfico corresponde con los buenos clientes a los que el modelo pronostica que efectivamente serán

buenos clientes, El cociente A/(A+B) es una medida de la **precisión** del modelo para captar a los buenos clientes, El modelo no consigue separar totalmente a los buenos de los malos clientes, pero al menos a los buenos clientes sí que consigue asignarles, en media, mayor *score* que o los malos clientes

Pregunta **7**

Correcta

Se puntúa 1,00 sobre 1,00

La **inferencia de denegados** es una fase fundamental del análisis para evitar que en nuestro scorecard aparezcan sesgos de selección. Por favor, revise bien el código que utilizamos en clase para la práctica de riesgos utilizando los datos de créditos alemanes "germancredit" e indique cuál o cuáles de las siguientes afirmaciones es correcta.

Seleccione una o más de una:

- a. En los modelos de admisión de clientes nunca es necesario hacer inferencia de denegados cuando los datos provienen de clientes alemanes como en este caso (`germancredit`)
- b. En los departamentos de riesgos, a la hora de construir un modelo de puntuación de riesgo de admisión, se utilizan los datos de todos los clientes a los que se concedió en el pasado algún crédito similar al que se está evaluando. No es necesario inferir nada, ni rechazados ni aceptados, porque se utilizan todos los datos disponibles y siempre se observa si fueron buenos o malos todos los clientes que solicitaron una tarjeta de crédito.
- c. A la hora de construir un modelo de puntuación de riesgo de admisión, hay que analizar bien la muestra de datos disponibles. Cuando en la base de datos no hay información de denegados debe utilizarse sólo la información de los aceptados y no existe ningún riesgo de cometer ningún sesgo de selección muestral ya que se utiliza toda la información disponible (la de los clientes aceptados)
- d. A la hora de construir un modelo de puntuación de riesgo de admisión, hay que analizar bien la muestra de datos disponibles. Cuando la base de datos disponible para entrenar el modelo no es representativa de toda la población sobre la que posteriormente se quiere aplicar dicho modelo, entonces existe el riesgo de cometer un sesgo de selección muestral. Por eso, cuando sólo se dispone de información sobre el impago de los clientes aceptados es necesario hacer inferencia de denegados ✔
- e. Da un poco igual hacer inferencia de denegados, los resultados nunca cambian demasiado cuando se incluye este análisis de los denegados

La inferencia de denegados intenta solucionar el problema del sesgo de selección muestral, esto es, que la muestra utilizada para entrenar el modelo no es representativa de la población sobre la que se quiere aplicar dicho modelo

La respuesta correcta es: A la hora de construir un modelo de puntuación de riesgo de admisión, hay que analizar bien la muestra de datos disponibles. Cuando la base de datos disponible para entrenar el modelo no es representativa de toda la población sobre la que posteriormente se quiere aplicar dicho modelo, entonces existe el riesgo de cometer un sesgo de selección muestral. Por eso, cuando sólo se dispone de información sobre el impago de los clientes aceptados es necesario hacer inferencia de denegados

Pregunta **8**

Correcta

Se puntúa 1,00 sobre 1,00

Considere que tiene que construir un modelo de puntuación del riesgo de crédito para nuevos clientes que solicitan una tarjeta de crédito (riesgo de admisión). Los datos para realizar dicho modelo de puntuación están en el fichero excel `DatosPráctica_Scoring.xlsx` disponible en la carpeta de archivos de clase. En este fichero excel `DatosPráctica_Scoringxlsx` tenéis la siguiente información (también la tenéis en la hoja **Descripción datos** del fichero excel):

- Default: Variable objetivo, indica si un cliente ha hecho impago durante la ventana de observación. Default = 1 if defaulted 0 if not (observed only when Cardhldr = 1)
- Cardhldr=1: Clientes aceptados, son clientes a los que se le ha dado la tarjeta de crédito (Cardhldr=1) y sobre los que sabemos si han impagado alguna vez o no (default= 1 o 0 respectivamente).
- Cardhldr=0: clientes rechazados, son clientes que solicitaron un crédito pero a los que se negó su solicitud (Cardhldr=0), son clientes a los que no se les concedió la tarjeta, y por tanto no sabemos si hubieran impagado o no (default= na) porque no se les concedió la tarjeta de crédito.
- (Cardhldr=na): solicitud de nuevos clientes que solicitan una tarjeta de créditos y que hay que puntuar. Estos clientes están al final del archivo, son los 34 nuevos potenciales clientes (Cardhldr=na) con identificador de cliente desde 1286 hasta 1319, que están solicitando una tarjeta de crédito. Vosotros tendréis que puntuar su calidad crediticia y concederles o no el préstamo.

Para construir vuestro modelo tenéis las siguientes posibles variables explicativas o predictoras del riesgo:

- ID: identificador de cada solicitante de tarjeta
- Age = Age n years plus twelfths of a year
- Income = Yearly income (divided by 10,000)
- Exp_Inc = Ratio of monthly credit card expenditure to yearly income
- Avgexp = Average monthly credit card expenditure
- Ownrent = 1 if owns their home, 0 if rent
- Selfempl = 1 if self employed, 0 if not.
- Depndt = number of dependents (personas a cargo)
- Inc_per = Income divided by (1+number of dependents)
- Cur_add = months living at current address
- Major = number of major credit cards held
- Active = number of active credit accounts

Suponga que toma todos los clientes **Aceptados** y hace un primer análisis exploratorio **sin depurar**, **ni transformar** ninguna de las variables explicativas, y utilizando todas las observaciones del fichero (los 994 clientes aceptados, esto es sin dividir la muestra para entrenamiento y test).

