2025/2026 1Q

Lista de Problemas 1

APA

Javier Béjar

Departament de Ciències de la Computació

Grau en Enginyeria Informàtica - UPC

UPC FIB Facultat d'Informàtica de Barcelona

UNIVERSITAT POLITÈCNICA DE CATALUNYA

DEPARTAMENT DE CIÉNCIES DE LA COMPUTACIÓ

FACULTAT D'INFORMÁTICA DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Primera edición, septiembre 2021

Esta edición, septiembre 2025

Instrucciones:

Para la entrega de grupo debéis elegir un problema del capítulo de problemas de grupo.

Para la entrega individual debéis elegir un problema del capítulo de problemas individuales.

Cada miembro del grupo debe elegir un problema individual diferente.

Debéis hacer la entrega subiendo la solución al racó.

Evaluación:

La nota de esta entrega se calculará como 1/3 de la nota del problema de grupo más 2/3 de la nota del problema individual.

Puntuación:

En esta primera lista, los errores leves se penalizarán con 0.5 puntos, los errores graves, sobre todo metodológicos se penalizarán con un punto. Explicaciones demasiado breves sobre lo que habéis hecho y los resultados también reducirán la nota. No hagáis una resolución mecánica de los problemas.

La penalización de los errores se doblará en la siguiente lista de problemas

Al realizar el informe correspondiente a los problemas explicad los resultados y las respuestas a las preguntas de la manera que os parezca necesaria. Se valorará más que uséis gráficas u otros elementos para ser más ilustrativos.

Entregad los resultados como un notebook (Colab/Jupyter). Podéis poner las respuestas a las preguntas en el notebook, este os permite insertar texto en markdown y en latex.

Aseguraos de que los notebooks mantienen la solución que habéis obtenido, no los entreguéis sin ejecutar.

Objetivos de aprendizaje:

1. Hacer un mínimo análisis exploratorio de un conjunto de datos
2. Hacer el preproceso de un conjunto de datos para usar regresión
3. Saber plantear problemas de regresión sencillos y resolverlos usando diferentes métodos
4. Interpretar los resultados de un problema de regresión

CAPÍTULO 1

Problemas de Grupo

Al resolver el problema explicad bien lo que hacéis y los resultados que obtenéis, no hacer ningún comentario o hacer comentarios superficiales tendrán una nota más baja.

Tenéis que mostrar que habéis entendido los métodos que estáis aplicando y como se utilizan, así que un corta y pega de problemas similares no es suficiente.

1. **Predicción del uso de bicicletas**

   El uso compartido de bicicletas es un servicio proporcionado por cualquier ciudad importante del mundo, por lo que comprender y predecir el comportamiento del sistema es un elemento clave. Vamos a trabajar con el conjunto de datos de bicicletas compartidas del repositorio de conjuntos de datos de UCI que recopila estadísticas agregadas de uso de bicicletas junto con otra información adicional relevante. Se pueden descargar los datos desde aquí `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`.

   El objetivo de este problema es predecir cuántas bicicletas se usarán al día siguiente a partir de los datos de días anteriores (usaremos el archivo `day.csv`). Podéis leer en el `Readme.txt` los detalles sobre las variables.

   a) El primer paso es preprocesar y preparar los datos antes de ajustar cualquier modelo. Hay algunas variables que no son útiles para el problema o que no tiene sentido usar. Eliminadlas del conjunto de datos y explicad por qué las eliminái s.

   Necesitaremos obtener variables que nos permitan predecir a partir de la historia del sistema. Tal como están los datos no podemos hacer eso, por lo que necesitaremos un poco de preproceso. La librería `pandas` permite generar una copia de una tabla de datos desplazada una serie de instantes temporales usando el método `shift`. Mirad como funciona y generad una copia de los datos desplazada un día y añadidla como nuevas columnas (fijaos que os saldrán datos perdidos ¿por qué?). Si queremos predecir el futuro habrá una serie de variables que no podemos saber. Partid la tabla de datos en las columnas que usaremos para predecir y las que podríamos predecir a partir de las otras. Fijaos que hay variables del día actual que sabemos, como el día de la semana que es.

