CUESTIÓN 1

Para estimar el número medio de portátiles por vivienda en una determinada zona rural se realiza un diseño en dos etapas:

Primera etapa: Se obtiene una muestra aleatoria simple con reemplazamiento de m=3 pueblos entre un total de M=300. El tamaño de cada pueblo (N_i) se mide por el número de viviendas.

Segunda etapa: Se selecciona una muestra aleatoria simple de 5 viviendas en cada uno de los pueblos seleccionados en primera etapa.

El número total de viviendas es de N=9000 .

Los datos disponibles son:

Unidades primarias (pueblos)	Tamaños (N_i)	Número de portátiles observados en las viviendas de la muestra m_i=5
1	50	1;0;3;2;2
2	60	2;2;1;2;3
3	65	2;4;3;2;1

Se pide:

(a) Calcular un estimador insesgado para el número medio de portátiles por vivienda en los siguientes casos:

- Muestreo con probabilidades iguales en las dos etapas.

- Muestreo con probabilidades proporcionales al tamaño en la primera etapa.

(b) Estimar el error de muestreo en ambos casos. ¿Cuál de los dos estimadores es preferible?

CUESTIÓN 2

Se formula un modelo (de regresión no lineal) de red neuronal de dos capas ocultas completamente conectada con las siguientes características:

i) entradas $\mathbf{x} \in \mathbf{R}^6$;
ii) salida $y \in \mathbf{R}$;
iii) primera capa oculta con cuatro nodos y función de activación logística;
iv) segunda capa oculta con dos nodos y función de activación ReLU.

Se pide:

(a) Representar el modelo gráficamente.

(b) Formular la transformación asociada a un nodo de la primera capa oculta. Formule la transformación asociada a un nodo de la segunda capa oculta.

(c) Representar la relación entre la respuesta y y la salida de la segunda capa oculta.

(d) Se dispone de datos $(\mathbf{x}_i, y_i)_{i=1}^n$ para entrenamiento de la red. Se desea ajustar la red mediante mínimos cuadrados con regularizador de caída de pesos (weight decay). Formule la función objetivo del problema que debería resolverse para ajustar el modelo.

(e) Describir analíticamente una iteración genérica del algoritmo del descenso del gradiente para ese problema. Describa qué rol tendría el método de retropropagación en tal algoritmo.

(f) Describir qué problemas se dan cuando el tamaño n de los datos es grande y el modelo de red neuronal es profundo. Describa cómo pueden mitigarse esos problemas con un método de descenso del gradiente estocástico.

(g) Se dispone de datos $(\mathbf{x}_i, y_i)_{i=1}^n$ para ajustar la red. Explique cómo procedería, y con qué propósito, a dividir el conjunto en uno de entrenamiento y otro de contraste para un mejor ajuste del modelo.

CUESTIÓN 3

Tenemos el siguiente esquema relacional, con información de una pequeña biblioteca:

Socio(cód_socio, nombre, dirección. email)
Libro(núm_registro, título, autor, isbn, ubicación)
EnPréstamo(núm_registro, cód_socio, fecha_préstamo)

Se pide diseñar consultas de álgebra relacional para averiguar la siguiente información:

(a) Qué títulos están actualmente en préstamo.

(b) Qué socios tienen más de tres ejemplares en préstamo.

(c) Nombre y direcciones de email de los socios que tienen libros desde hace más de 30 días. (Admitimos la operación "fecha + 30" o "fecha – 30" para indicar fechas posteriores o anteriores en 30 días a una fecha dada y las operaciones fecha1 < fecha2 para expresar que una fecha es anterior a otra.)

(d) Un socio devuelve un libro. ¿Cuál será la operación de álgebra relacional que actualiza nuestro esquema?

CUESTIÓN 4

Tenemos un archivo de texto plano, llamado "población.csv", con datos de poblaciones del mundo. He aquí el inicio de su contenido:

```
7.975.105.155
China,1.425.887.337
India,1.417.173.173
Estados Unidos,338.289.857
Indonesia,275.501.339
```

La primera línea es la población mundial total; las siguientes, las de cada país. El nombre del país puede contener más de una palabra (ejemplos: Corea del Norte, Islas Vírgenes de los Estados Unidos). Las cifras de poblaciones vienen en el formado mostrado, con puntos separando los miles, millones, etc. Los nombres de los países y las cifras vienen separados con una coma.

Deseamos elegir el nombre de un país aleatoriamente, con probabilidades proporcionales a sus poblaciones. Se pide diseñar un programa o algoritmo para ello.

