Laboratorio de Datos
Recuperatorio
3/7/2023

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

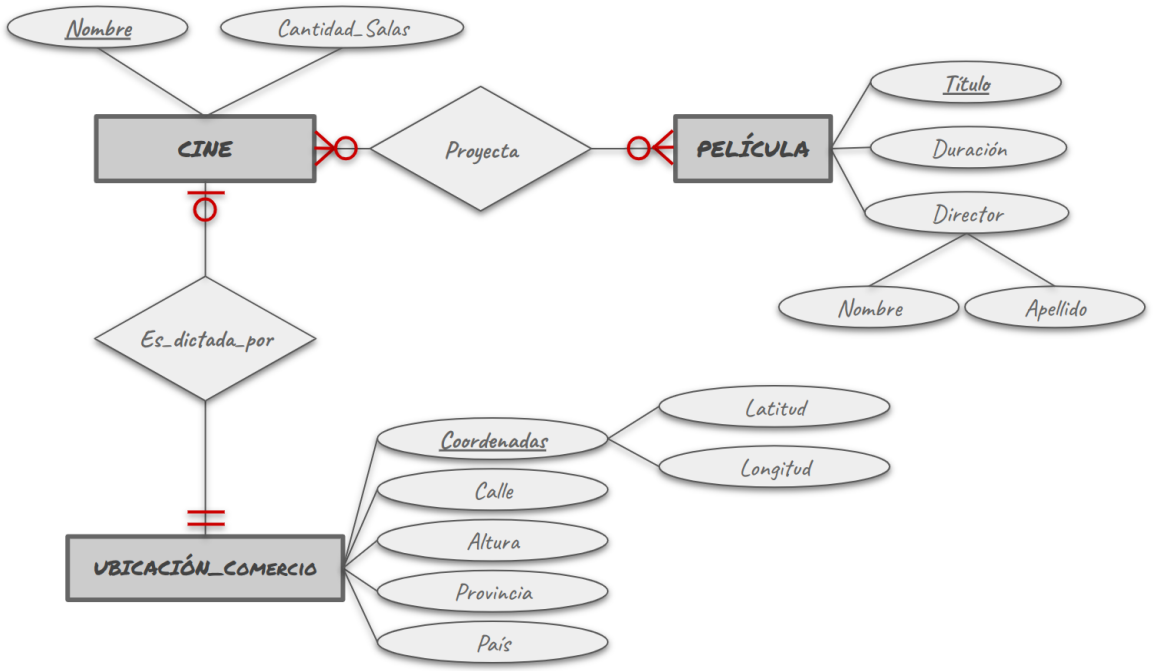
Nombre y apellido:
LU:

Grilla de puntajes. No completar.

Ej 1	Ej 2	Ej 3	Ej 4	Ej 5	Ej 6	Ej 7	Ej 8	Total	Condición

Aclaraciones: El parcial NO es a libro abierto. Para aprobar se requieren al menos 60 puntos. Cualquier decisión de interpretación que se tome debe ser aclarada y justificada. Todas las respuestas tienen que estar justificadas de manera concisa. Agregue nombre, apellido, LU y nro. de hoja (empezando a numerar en las hojas de respuesta) en el extremo superior izquierdo de cada hoja.

1. (15 p) Dado el siguiente DER mapearlo al modelo relacional. No olvide indicar en todos los casos nombre de esquema, sus atributos, clave primaria y foreign keys.



2. (15 p) Dado el siguiente esquema, correspondiente a la manera en que almacena los datos de sus entradas un museo, decir si está en 2FN y/o en 3FN. En caso de no estarlo proponer una descomposición que se encuentre en 3FN, que preserve las dependencias funcionales y sea lossless join. Marcar las claves primarias (PK) y las dependencias funcionales en los esquemas surgidos por la descomposición.

Esquema
ENTRADA_MUSEO(**TipoDocumento**, **Numero_Documento**, **Fecha**, ValorEntrada,
NombreVisitante, TipoDeAcceso,
TiempoPermitidoDePermanencia)

Dependencias Funcionales
TipoDocumento + Fecha -> ValorEntrada
TipoDocumento + Numero_Documento -> NombreVisitante
TipoDocumento + Numero_Documento + Fecha -> TipoDeAcceso
TipoDeAcceso -> TiempoPermitidoDePermanencia

3. (15 p) Dadas las siguientes tablas EMPLEADO y PROYECTO con el contenido que se muestra a continuación, si se ejecutan las siguientes consultas SQL ¿qué se obtiene como resultado?. Escribir la tabla resultante con su contenido, es decir tanto filas como columnas.