Dado que tenemos que predecir el futuro, no podemos partir los datos en entrenamiento y test tal como lo hacemos habitualmente (explicad por qué no podemos hacerlo). Vamos a seleccionar los primeros 500 ejemplos para entrenamiento y el resto para test.

Haced una exploración mínima del conjunto de datos de entrenamiento observando las relaciones entre las variables, especialmente con las variables objetivo. Describid las cosas que hayáis visto que os parezcan interesantes. Estandarizad los datos antes de entrenar los modelos.

b) Dado que hay varias variables que podemos predecir para el siguiente día vamos a escoger `temp`, `hum`, `windspeed` y `cnt`. Ajustad una regresión lineal a los datos y calculad la *calidad* del modelo empleando validación cruzada y con los datos de test. Para hacer la validación cruzada tenéis que usar el método `TimeSeriesSplit` con 5 particiones en el parámetro `cv` de la validación. ¿Por qué no es válida la validación cruzada tradicional? Comentad los resultados obtenidos con cada una de las variables. Calculad el error de validación cruzada y el del test con el *mean absolute error* (tendréis que mirar la documentación de *scikit-learn* para ver como se hace)

c) Probablemente la regularización ayude a obtener mejores resultados. Usad Ridge regression y LASSO para predecir cada variable. Tendréis que usar el mismo método para hacer la validación cruzada. Comentad los resultados obtenidos con cada una de las variables

d) Se nos podría ocurrir que dado que podemos predecir algunas variables para el día siguiente, esta nos podría valer como substituto e introducirla en el modelo para poder predecir mejor la variable `cnt`. Añadid las predicciones de la variable que mejor se predice a los datos y ajustad de nuevo la regresión lineal, la Ridge Regression y el LASSO. Explicad lo que sucede y por qué.

e) Para entender el modelo tenemos que analizar con detalle los resultados. Representad las predicciones del mejor modelo para el test contra los valores reales, analizad el gráfico de los residuos, explicad los resultados. Mirad los pesos que le asigna el LASSO a las variables ¿hay algunas que son descartadas? Eliminad las variables que no considera relevantes LASSO y ajustad de nuevo una regresión lineal. Comparad los resultados.

Ajustad una regresión lineal con todas las variables usando el método `OLS` de la librería `statsmodels` (los pesos serán los mismos) y analizad la significatividad que asigna el método al coeficiente de cada variable. Eliminad las variables que no tienen una significatividad menor que 0.05 y repetid la regresión. Comparad y comentad los resultados de los dos modelos reducidos.

f) Hemos asumido que sabiendo los datos del día anterior era suficiente para predecir los del día siguiente. Ajustad el modelo LASSO a los datos con dos y tres días antes. Comparad los resultados de los modelos entre ellos y con el que solo usa el día anterior. Mirad las variables que descarta LASSO y explicad lo que observéis.

2. **Aire limpio, aire puro**

La contaminación del aire es algo serio en las grandes ciudades como Barcelona, el poder relacionarlo con otras variables puede ayudar a comprender mejor sus fuentes y las circunstancias que le afectan.

El ayuntamiento de Barcelona recolecta diversos datos sobre la ciudad en su portal de datos abiertos[1]. Vamos a trabajar con un extracto de esos datos para los años 2022-2024, eligiendo un subconjunto de variables que son medidas diarias que representan diferentes características que tienen alguna relación con la contaminación como el número de matriculaciones de vehículos, el

[1] https://portaldades.ajuntament.barcelona.cat/

número de personas que llegan a Barcelona desde ciudades cercanas, el precio de la electricidad, el volumen de tránsito e información meteorológica medida en diferentes puntos de la ciudad (temperatura, viento, precipitación). El objetivo es buscar la relación con la cantidad de dióxido de nitrógeno (NO2) medido en l'Eixample.