Por favor indique a continuación cual es el Valor de la información de cada una de esas variables potencialmente predictoras

a. Age 0,2249769257736427 ✔

b. Income 0,22299861273580315 ✔

c. Exp_Inc 0,18713394146082002 ✔

d. Avgexp 0,27166297141159035 ✔

e. Ownrent 0,01197553855764489 ✔

f. Selfempl 0,011894593992974734 ✔

g. Depndt 0,00512071890004416 ✔

h. Inc_per 0,1562402432592817 ✔

i. Cur_add 0,1928572971306474 ✔

j. Major 0,03055742558083851 ✔

k. Active 0,8041765978543565 ✔

Pregunta **9**

Sin contestar

Se puntúa como 0 sobre 1,00

Considere que tiene que construir un modelo de puntuación del riesgo de crédito para nuevos clientes que solicitan una tarjeta de crédito (riesgo de admisión). Los datos para realizar dicho modelo de puntuación están en el fichero excel `DatosPráctica_Scoring.xlsx` disponible en la carpeta de archivos de clase En este fichero excel `DatosPráctica_Scoringxlsx` tenéis la siguiente información (también la tenéis en la hoja **Descripción datos** del fichero excel):

- Default: Variable objetivo, indica si un cliente ha hecho impago durante la ventana de observación. Default = 1 if defaulted 0 if not (observed only when Cardhldr = 1)
- Cardhldr=1: Clientes aceptados, son clientes a los que se le ha dado la tarjeta de crédito (Cardhldr=1) y sobre los que sabemos si han impagado alguna vez o no (default= 1 o 0 respectivamente).
- Cardhldr=0: clientes rechazados, son clientes que solicitaron un crédito pero a los que se negó su solicitud (Cardhldr=0), son clientes a los que no se les concedió la tarjeta, y por tanto no sabemos si hubieran impagado o no (default= na) porque no se les concedió la tarjeta de crédito.
- (Cardhldr=na): solicitud de nuevos clientes que solicitan una tarjeta de créditos y que hay que puntuar. Estos clientes están al final del archivo, son los 34 nuevos potenciales clientes (Cardhldr=na) con identificador de cliente desde 1286 hasta 1319, que están solicitando una tarjeta de crédito. Vosotros tendréis que puntuar su calidad crediticia y concederles o no el préstamo.

Para construir vuestro modelo tenéis las siguientes posibles variables explicativas o predictoras del riesgo:

- ID: identificador de cada solicitante de tarjeta
- Age = Age n years plus twelfths of a year
- Income = Yearly income (divided by 10,000)
- Exp_Inc = Ratio of monthly credit card expenditure to yearly income
- Avgexp = Average monthly credit card expenditure
- Ownrent = 1 if owns their home, 0 if rent
- Selfempl = 1 if self employed, 0 if not.
- Depndt = number of dependents (personas a cargo)
- Inc_per = Income divided by (1+number of dependents)
- Cur_add = months living at current address
- Major = number of major credit cards held
- Active = number of active credit accounts

Con toda esta información, y considerando tanto a los clientes **Aceptados**, como a los **Rechazados** (haciendo **inferencia de rechazados**), realice un pronóstico para las solicitudes de los nuevos clientes (los clientes con un ID desde el 1286 hasta 1319). Se recomienda realizar un análisis exploratorio previo para depurar los datos que se van a utilizar en la estimación de los modelos (**depuración** de datos atípicos, variables con poca variabilidad, transformación de variables), y dividir la muestra para entrenamiento y test para validar sus modelos).

Por favor indique a continuación a cuales de las siguientes 25 nuevas solicitudes **SÍ concede el crédito** (marque todas las solicitudes a las que concedería el crédito)

Seleccione una o más de una:

- a. ID: 1286
- b. ID: 1287
- c. ID: 1288
- d. ID: 1289
- e. ID: 1295
- f. ID: 1297
- g. ID: 1298
- h. ID: 1299
- i. ID: 1300
- j. ID: 1301
- k. ID: 1302

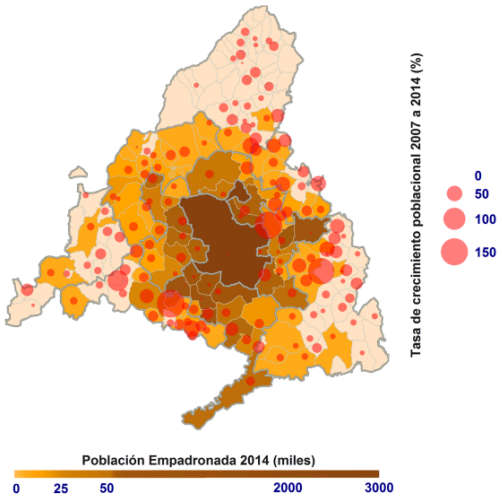
- l. ID: 1303
- m. ID: 1304
- n. ID: 1305
- o. ID: 1306
- p. ID: 1307
- q. ID: 1308
- r. ID: 1309
- s. ID: 1310
- t. ID: 1311
- u. ID: 1312
- v. ID: 1316
- w. ID: 1317
- x. ID: 1318
- y. ID: 1319