El diseño pedido puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. En la simulación de la variable aleatoria que se necesite, se podrá asumir que se tiene definida y disponible una función básica que simule una variable aleatoria uniforme continua U(0, 1). No se podrá usar ninguna otra función aleatoria.

CUESTIÓN 5

Sea una población formada por N=40 empresas. Se quiere estimar el total de su cifra de negocios (CN) y para ello se obtiene una muestra aleatoria simple (sin reemplazamiento) de n=4 empresas.

Se conoce el número de ocupados (x) de cada una de las empresas de la población. El total de ocupados en la población es X=160.

Los datos disponibles son los siguientes:

Unidades de la Muestra (i)	Y (CN, miles de euros)	X (ocupados)
1	100	3
2	200	5
3	500	7
4	300	5

Se pide:

(a) Estimar el total de la Y (CN) por el método de la razón.

(b) Estimar el sesgo del estimador del apartado anterior.

(c) Estimar su varianza.

(d) Explicar si es o no relevante el efecto del sesgo en el intervalo de confianza a partir de los resultados de los apartados anteriores.

CUESTIÓN 6

Suponga que las observaciones (x, y) de un problema pueden modelizarse mediante regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, donde $\epsilon_1, ..., \epsilon_n$ se distribuyen independientemente según una $N(0, \sigma^2)$. Para simplificar, suponga que σ es conocida e igual a 1.

Se pide:

(a) Escribir la log-verosimilitud de los datos.

(b) Suponga que, a priori, (β_0, β_1) son independientes e idénticamente distribuidas según una distribución normal de media 0 y varianza c^2. Calcule la distribución a posteriori de los parámetros.

(c) Basándose en tal distribución a posteriori, explique cómo contrastaría la hipótesis nula $H_0 : \beta_1 = 0$.

(d) Calcular la distribución predictiva de y para una nueva observación x_{n+1}.

(e) Dar un intervalo predictivo de probabilidad 0.88 para y dado x_{n+1}, basado en la salida de una rutina `rnorm(s)` que permite generar `s` observaciones de una normal estándar.

CUESTIÓN 7

Un supermercado necesita diseñar una base de datos para gestionar su inventario de productos y controlar las ventas. Cada producto se registra con su código, una descripción sucinta, su precio y las cantidades de dicho producto mínima (para reponer existencias cuando baje de dicha cantidad), real (la que hay realmente en los estantes) y máxima (la que cabe en los estantes). Cada cliente se registra con su email, su nombre y su domicilio). Cada vez que un cliente acude al supermercado, al pagar se genera un único tique, con la fecha, la lista de productos y la cantidad de cada uno, y el importe total de la factura. El supermercado registra, además, lógicamente, el cliente que ha realizado dicha compra.

Se pide:

(a) Realizar un diseño E/R de dicha base de datos.

(b) Definir, con este diseño, los siguientes conceptos, si están presentes, o inventa un ejemplo sencillo pero apropiado para cada uno de ellos:

- Entidades fuertes, entidades débiles.
- Atributos, claves.
- Relación uno a uno, uno a muchos, muchos a muchos.

(c) Participación total o parcial en una relación.

(d) Atributos de una relación.

(e) Atributo multivalorado, atributo calculado.

CUESTIÓN 8

Tenemos un archivo de texto plano, llamado "pps.csv" (abreviando país, población, superficie, separados por comas), con datos de poblaciones y superficies de países del mundo. He aquí el inicio de su contenido:

```
China,1.425.887.337,9.562.900
India,1.417.173,173 3,287.300
Estados Unidos,338.289.857,9.831.500
Indonesia,275.501.339,1.913.600
```

Cada línea consigna el nombre de cada país, seguido de su población y su superficie en kilómetros cuadrados. El nombre del país puede contener más de una palabra (ejemplos: Corea del Norte, Islas Vírgenes de los Estados Unidos). Las cifras de poblaciones vienen en el formado mostrado, con puntos separando los miles, millones, etc. Los nombres de los países y las cifras vienen separados con una coma. En las cantidades, se usa el punto para separar miles, millones, etc.

Se pide:

(a) Diseñar un programa o algoritmo para obtener el coeficiente de correlación entre ambas variables, superficie y población.

(b) Como sabes, es posible realizar esto recorriendo el archivo dos veces, una para calcular las medias y otra para recopilar el resto de la información. Pero la lectura de archivos es bastante costosa computacionalmente. Por eso, en este segundo apartado se pide diseñar un programa o algoritmo para obtener obligatoriamente este coeficiente leyendo el archivo en una sola pasada. Si ya lo has hecho así en tu primer apartado, basta con que en éste lo menciones.

El diseño puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. No está permitido usar ninguna función predefinida que calcule la media ni otras medidas estadísticas.