Podéis obtener estos datos mediante la función `load_BCN_NO2` de la librería `apafib`. Resolved los siguientes apartados ilustrando los resultados de la manera que os parezca más adecuada.

a) Dividid el conjunto de datos en entrenamiento y test (60 %/40 %). Haced una exploración mínima del conjunto de datos de entrenamiento observando las relaciones entre las variables, especialmente con la variable objetivo. Describid las cosas que hayáis visto que os parezcan interesantes. Transformad las variables adecuadamente para poder ajustar un modelo de regresión tanto el conjunto de entrenamiento como el de test.

b) Aplicad algún método de reducción de dimensionalidad a los datos de entrenamiento y comentad lo que se pueda apreciar en la visualización. Pensad en qué podéis representar sobre la transformación.

c) Ajustad una regresión lineal, una regresión Ridge y una regresión LASSO a los datos ¿Os parece suficientemente bueno el resultado? Representad los valores de la variable objetivo para el conjunto de test contra las predicciones y representad los residuos. ¿Qué modelo os parece mejor?

d) Habréis visto que las variables meteorológicas se toman en diferentes puntos de la ciudad, pero podéis comprobar que las medidas son bastante parecidas (comprobadlo y mostrad que es así adecuadamente). ¿Es posible reducir el número de observatorios a solo uno? ¿Cuál es el mejor? Comprobadlo con el modelo que haya dado el mejor resultado.

e) Otra manera de ver la relevancia de las variables en el modelo es comprobar la significatividad de los coeficientes de regresión. Ajustad una regresión lineal con todas las variables usando el método `OLS` de la librería `statsmodels` (los pesos serán los mismos que en la regresión lineal) y analizad la significatividad que asigna el método al coeficiente de cada variable. Explicad lo que habéis visto ¿hay alguna posible razón para que algunas variables no sean importantes? Eliminad las variables que no son significativas y ajustad de nuevo el mejor modelo. Comentad los resultados.

f) Si representáis las predicciones del mejor modelo contra los valores reales probablemente veréis que no todas las predicciones son homogéneas. Eliminad todas las variables que el modelo `OLS` considera no significativas y usad la función `PolynomialFeatures` de scikit-learn para añadir características al conjunto de datos que correspondan a polinomios de grado 2 para esas variables (usad *interaction_only=True* para que solo se tengan en cuenta las interacciones entre las variables). Ajustad de nuevo una regresión lineal y un modelo LASSO para estas variables y evalúad la calidad de los modelos. Representad las predicciones respecto a los valores reales y comentad lo que observáis.

3. **El precio de las cosas y más**

Asumimos que los precios de las cosas están ligados entre ellos de manera más o menos compleja, eso mueve todo lo que está relacionado con la economía, como por ejemplo la bolsa. A veces hay cosas que pueden estar relacionadas también de manera más o menos coherente y permite descubrir factores desconocidos pueden sorprendernos (o ser simplemente espurios).

El ayuntamiento de Barcelona recolecta diversos datos sobre la ciudad en su portal de datos abiertos[1]. Vamos a trabajar con un extracto de esos datos para los años 2022-2023, eligiendo un subconjunto de variables que tienen que ver con la economía, como el precio de productos de la cesta de la compra o el número de matriculaciones de vehículos y, por lo tanto, tendrán

alguna relación con el índice IBEX. De manera exploratoria añadiremos una serie de variables que no parecen tener relación y que si es el caso no tendrán un peso en el modelo, como son la temperatura medida en diferentes puntos de la ciudad y el nivel de ruido a diferentes horas del día.

