## 1. Función de Producción de Cobb Douglas

La base de datos nativa de R RDPerfComp, que se encuentra en la liberia pder, contiene datos de 509 empresas de USA corresondientes a los años 1982 a 1989, tomados de Blundell, Richard and Stephen Bond (2000) “GMM estimation with persistent panel data: an application to production functions”, Econometric Reviews, 19(3), 321-340.

La estructura de la base de datos es la siguiente:

* + id: identificador de la empresa
  + year: año
  + y: producción en logaritmo
  + n: Empleo en logaritmo
  + k: Capital en logaritmo
* Realice un histograma de la variable y, y estime un modelo GLM para la producción utilizando diferentes familias y links acordes con el histograma. Lógicamente, habrá familias que no se adecuen a estos datos, exponga cuales de estas familias y funciones de ligadura no deben de utilizarse e indique el porqué.
* Realice lo mismo, pero con los datos en niveles (no en logaritmos).
* Comente los resultados de las regresiones que ha realizado y compare la bondad del ajuste de ambos modelos.
* Presente el sumario de resultados y el gráfico de XY con los puntos reales y los datos estimados, especificando en el título la función estimada.
* Realice un pronóstico de y correspondiente a una empresa con 30 empleados. En el pronóstico presente los intervalos de confianza al 95% y 90%. Presente el pronóstico en niveles de producción.

## 2. Diseño experimental

Un investigador se propone estudiar el desarrollo de la aptitud en análisis de datos de un determinado grupo de estudiantes. A tal propósito, confecciona una serie de tareas estandarizadas, consistentes en el desarrollo de cuatro modelos de análisis. Estas tareas son presentadas a los estudiantes, cuando realizan las evaluaciones. Las evaluaciones son programadas de forma secuencial a lo largo del curso.

Por último, el rendimiento en la resolución de los modelos es valorado con una escala de 10 puntos. Dado que el investigador considera de interés estudiar la posible diferencia atribuible al género, elige dos muestras iguales de escolares de uno y otro género. Los resultados se muestran en el fichero Notas.xlsx.

* ¿Qué tipo de modelo es el más adecuado para este análisis? Realice un dibujo representando dicho modelo.
* Haga un análisis descriptivo de los datos teniendo en cuenta los factores definidos.
* Estime el modelo anteriormente planteado, comente los resultados obtenidos y establezca las conclusiones pertinentes derivadas de este estudio.

## 3. Análisis factorial y Análisis Cluster

En el fichero de calidad\_vida.csv figuran los datos de 2017 por Comunidades Autónomas relativos a la DIMENSIÓN 1. CONDICIONES MATERIALES DE VIDA, de los indicadores de calidad de vida que publica el Institituto Nacional de Estadística (<https://www.ine.es/ss/Satellite?L=0&c=INEPublicacion_C&cid=1259937499084&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&param1=PYSDetalleGratuitas&param4=Ocultar#top>)

Los Subdominios e Indicadores parciales de la Dimensión 1 son los siguientes:

1.1. Condiciones económicas

1.1.1. Renta media y mediana

1.1.2. Población en riesgo de pobreza relativa según distintos umbrales

1.1.4. Desigualdad (S80/S20)

1.1.5. Satisfacción con la situación económica del hogar

1.2. Condiciones materiales

1.2.1. Dificultades para llegar a fin de mes

1.2.2. Carencia material

1.2.3. Población que vive en hogares con determinadas deficiencias en la Vivienda

1.2.4. Población con falta de espacio en la vivienda

1.2.5. Población con gasto elevado en vivienda

1.2.6. Satisfacción con la vivienda

1.3. Seguridad económica

1.3.2. Incapacidad de hacer frente a gastos económicos imprevistos

1.3.3. Retrasos en los pagos

* Lea el fichero, y realice las pruebas pertinentes para ver si el conjunto de datos puede ser objeto de análisis factorial. De ser cierto, plantee una análisis factorial con la libreria “psych”, respondiendo que criterio ha elegido para seleccionar los factores y que porcentaje de varianza explican los factores elegidos. Interprete el significado de dichos factores.
* Realice el mismo análisis con la librería rela. ¿Son equivalentes los resultados obtenidos? Exponga las diferencias.
* A la vista de ambos análisis: ¿Agruparía las variables en los subdominios que utiliza el INE?
* Por último, incorpore a la base de datos la/s variable/s nueva/s con las puntuaciones factoriales y realice con ellas un clúster de las CCAA utilizando el método de Ward. Determine el número más adecuado de clústeres, e interprete los resultados.