Las respuestas correctas son: ID: 1286, ID: 1288, ID: 1289, ID: 1295, ID: 1297, ID: 1298, ID: 1299, ID: 1300, ID: 1302, ID: 1303, ID: 1305, ID: 1308, ID: 1309, ID: 1311, ID: 1312, ID: 1316, ID: 1317, ID: 1319

Pregunta **10**

Parcialmente correcta

Se puntúa 0,86 sobre 1,00



Teniendo en cuenta los principales objetivos del análisis de la **Estadística Espacial**, indique cuál de las siguientes aplicaciones puede considerarse que forman parte de las aplicaciones de la Estadística Espacial (marque **todas** las opciones que considere necesarias)

Seleccione una o más de una:

- a. Detección de heterocedasticiad en un modelo de regresión para detectar fraude en tarjetas de crédito
- b. Representación de la distribución espacial de los precios de la vivienda en Madrid ✔
- c. Medición de áreas y distancias ✔
- d. Calcular la ruta óptima para llegar al trabajo
- e. Análisis de la distribución de las células cancerígenas en un tejido humano ✔
- f. Búsqueda de patrones de distribución espacial ✔
- g. `Krigeado` de las series de precios de la vivienda en la ciudad de Madrid ✔
- h. Estimación de modelos econométricos como el `Durbin spatial model` o el `spatial error model` ✔

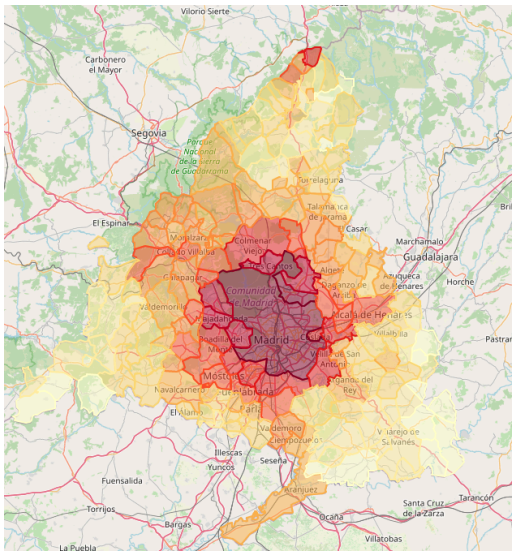
La estadística espacial hace referencia al análisis estadístico de datos espaciales, esto es, datos que tienen asociados una localización, una posición en el espacio. Cuando la referencia de posición es geográfica serán datos geoespaciales (datos georeferenciados), pero no necesariamente el análisis espacial tiene que ver con datos geoespaciales. Por tanto todos los análisis anteriores que tengan una referencia espacial podrían considerarse como integrantes de la estadística espacial.

Las respuestas correctas son: Representación de la distribución espacial de los precios de la vivienda en Madrid, Medición de áreas y distancias, Calcular la ruta óptima para llegar al trabajo, Análisis de la distribución de las células cancerígenas en un tejido humano, Búsqueda de patrones de distribución espacial, `Krigeado` de las series de precios de la vivienda en la ciudad de Madrid, Estimación de modelos econométricos como el `Durbin spatial model` o el `spatial error model`

Pregunta **11**

Correcta

Se puntúa 1,00 sobre 1,00



Utilizando la cartografía sobre municipios de España **Munic04_ESP** proporcionada en clase, el siguiente código sirve para explorar la distribución de los precios medios de la vivienda (***PrecioIn16***) de los municipios de la Comunidad de Madrid

```
>>>import geopandas as gpd

>>>gdfm =gpd.read_file("cartografias/Munic04_ESP.shp")
>>>gdfm_Madrid =gdfm[gdfm['COD_PROV']=='28']


>>>gdfm_Madrid.explore(column='PrecioIn16',
>>>                    scheme='NaturalBreaks',
>>>                    k=9, cmap='YlOrRd',
>>>                    legend=False,
>>>                    style_kwds=dict(fillOpacity=0.8))
```

Marque **todas** las opciones que considere correctas

Seleccione una o más de una:

- a. Lo siento pero el código no proporciona información de la Comunidad de Madrid sino de toda España
- b. Después del Municipio de Madrid (que tiene un precio medio de la vivienda *PrecioIn16*= 2923.45€), **Pozuelo de Alarcón** es el municipio con la vivienda más cara
- c. Después del Municipio de Madrid (que tiene un precio medio de la vivienda *PrecioIn16*= 2923.45€), **San Sebastián de los Reyes** es el municipio con la vivienda más cara
- d. Después del Municipio de Madrid (que tiene un precio medio de la vivienda *PrecioIn16*= 2923.45€), **Alcobendas** es el municipio con la vivienda más cara ✔
- e. Fuera del área metropolitana de la capital, el municipio más caro de la Sierra es **San Lorenzo del Escorial**
- f. Fuera del área metropolitana de la capital, el municipio más caro de la Sierra es **Navacerrada**
- g. Fuera del área metropolitana de la capital, el municipio más caro de la Sierra es **Somosierra** ✔
- h. Debe haber un error en el mapa, porque una parte del Municipio de **Santa María de la Alameda** se representa en la provincia de Segovia, junto al Espinar

Las respuestas correctas son: Después del Municipio de Madrid (que tiene un precio medio de la vivienda *PrecioIn16*= 2923.45€), **Alcobendas** es el municipio con la vivienda más cara, Fuera del área metropolitana de la capital, el municipio más caro de la Sierra es **Somosierra**