Podéis obtener estos datos mediante la función `load_BCN_precios` de la librería `apafib`. Resolved los siguientes apartados ilustrando los resultados de la manera que os parezca más adecuada.

a) Dividid el conjunto de datos en entrenamiento y test (60 %/40 %). Haced una exploración mínima del conjunto de datos de entrenamiento observando las relaciones entre las variables, especialmente con la variable objetivo. Describid las cosas que hayáis visto que os parezcan interesantes. Transformad las variables adecuadamente para poder ajustar un modelo de regresión tanto el conjunto de entrenamiento como el de test

b) Aplicad algún método de reducción de dimensionalidad a los datos de entrenamiento y comentad lo que se pueda apreciar en la visualización. Pensad en qué podéis representar sobre la transformación.

c) Ajustad una regresión lineal, una regresión Ridge y una regresión LASSO a los datos ¿Os parece suficientemente bueno el resultado? Representad los valores de la variable objetivo para el conjunto de test contra las predicciones y representad los residuos. ¿Qué modelo os parece mejor?

d) Comprobad si el modelo LASSO identifica las variables espurias como no significativas. Eliminadlas del modelo y volved a ajustar el mejor modelo que os ha salido ¿Cuál es el mejor? Comentad los resultados.

e) Otra manera de ver la relevancia de las variables en el modelo es comprobar la significatividad de los coeficientes de regresión. Ajustad una regresión lineal con todas las variables usando el método `OLS` de la librería `statsmodels` (los pesos serán los mismos que en la regresión lineal) y analizad la significatividad que asigna el método al coeficiente de cada variable. Explicad lo que habéis visto ¿hay alguna posible razón para que algunas variables no sean importantes? Eliminad las variables que no son significativas y ajustad de nuevo el mejor modelo. Comentad los resultados.

f) A veces las interacciones entre las variables son importantes para obtener un mejor modelo. Partid del conjunto de datos del que habéis quitado las variables no significativas y usad la función `PolynomialFeatures` de scikit-learn para añadir características al conjunto de datos que correspondan a polinomios de grado 2 para esas variables. Ajustad de nuevo una regresión lineal y un modelo Ridge para estas variables y evaluad la calidad de los modelos. Representad las predicciones respecto a los valores reales y comentad lo que observáis.

CAPÍTULO 2

Problemas Individuales

Al resolver el problema explicad bien lo que hacéis y los resultados que obtenéis, no hacer ningún comentario o hacer comentarios superficiales tendrán una nota más baja.

Tenéis que mostrar que habéis entendido los métodos que estáis aplicando y como se utilizan, así que un corta y pega de problemas similares no es suficiente.

Para obtener los datos de algunos de estos problemas necesitaréis instalaros la última versión de la librería `apafib`. La podéis instalar localmente haciendo:

```
pip install --user --upgrade apafib
```

Para usar las funciones de carga de datos solo tenéis que añadir su importación desde la librería, en vuestro script o notebook, por ejemplo

```
from apafib import load_medical_costs
```

La función por lo general os retornará un `DataFrame` de Pandas con los datos. Si no es así el enunciado explicará que retorna.

1. **La medicina es cara**

   El coste de los seguros médicos varía bastante según las circunstancias de cada persona, pero a veces averiguar como se calcula realmente no es tan sencillo. El conjunto de datos `Medical Cost Personal Dataset`[1]. tiene la descripción de las características de un grupo de personas y los cargos de su seguro médico. Nos interesa predecir esta última variable (`charges`).

   Trabajaremos con una versión de este conjunto que podéis obtener mediante la función `load_medical_costs` de la librería `apafib`. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

[1] Tenéis información sobre sus atributos en https://www.kaggle.com/datasets/mirichoi0218/insurance

a) Divide el conjunto de datos en entrenamiento y test (70 %/30 %). Haz una exploración mínima del conjunto de datos de entrenamiento observando las relaciones entre las variables, especialmente con la variable objetivo. Describe las cosas que hayas visto que te parezcan interesantes. Aplica adecuadamente PCA a los datos y representa la variable respuesta en los dos primeros componentes. Explica lo que observes sobre los componentes calculados y la representación de los datos en dos dimensiones. Transforma las variables adecuadamente para poder ajustar un modelo de regresión tanto el conjunto de entrenamiento como el de test.

b) Ajusta un modelo de regresión lineal para predecir la variable objetivo usando la librería `scikit-learn` y estima la calidad de la regresión. ¿Te parece suficientemente bueno el resultado? Representa los residuos de la regresión y las predicciones contra los valores reales, y comenta que aparece. Ajusta la regresión con todos los datos de entrenamiento usando la librería `statsmodels`. Comenta la significatividad de los pesos.