## 4. Análisis de Correspondencias Múltiple

El conjunto de datos de hardware de Hartigan screws.sav contiene información sobre las características de los tornillos, pernos, tuercas y tachuelas. La siguiente tabla muestra las variables, junto con sus etiquetas de variable y las etiquetas de valor asignadas a las categorías de cada variable.

| *Hartigan hardware dataset* | | |
| --- | --- | --- |
| **Variable name** | **Variable label** | **Value label** |
| *thread* | *Thread* | *Yes\_Thread*, *No\_Thread* |
| *head* | *Head form* | *Flat*, *Cup*, *Cone*, *Round*, *Cylinder* |
| *indhead* | *Indentation of head* | *None*, *Star*, *Slit* |
| *bottom* | *Bottom shape* | *sharp*, *flat* |
| *length* | *Length in half inches* | *1/2\_in*, *1\_in*, *1\_1/2\_ in*, *2\_in*, *2\_1/2\_in* |
| *brass* | *Brass* | *Yes\_Br*, *Not\_Br* |
| *object* | *Object* | *tack, nail1, nail2, nail3, nail4, nail5, nail6, nail7, nail8, screw1, screw2, screw3, screw4, screw5, bolt1, bolt2, bolt3, bolt4, bolt5, bolt6, tack1, tack2, nailb, screwb* |

Analice estos datos a través de un Análisis de Correspondencias Múltiple. Comente los resultados e interprete las dimensiones obtenidas.

**5. Escalamiento Multidimensional y Análisis Cluster**

Con los datos del fichero Datos CCAA.xlsx realice el escalamiento multidimensional que crea más oportuno. Interprete las dimensiones obtenidas.

Utilizando las coordenadas en las dimensiones obtenidas para cada comunidad autónoma, realice un análisis cluster, seleccione el número de clústeres más adecuado y describa cada grupo en función de las características de las comunidades autónomas incluidas en él.

**6. Análisis de Correlación Canónica**

El fichero Datos CCAA2.xlsx contiene información relativa al uso de la economía colaborativa procedente de la Encuesta sobre equipamiento y uso de tecnologías de información y comunicación en los hogares 2018.

En concreto, las tres primeras columnas hacen referencia al porcentaje de usuarios de internet que utilizan webs o apps de cara a obtener un servicio de alojamiento, y las tres siguientes hacen referencia al servicio de transporte.

Analice esta información mediante un análisis de correlación canónica y describa y comente los resultados obtenidos.

**7. Machine Learning**

Se ha llevado a cabo una campaña de marketing por parte de una empresa sobre 9.134 de sus clientes en cuatro estados de Estados Unidos para ofrecerles un determinado producto. Se dispone de información demográfica, social y de otro tipo de datos relacionados con su gestión económica.

Las 24 variables contenidas en la base de datos son las siguientes:

* Customer (Código de Cliente)
* State (Estado)
* Customer Lifetime Value (Valor de vida del cliente)
* Response (Respuesta)
* Coverage (Cobertura)
* Education (Nivel de educación)
* Effective To Date (Vigente hasta la fecha)
* EmploymentStatus (Estatus de Empleo)
* Gender (Género)
* Income (Ingresos)
* Location Code (Código de localización)
* Marital Status (Estado civil)
* Monthly Premium Auto (Auto Premio Mensual)
* Months Since Last Claim (Meses desde la última reclamación)
* Months Since Policy Inception (Meses desde el inicio de la poliza)
* Number of Open Complaints (Número de quejas abiertas)
* Number of Policies (Número de pólizas)
* Policy Type (Tipo de póliza)
* Policy (Pólizas)
* Renew Offer Type (Renovación tipo de oferta)
* Sales Channel (Canal de ventas)
* Total Claim Amount (Cantidad total de reclamación)
* Vehicle Class (Tipo de vehículo)
* Vehicle SizesSEE (Tamaño del vehículo)

El objetivo general de este ejercicio es aplicar diferentes métodos y algoritmos de clasificación con el objetivo de seleccionar aquel o aquellos métodos que sean óptimos a la hora de predecir la variable Response (valores: Yes, No, es decir si la campaña ha sido un éxito o no). De los siguientes modelos, estime al menos uno de cada tipo:

1. Modelo Lineal Generalizado (Regresión Logística, Regresión Probit).
2. Árboles de decisión. (CHAID, CART, C5, Random Forest).
3. Redes Neuronales. (Perceptron Multicapa, Función de Base Radial).

Utilice el 80% de la muestra para el entrenamiento de los modelos y el restante 20% para test.

Compare los resultados obtenidos por los diferentes modelos construidos, y seleccione el que finalmente utilizaría, utilizando como criterio el área bajo la curva ROC.