Pregunta **12**

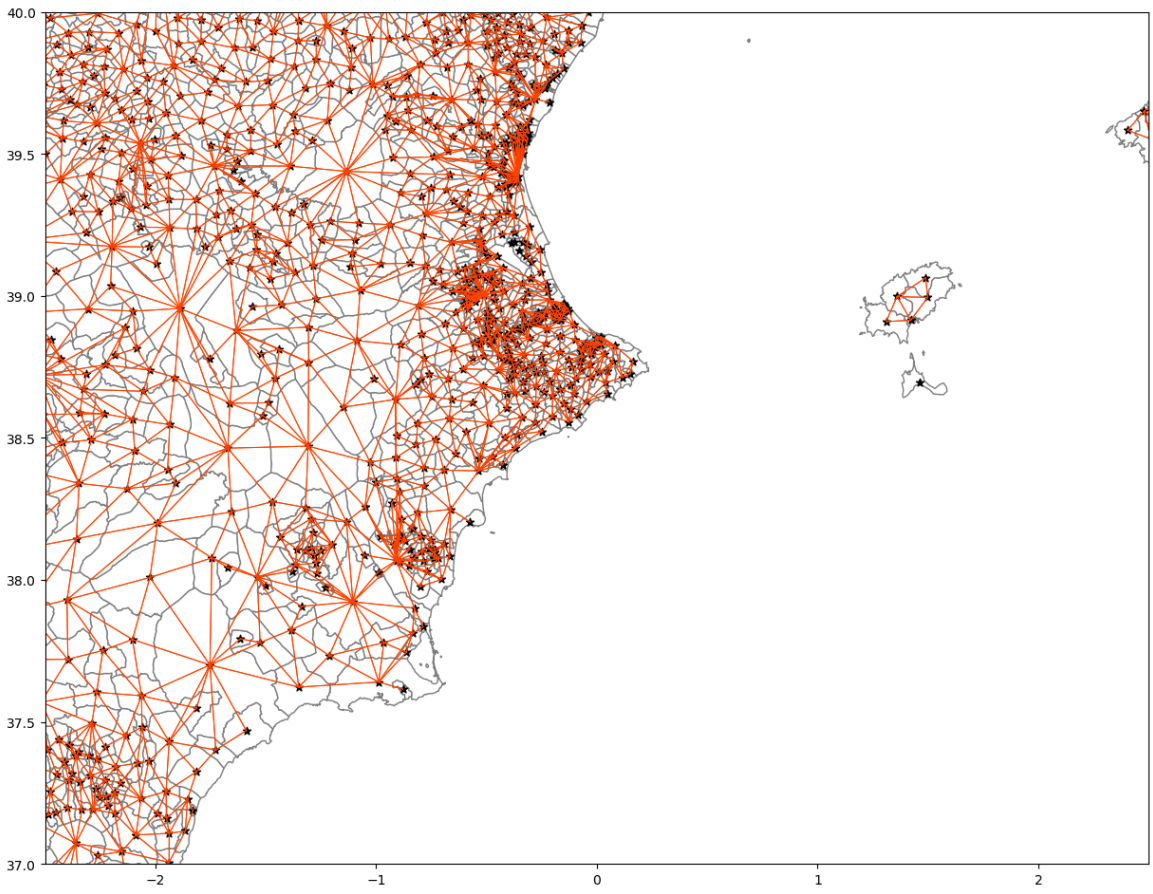
Correcta

Se puntúa 1,00 sobre 1,00

Para estimar retardos espaciales, correlaciones espaciales y regresiones espaciales es necesario calcular la matriz de peso espaciales, matriz de contigüidades o matriz de vecindades.

Atendiendo al código siguiente, y leyendo la cartografía de municipios de España de la práctica de clase `Munic04_ESP` ¿Qué tipo de matriz es la que se ha construido y representado en la figura?

```
>>> import geopandas as gpd
>>> import matplotlib.pyplot as plt
>>> from pysal.lib import weights


>>> gdfm =gpd.read_file("cartografias/Munic04_ESP.shp")

>>> wq = weights.contiguity.Queen.from_dataframe(gdfm)

>>> ax = gdfm.plot(
>>>         edgecolor="grey", facecolor="w",
>>>         figsize=(15,15))

>>> plt.xlim(-2.5,2.5)
>>> plt.ylim(37,40)

>>> wq.plot(
>>>     gdfm,
>>>     edge_kws=dict(linewidth=0.75, color="orangered"),
>>>     node_kws=dict(marker="*"),
>>>     ax=ax)

>>> plt.show()
```



Marque todas las opciones que considere que **son correctas**

Seleccione una o más de una:

- a. Una matriz de contigüidades tipo **torre**
- b. Una matriz de pesos tipo **alfil**
- c. Una matriz de pesos por contigüidades tipo **reina** ✔
- d. Hay algún error en el código. Ese script no representa ese mapa, representa un mapa de Cataluña
- e. Una matriz de pesos por **k-vecinos más próximos**
- f. Una matriz de pesos por **distancias** (inversamente proporcional a la distania)
- g. Ninguna de las restantes opciones es correcta