c) La relación entre las variables del conjunto de datos y la variable objetivo podrían ser no lineal. Usa la función `PolynomialFeatures` de `scikit-learn` para añadir características al conjunto de datos que correspondan a polinomios de grado 2. Ajusta a estos nuevos datos un modelo de regresión lineal y comenta los resultados comparándolos con la regresión original.

d) Hemos añadido una gran cantidad de variables a los datos, eso quiere decir que el modelo será mucho más complejo. Aplica regresión LASSO para ver si es posible reducir el número de variables que intervienen en el modelo. Compara la calidad del modelo con los otros. Analiza la importancia de los pesos de este modelo ¿Qué importancia tienen las interacciones entre las variables?

2. **Barcelona motor del IBEX**

   La predicción bursátil es un problema complejo, pero a veces se pueden observar relaciones con variables que aparentemente no deberían influenciar. El portal de datos abiertos del ayuntamiento de Barcelona recoge informaciones diarias sobre la ciudad[4] y esto nos ofrece la oportunidad de averiguar si lo que pasa en Barcelona tiene alguna influencia en el mercado del IBEX. Vamos a trabajar con un extracto de esos datos para el año 2021, con un subconjunto de variables que hemos elegido según nuestro criterio *experto* para desentrañar esa influencia. El objetivo es aproximar el valor de la cotización del IBEX a partir de las otras variables.

   Puedes obtener estos datos mediante la función `load_BCN_IBEX` de la librería `apafib`. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

   a) Divide el conjunto de datos en entrenamiento y test (80 %/20 %). Haz una exploración mínima del conjunto de datos de entrenamiento observando las relaciones entre las variables, especialmente con la variable objetivo. Describe las cosas que hayas visto que te parezcan interesantes. Transforma las variables adecuadamente para poder ajustar un modelo de regresión tanto para el conjunto de entrenamiento como para el de test.

   b) Aplica Análisis de Componentes Principales (PCA) adecuadamente al conjunto de entrenamiento y visualízalo en 2D representando la variable objetivo. ¿Crees que puede haber una relación entre las variables del conjunto de datos y la variable objetivo? ¿Por qué?

   c) Ajusta una regresión lineal, una regresión Ridge y una regresión LASSO a los datos ¿Te parece suficientemente bueno el resultado? ¿Qué modelo te parece mejor? Representa los residuos de la mejor regresión y los valores de la variable objetivo del conjunto de test contra sus predicciones. ¿Tienen sentido las variables con más peso que aparecen en los modelos para la variable que queremos predecir? Elimina las variables que tienen menos

peso en los modelos del conjunto de datos y reajusta el modelo de regresión lineal ¿Cómo ha cambiado el peso de las variables que quedan?

d) Al ser un problema complejo, igual hay interacciones entre variables que explican mejor la variable objetivo. Usa la función `PolynomialFeatures` de scikit-learn para añadir al conjunto de datos original características que correspondan a polinomios de grado 2. Vuelve a ajustar la regresión Ridge y la regresión LASSO. ¿Han mejorado los modelos y los residuos? Fíjate en las variables a las que LASSO no les ha dado un peso 0. ¿Se corresponden con interacciones entre variables?

3. **Blowing in the wind**

Una forma sencilla de predecir series temporales es utilizar regresión lineal sobre un número de instantes temporales de la serie. Estaremos prediciendo el futuro en función de las observaciones pasadas en una ventana de tiempo, de manera que:

$$f(x_t) = c + \left[\sum_{i=1}^{p} w_{t-i} \cdot x_{t-i}\right] + \epsilon_t$$

Donde c es una constante y ϵ_t es ruido gausiano. Este modelo es denominado auto regresivo (AR).

El conjunto de datos `Wind Speed Prediction Dataset`[2] tiene mediciones de diferentes variables tomadas por una estación meteorológica durante 15 años. El objetivo es predecir el valor de la variable `WIND` usando ventanas de datos pasados.