La respuesta correcta es: Una matriz de pesos por contigüidades tipo **reina**

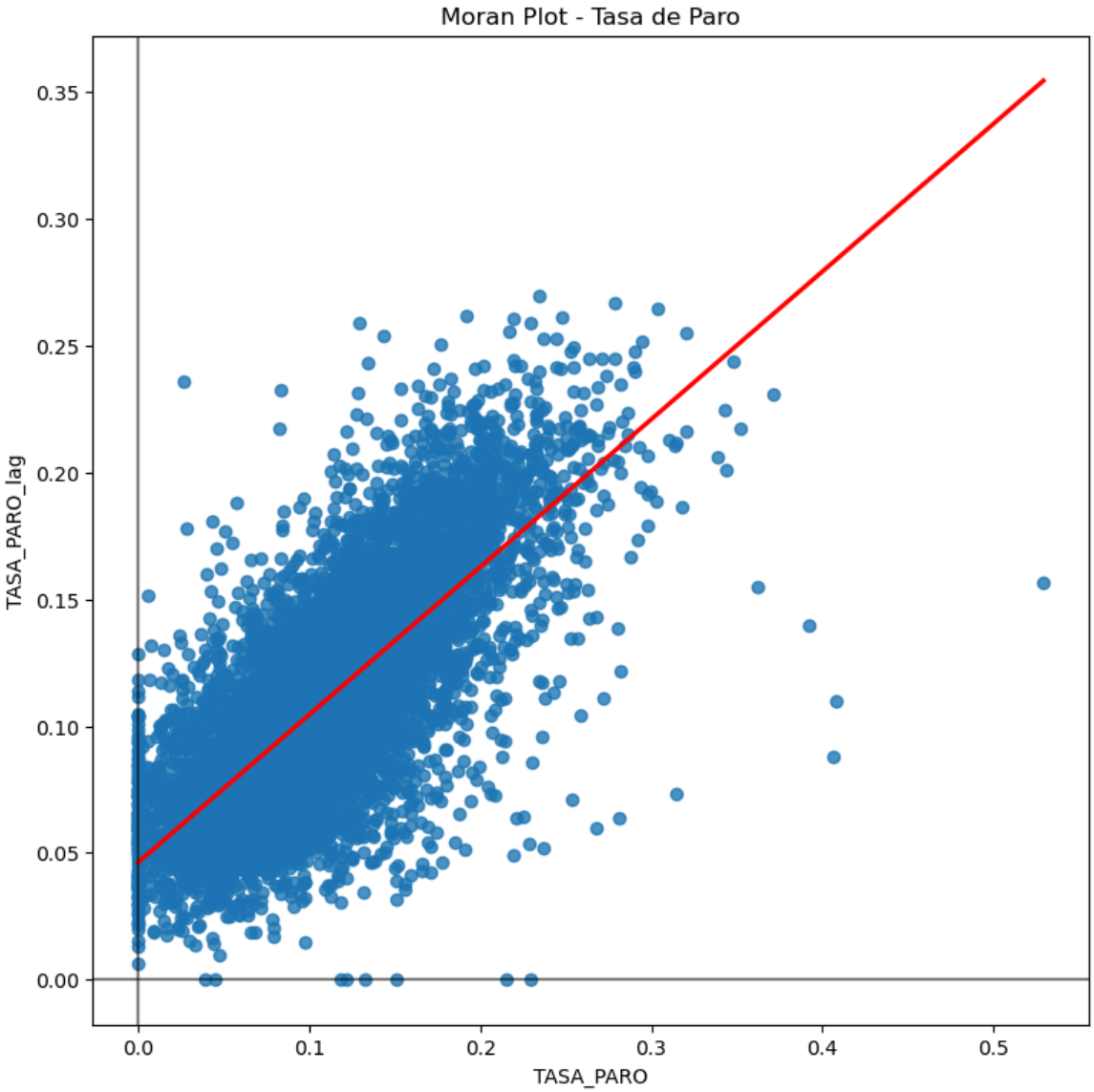
Pregunta **13**

Parcialmente correcta

Se puntúa 0,67 sobre 1,00

Utilizando la cartografía de municipios de España de la práctica de clase `Munic04_ESP` se ha estimado el siguiente `Gráfico de MORAN` para la *Tasa de paro municipal*

```
>>> import geopandas as gpd
>>> import matplotlib.pyplot as plt
>>> import seaborn as sns
>>> from pysal.lib import weights
>>> from pysal.explore import esda


>>> gdfm =gpd.read_file("cartografias/Munic04_ESP.shp")

>>> wq = weights.contiguity.Queen.from_dataframe(gdfm)
>>> wq.transform = "R"

>>> gdfm["TASA_PARO_lag"] = weights.spatial_lag.lag_spatial(wq, gdfm["TASA_PARO"])


>>> f, ax = plt.subplots(1, figsize=(9, 9))
>>> sns.regplot(
>>>     x="TASA_PARO",
>>>     y="TASA_PARO_lag",
>>>     ci=None,
>>>     data=gdfm,
>>>     line_kws={"color": "r"})
>>> ax.axvline(0, c="k", alpha=0.5)
>>> ax.axhline(0, c="k", alpha=0.5)
>>> ax.set_title("Moran Plot - Tasa de Paro")
>>> plt.show()
```



suponga que se realiza también el test de Moran global

```
>>> moran = esda.moran.Moran(gdfm["TASA_PARO"], wq)
```

Marque a continuación todas las opciones que considere que **son correctas**

Seleccione una o más de una:

- a. El gráfico de Moran muestra un gráfico de dispersión entre la tasa de paro municipal y la tasa de renta percápita (ley de okun)
- b. El gráfico de Moran muestra evidencia de heterogeneidad espacial (posiblemente tendremos un problema de heterocedasticidad si hacemos un modelo de regresión) ✖
- c. El gráfico de Moran muestra una autocorrelación espacial positiva ✔
- d. El gráfico de Moran muestra una autocorrelación espacial negativa
- e. El gráfico de Moran no muestra una clara autocorrelación espacial ni positiva ni negativa, cuestión ésta corroborada con la estimación de una I de Moran no significativa `p-valor=0.001`
- f. La I de Moran resulta ser `0.7396939`y estadístiamente significativa por lo que hay autocorrelación espacial (pendiente positiva del gráfico de Moran) ✖
- g. Este gráfico muestra evidencia de Autocorrelación espacial local significativa sólo para los Municipios en el cuadrante Low-Low y High-High

Ejecuta el test de Moran para completar la información que suministra el gráfico de Moran

La respuesta correcta es: El gráfico de Moran muestra una autocorrelación espacial positiva

Pregunta **14**

Correcta

Se puntúa 1,00 sobre 1,00

Utilizando la cartografía de municipios de España de la práctica de clase `Munic04_ESP` se han estimado los siguientes **DOS** modelos de regresión para la LEY DE OKUN (relación entre las *tasas de paro* y las *rentas per cápita* municipales)

MODELO (A)

$$Tasa\ Paro_i = \beta_0 + \beta_1 \cdot Renta\ Pc_i + u_i\ ,\ con\ u_i\ \sim\ R.B.$$

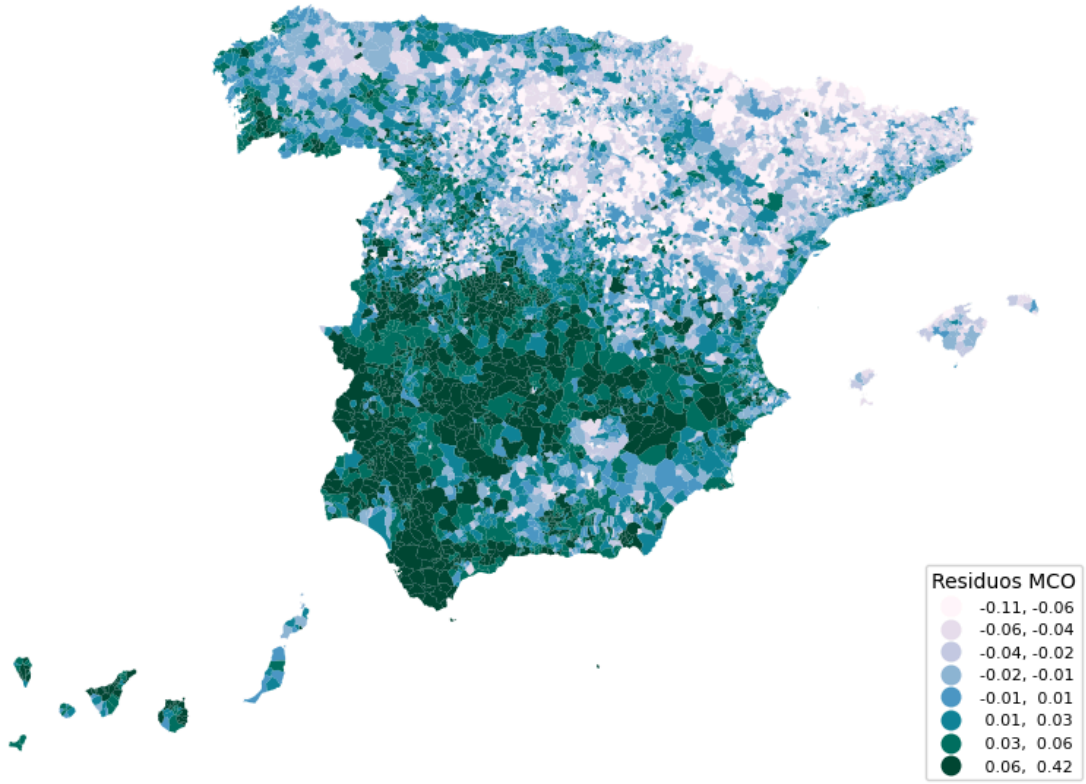
MODELO (B)