Trabajaremos con una versión de este conjunto que podéis obtener mediante la función `load_wind_prediction` de la librería `apafib`. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

a) Primero haremos una limpieza de los datos. Verás que existen datos perdidos en alguna de las variables. Dado que son series temporales una manera sencilla de hacer la imputación es usar el valor del último instante válido de la serie. En este caso puedes usar la función `fillna` de Pandas usando el parámetro `method='ffill'`. Hay tres variables `IND` que no están descritas en el conjunto de datos, no sabiendo lo que significan lo mejor es eliminarlas. Tienes también la variable `DATE`, que de por sí no tiene mucha utilidad, pero podemos pensar que saber el mes del año podría ser útil para predecir el viento. Averigua como transformar con Pandas esa variable en formato `datetime` y como extraer el mes. Una vez obtenido ese valor puedes deshacerte de la variable.

Para generar los datos necesitaremos ventanas temporales de cierta longitud. `Pandas` permite generar una copia de una tabla de datos desplazada una serie de instantes temporales usando el método `shift`. Genera tres conjuntos de datos con longitud de ventana 2, 4 y 6 de manera que puedas predecir el viento de un dia a partir de las variables del día anterior, tres días y cinco días. Tendrás que descartar todas las columnas que corresponden al último instante temporal de manera que quede solo la variable `WIND` que es la que has de predecir. Elimina todos los valores perdidos que ha generado esta transformación.

La validación en series temporales no puede usar validación cruzada. Tendrás que generar un conjunto de entrenamiento, uno de validación y uno de test para evaluar los modelos. Parte el conjunto de datos de manera que tengas un 70 % de entrenamiento, 15 % de validación y 15 % de test. Elimina las últimas filas del conjunto de entrenamiento y validación para que no haya instantes temporales compartidos entre las particiones.

[2] Tenéis información sobre sus atributos en https://www.kaggle.com/datasets/fedesoriano/wind-speed-prediction-dataset.

b) La calidad de un modelo en series temporales se puede medir de diferentes maneras. Para este caso, usa el error absoluto medio (MAE). Esto tiene la ventaja de que el error esta en las unidades de la variable respuesta, en este caso *m/s*. Entrena regresiones lineales, Ridge y LASSO con los diferentes conjuntos de datos y compara su calidad y las características de los modelos. El ajuste de hiperparámetros no lo puedes hacer mediante validación cruzada, has de utilizar el conjunto de datos de validación. Selecciona el mejor modelo adecuadamente. Analiza los pesos de la regresión LASSO ¿es el mes relevante para la predicción? ¿Cómo se comportan los pesos respecto a la longitud de la ventana? Representa el qqplot de los residuos del mejor modelo y comprueba si son gausianos. Comenta los resultados.

c) Al predecir el viento nos interesa saber la incertidumbre de la predicción. La regresión cuantil[3] es un modelo que permite estimar el intervalo de predicción de una regresión. La regresión se realiza para que las predicciones estén por debajo de un cuantil de probabilidad determinado. Haciendo la regresión a diferentes cuantiles podemos tener el intervalo de valores posibles para una predicción. El cuantil 0.5 corresponde a la media de la predicción de la regresión. Ajusta una regresión cuantil para los cuantiles 0.1, 0.5 y 0.9. Tendrás que ajustar el peso de la regularización de esta regresión para obtener el mejor modelo usando la muestra de validación. Selecciona los primeros y últimos 100 instantes de la serie de test y representa las predicciones de los diferentes cuantiles. Calcula la media y varianza de la diferencia entre la predicción del cuantil 0.1 y 0.9 para estos dos intervalos para ver si hay una diferencia. Comenta los resultados.

Representa la predicción del mejor modelo para una pequeña ventana ($\approx$ 100) de datos de la muestra de test. ¿Crees que la regresión está haciendo una buena aproximación de la serie temporal? ¿Qué características debería cumplir esta serie para que la regresión lineal fuera un buen modelo para predecirla?

4. **Todo es incierto en este mundo, menos la muerte y los impuestos**

El centro de investigaciones sociológicas (CIS) es un organismo que tiene por finalidad el estudio científico de la sociedad y es una gran fuente de datos. Entre los diferentes estudios que se realizan anualmente hay uno sobre política fiscal preguntando entre otras cosas, la percepción que se tiene sobre el destino de los impuestos, el esfuerzo fiscal realizado y la tolerancia al fraude. En este problema vamos a analizar los resultados de esta encuesta para un año concreto intentando predecir la variable que corresponde a la edad del entevistado. La clave de las variables y su significado la puedes encontrar en la página web de los problemas.

Trabajaremos con una versión reducida de este conjunto que puedes obtener mediante la función `load_CIS_Impuestos` de la librería `apafib` esta retornará un dataframe de Pandas. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

a) Divide el conjunto de datos en entrenamiento y test (80 %/20 %). Comprobarás que la mayoría de las variables del conjunto de datos son categóricas. Explorando los datos también encontrarás que hay bastantes valores perdidos. En este caso los valores perdidos corresponden a valores *no sabe/no contesta* y en el caso de los valores categóricos esto puede ser información interesante. Haremos la imputación de manera diferente para categóricos y para numéricos. Para los categóricos, la función de Pandas que la calcula las columnas de la codificación *One Hot* permite indicar que se quiere una columna más para el valor perdido, lo haremos así. Para los numéricos, haremos una imputación usando el `KNNImputer` después de imputar los categóricos.

[3] Está implementada en `scikit learn` como `QuantileRegressor`.

Una vez tenemos el conjunto de datos preprocesado puedes aplicar adecuadamente PCA para ver su comportamiento. Analiza sus resultados y representa la variable objetivo sobre el PCA en dos dimensiones. Explica lo que has observado.

b) Transforma las variables adecuadamente para poder ajustar un modelo de regresión tanto el conjunto de entrenamiento como el de test. Ajusta un modelo de regresión lineal para predecir la variable objetivo y estima la calidad de la regresión. ¿Te parece suficientemente bueno el resultado? Analiza los pesos del modelo. Representa los residuos y las predicciones sobre el test contra los valores reales. Comenta qué aparece.

c) El número de variables que hay en el conjunto de datos es relativamente grande, así que podría haber riesgo de sobre especialización. Aplica regresión Ridge y LASSO al conjunto de datos. Compara los resultados. ¿Alguno de los modelos es mejor que la regresión original?

d) Una cosa interesante en estos estudios es analizar como las variables explican la variable objetivo. ¿Hay pesos prominentes en el modelo que tengan algún sentido para el cálculo de la edad? Explica la relación que podrían tener esas variables con la variable objetvo ¿Tienen los valores no sabe/no contesta alguna importancia en el modelo?

5. **Músicas del mundo**

Adivinar el lugar en el mundo en el que se ha tomado una fotografía es una tarea difícil, pero es aún más el averiguar de donde proviene una música. Vamos a trabajar con un conjunto de datos derivado de estos datos del repositorio de UCI[4]. En estos se han calculado una serie de características de diferentes grabaciones de músicas y se ha anotado la longitud y la latitud de su origen.

Puedes obtener este conjunto de datos mediante la función `load_world_music` de la librería `apafib` esta retornará un dataframe de Pandas. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

a) Divide el conjunto de datos en entrenamiento y test (70 %/30 %). Verás que todas las variables son continuas. Hay algunos valores perdidos, impútalos usando el `KNNImputer` de `scikit-learn`. Una vez tenemos el conjunto de datos preprocesado puedes aplicar PCA para ver su comportamiento. Analiza sus resultados y representa las dos variables objetivo sobre el PCA en dos dimensiones. Explica lo que has observado.