$$Tasa\ Paro_i = \beta_0 + \beta_1 \cdot Renta\ Pc_i + u_i$$

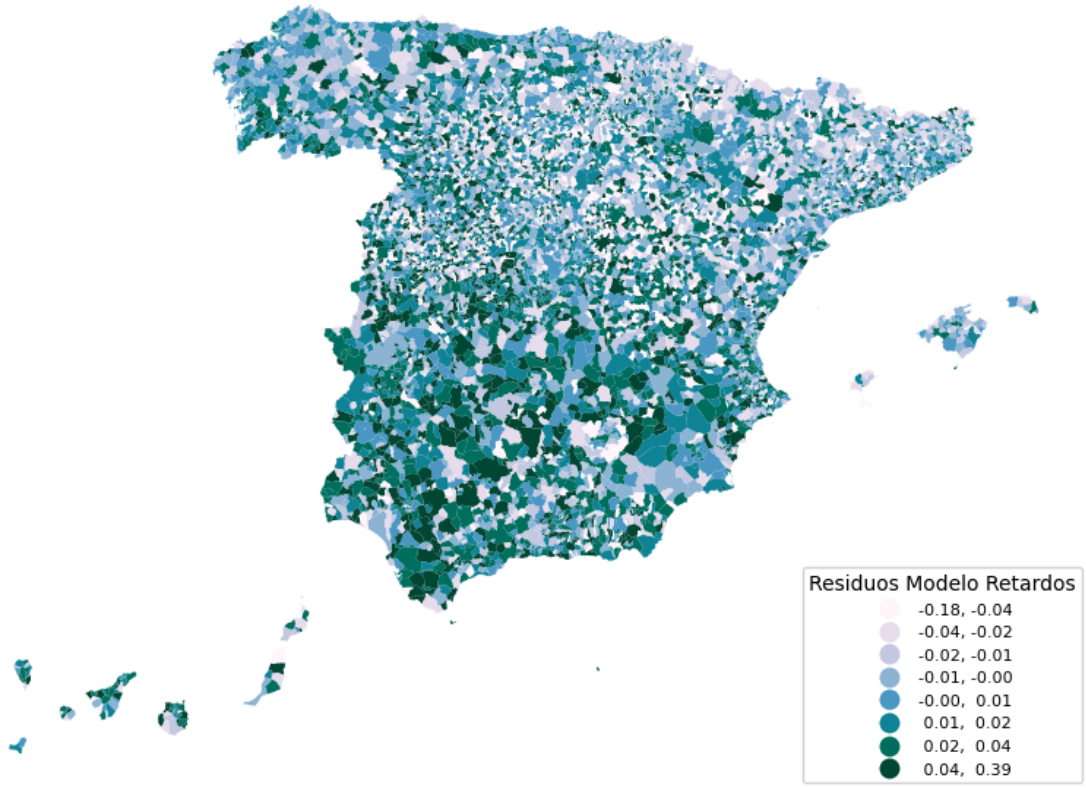
$$u_i = \lambda \cdot W \cdot \epsilon_i + \epsilon_i\ \ ,\ con\ \epsilon_i\ \sim\ R.B.$$

```
>>> import geopandas as gpd
>>> import matplotlib.pyplot as plt
>>> import seaborn as sns
>>> from pysal.lib import weights
>>> from pysal.explore import esda
>>> from pysal.model import spreg

>>> gdfm =gpd.read_file("cartografias/Munic04_ESP.shp")

>>> wq = weights.contiguity.Queen.from_dataframe(gdfm)
>>> wq.transform = "R"



# MODELO (A)

>>> modelo_A = spreg.OLS(
>>>     # Dependent variable
>>>     gdfm[["TASA_PARO"]].values,
>>>     # Independent variables
>>>     gdfm[["RENTPCAP07"]].values,
>>>     # Dependent variable name
>>>     name_y="TASA_PARO",
>>>     # Independent variable name
>>>     name_x=["RENTA_PERCAPITA"])

>>> gdfm["residual"] = modelo_A.u

>>> moran = esda.moran.Moran(gdfm["residual"], wq)
>>> print("I de moran:", moran.I.round(3))
>>> print("p-valor:", moran.p_sim)


# MODELO (B)

>>> modelo_B = spreg.GM_Error_Het(
>>>     # Dependent variable
>>>     gdfm[["TASA_PARO"]].values,
>>>     # Independent variables
>>>     gdfm[["RENTPCAP07"]].values,
>>>     # Spatial weights matrix
>>>     w=wq,
>>>     # Dependent variable name
>>>     name_y="TASA_PARO",
>>>     # Independent variable name
>>>     name_x=["RENTA_PERCAPITA"])

>>> gdfm["mLagresidual"] = modelo_B.e_filtered

>>> moran = esda.moran.Moran(gdfm["mLagresidual"], wq)
>>> print("I de moran:", moran.I.round(3))
>>> print("p-valor:", moran.p_sim)
```

El mapa de los residuos de este **MODELO (A)** es el siguiente



y El mapa de los residuos del **MODELO (B)** es este otro



Marque a continuación todas las opciones que considere que **son correctas**

Seleccione una o más de una:

- a. El **Modelo A** se ha estimado por mínimos cuadrados ordinarios y aunque pueda tener problemas de heterocedasticidad (no contrastados en el código mostrado), presenta problemas de autocorrelación espacial porque el **test de Moran** muestra un I de moran de `0.58` en los residuos y además el mapa de los residuos muestra concentraciones similares a la del mapa de la serie de tasa de paro ✔

b. El **Modelo A** se ha estimado por mínimos cuadrados ordinarios y aunque pueda tener problemas de heterocedasticidad (no contrastados en el código mostrado), presenta problemas de autocorrelación espacial porque el **test de Moran** muestra un I de moran de `0.86` en los residuos y además el mapa de los residuos muestra concentraciones similares a la del mapa de la serie de tasa de paro

c. El **Modelo B** es un modelo que incorpora un retardo espacial en el término de error (modelo de error espacial) y parece que recoge bien toda la estructura de dependencia espacial existente en la tasa de paro, porque el mapa de sus residuos parecen ruido blanco y el test de Moran de estos residuos proporciona una I de Moran de `-0.066` estadísticamente significativa (`p-value=0.001`) ✔

d. El **Modelo B** es un modelo que incorpora un retardo espacial de la Tasa de Paro y el coeficiente asociado a este retardo espacial ('0.7594137') es significativo y positivo. Además el mapa de sus residuos parecen un ruido blanco, y el test de Moran sobre sus residuos proporciona resultados de I de Moran de `-0.066` compatibles con esta hipótesis de ruido blanco

e. El **Modelo B** es un modelo que incorpora un retardo espacial en la Renta Percápita (modelo de Durbin) y parece que recoge bien la estructura de dependencia espacial en la Tasa de Paro, porque el mapa de sus residuos parecen un ruido blanco y el test de Moran de estos residuos proporciona resultados de I de Moran de `-0.066` compatibles con esta hipótesis de ruido blanco

f. Ninguna de las restantes opciones es correcta

Revisa los materialesde clase para saber qué tipo de modelos se está estimando y ejecuta los códigos en tu ordenador

Las respuestas correctas son: El **Modelo A** se ha estimado por mínimos cuadrados ordinarios y aunque pueda tener problemas de heterocedasticidad (no contrastados en el código mostrado), presenta problemas de autocorrelación espacial porque el **test de Moran** muestra un I de moran de `0.58` en los residuos y además el mapa de los residuos muestra concentraciones similares a la del mapa de la serie de tasa de paro, El **Modelo B** es un modelo que incorpora un retardo espacial en el término de error (modelo de error espacial) y parece que recoge bien toda la estructura de dependencia espacial existente en la tasa de paro, porque el mapa de sus residuos parecen ruido blanco y el test de Moran de estos residuos proporciona una I de Moran de `-0.066` estadísticamente significativa (`p-value=0.001`)

Pregunta **15**

Parcialmente correcta

Se puntúa 0,89 sobre 1,00

En esta pregunta vamos a trabajar con los datos de viviendas de la ciudad de Madrid tomados del artículo de José María Montero, Román Mínguez y Gema Fernández-Avilés (2018), Housing price prediction: parametric vs semiparametric spatial hedonic models J. Geogr Systems, vol 20, pp 27-55 (https://doi.org/10.1007/s10109-017-0257-y).

el objetico es cuantificar la posible autocorrelación espacial en el precio por metro cuadrado del "centro histórico" de Madrid.

Los datos están en el archivo `Data_Housing_Madrid.csv` disponible en la carpeta de archivos de clase. Las variables que nos van a interesar son:

- house.price: precio de la vivienda (euro/m2)
- historical: indica si una vivienda pertenece o no al casco histórico de Madrid
- longitude: longitud en coordenadas geográficas
- latitude: latitud en coordenadas geográficas

utilice las coordenadas geográficas para construir un data.frame espacial con geometría

```
# lectura del fichero de datos
>>> df = pd.read_csv('datos/Data_Housing_Madrid.csv')

#construcción de GeoDataFrame
>>>gdfm = gpd.GeoDataFrame(df, geometry=gpd.points_from_xy(df.longitude, df.latitude), crs='EPSG:4326')
```

Considere sólo las viviendas **del centro histórico** y construya una matriz de pesos por un método híbrido en el que se utiliza un radio máximo (por ejemplo 250 metros, más allá de esa distancia respecto a la vivienda objetivo se asume que el precio del resto de viviendas no ejercen influencia sobre la vivienda objetivo) y donde la importancias relativas de las viviendas que si cauen dentro del radio de influencia decrece con la distancia.

```
# Construcción de la matriz de pesos espaciales DistanceBand
>> w_hy = weights.distance.DistanceBand.from_dataframe(
    gdfm_historical, threshold = 0.00225, alpha=-1, binary=False)

# 250 m a la redonda son aproximadamente 0.00225 grados= 250/1000*(360/40000)
```

Conteste a las preguntas numéricas y marque de las tres últimas las respuestas que considere que son correctas

a. Indique el número total de viviendas en el dataset 10512 ✔

b. Indique el número de viviendas que pertenecen al "casco histórico" 3633 ✔

c. precio mediano de las viviendas en el "casco histórico" 4444 ✔

d. precio máximo de las viviendas en el "casco histórico" 20131 ✔

e. número de viviendas para las que no se han encontrado vecinas próximas 106 ✔

f. cual es el número mediano de viviendas próximas que se han encontrado para una vivienda en el "casco histórico" 14 ✖

g. Cúal es el valor de la I de Moran (redondeado a 3 dígitos decimales) 0,326 ✔

h. ¿Cúal es el pvalor del contraste de ausencia de autocorrelación espacial global por I de Moran? (redondeado a 3 dígitos decimales) 0,001 ✔

i.
- ☐ según estos resultados, analizar el precio de las viviendas vecinas NO ayuda a estimar el precio de una vivienda en el centro histórico de Madrid
- ☑ según estos resultados, analizar el precio de las viviendas vecinas Sí ayuda a estimar el precio de una vivienda en el centro histórico de Madrid ✔
- ☐ según estos resultados, no es posible determinar si el precio de las viviendas vecinas ayudan o no ayudan a estimar el precio de una vivienda en el centro histórico de Madrid

Se puntúa 1,00 sobre 1,00

La respuesta correcta es:

- según estos resultados, analizar el precio de las viviendas vecinas Sí ayuda a estimar el precio de una vivienda en el centro histórico de Madrid

◄ Vídeos Spacial

Ir a...

Encuesta valoración profesor Lorenzo Escot ►