b) Transforma las variables adecuadamente para poder ajustar un modelo de regresión tanto el conjunto de entrenamiento como el de test. Ajusta un modelo de regresión lineal para predecir las dos variables objetivo y estima la calidad de la regresión. ¿Te parece suficientemente bueno el resultado? Analiza los pesos de los modelos. Representa los residuos y la predicción sobre el test contra los valores reales. Comenta qué aparece. ¿Qué variable objetivo es más fácil de predecir?

c) El número de variables es relativamente grande y eso puede afectar a la efectividad de los modelos. Ajusta regresiones Ridge y LASSO a los datos y analiza los resultados. ¿Qué modelo es mejor para cada variable objetivo? Estudia cómo LASSO ha simplificado los modelos, ¿para qué variable objetivo hacen falta más variables?

d) Es complicado el tener que predecir dos variables de manera separada. Una alternativa sería el poder representarlas como una. Dado que tenemos localizaciones una posibilidad sería el calcular la distancia a un punto arbitrario (por ejemplo, 0 latitud y 0 longitud) y usarla como

[4]El conjunto de datos original es más difícil y no registra la localización de las músicas con gran precisión.

variable objetivo aunque no represente lo mismo[5]. Podemos calcular la distancia entre dos puntos en la superficie de una esfera usando la distancia de Haversine. Calcula esta distancia para cada ejemplo (longitud y latitud están en radianes) y utiliza regresión lineal y LASSO. Compara la predicción con la media del error de las dos variables originales. Comenta los resultados.

6. **¿Quieres saber cuanto vale tu coche?**[6]

   El estimar el precio de un coche usado es algo complejo y depende de muchos factores. Vamos a trabajar con un conjunto de datos sobre características de coches usados (fabricante, modelo, estado, costes...) que las asocia con el precio al que se vendieron.

   Puedes obtener este conjunto de datos mediante la función `load_car_sales` de la librería `apafib` esta retornará un dataframe de Pandas. Resuelve los siguientes apartados ilustrando los resultados de la manera que te parezca más adecuada.

   a) Divide el conjunto de datos en entrenamiento y test (70 %/30 %). Verás que hay una mezcla de variables continuas y categóricas. Trata adecuadamente las variables categóricas teniendo en cuenta lo que significan. Hay algunos valores perdidos en las variables continuas, impútalos usando el `KNNImputer` de `scikit-learn`. Una vez tenemos el conjunto de datos preprocesado puedes aplicar PCA adecuadamente para ver su comportamiento. Analiza sus resultados y representa la variable objetivo sobre el PCA en dos dimensiones. Explica lo que has observado.

   b) Transforma las variables adecuadamente para poder ajustar un modelo de regresión tanto el conjunto de entrenamiento como el de test. Ajusta un modelo de regresión lineal para predecir la variable objetivo y estima la calidad de la regresión. ¿Te parece suficientemente bueno el resultado? Analiza los pesos de los modelos. Representa los residuos y comenta qué aparece.

   c) Ajusta regresiones Ridge y LASSO a los datos. Analiza la diferencia de pesos entre la regresión lineal y Ridge. Analiza los pesos de LASSO y comenta qué variables han desaparecido de la regresión. ¿Tienen sentido las variables con mayor y menor peso para la predicción del precio? O lo que es lo mismo, ¿se podría explicar en función de lo que considera el modelo importante por qué se da un precio a un coche?

   d) Habrás observado algo bastante particular en los residuos de los modelos. Podemos intentar arreglarlo introduciendo interacciones entre las variables. Usa la función `PolynomialFeatures` de scikit-learn para añadir al conjunto de datos original características que correspondan a polinomios de grado 2. Vuelve a ajustar la regresión Ridge y la regresión LASSO. ¿Han mejorado los modelos? ¿Se ha arreglado el problema con los residuos? ¿Como de simple es el modelo LASSO?

[5] Estamos perdiendo información con esta transformación, ya que consideramos iguales ejemplos en círculos concéntricos alrededor de ese punto arbitrario, pero podría servir para estudiar la expansión de las características de la música desde un origen.

[6] Si no he visto ese anuncio mil veces no lo he visto ninguna, que pesados.