Empleado

Tipo_doc	Nro_doc	Nombre	Fecha_ingreso	ID_Proyecto
DNI	12345678	Juan Pérez	2021-05-10	2
DNI	23456789	María González	2022-01-15	4
DNI	34567890	Alejandro Rodríguez	2020-11-30	2
DNI	45678901	Lucía Martínez	2023-03-18	3

Proyecto

ID_Proyecto	ID_SubProyecto	Nombre_Proy
1	A	PETRÓLEO
2	A	PANDAS
2	H	BALLENAS
4	C	CO2

i)
```
SELECT e.Tipo_doc, e.Nro_doc, e.Nombre, p.Nombre_Proy
FROM Empleado AS e
LEFT OUTER JOIN Proyecto AS p
ON e.ID_Proyecto=p.ID_Proyecto
WHERE ID_Proyecto <> 5
```

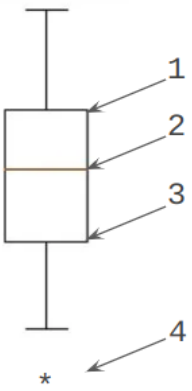
ii)
```
SELECT nombre
FROM Empleado
WHERE fecha_ingreso = ( SELECT MIN(fecha_ingreso)
                        FROM Empleado)
```

4. (10 p) Dados los siguientes problemas de calidad de datos clasifíquelos en función del atributo de calidad que se ve afectado y a si es problema de modelo o de datos (instancia).
   a. No es posible almacenar el sistema de referencia.
   b. Hay inconsistencias entre nombres de personas con un mismo DNI en distintos sistemas.
   c. La ubicación almacenada en un sistema de varios pozos petroleros no coincide con la ubicación real.
   d. Los teléfonos de los pacientes de un consultorio médico no están actualizados.

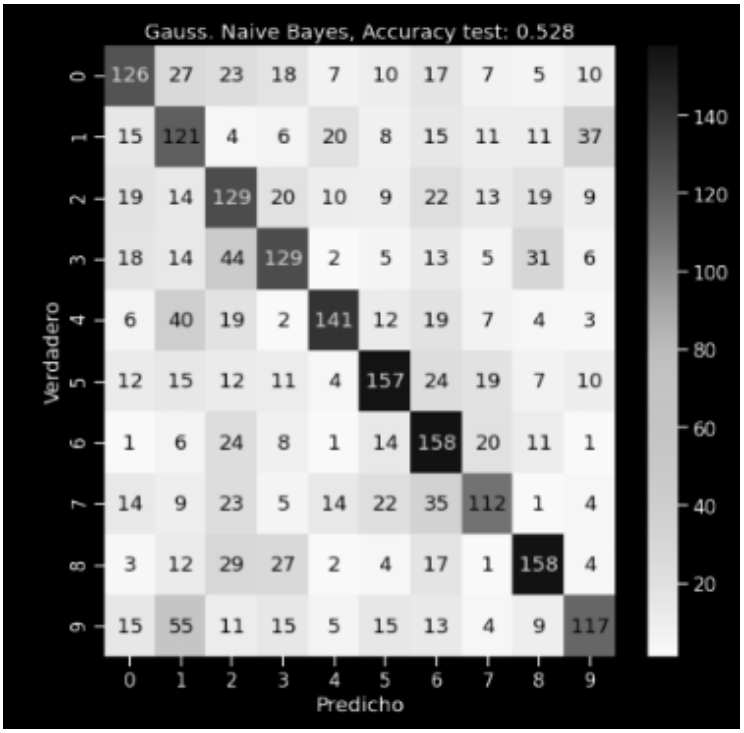
Nombre y apellido:
LU:

5. (10 p) Dada la siguiente caja de un boxplot, qué es lo que se representa con:

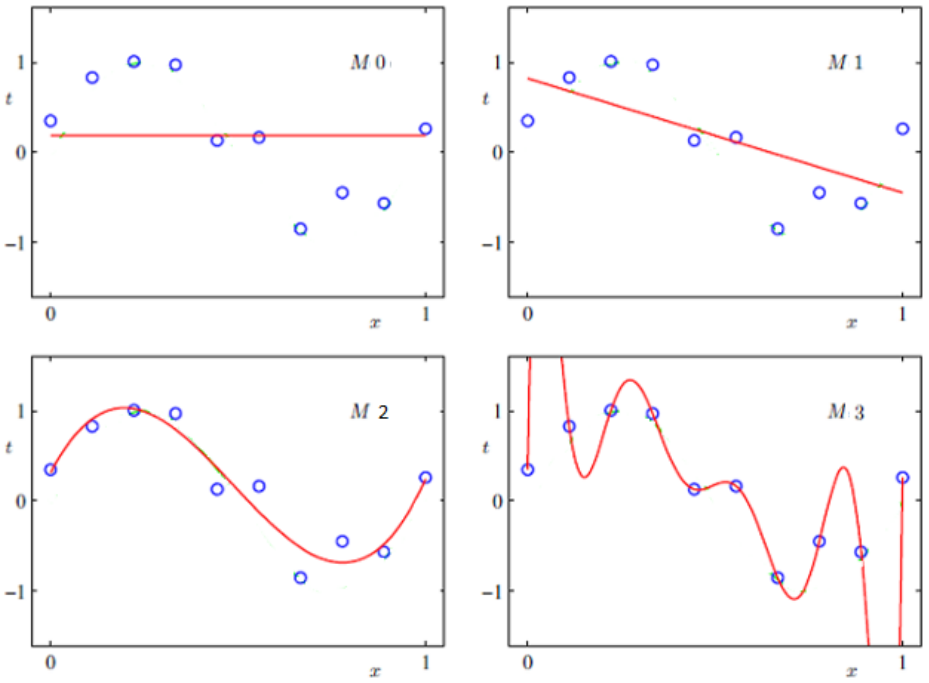
   i) la línea horizontal señalada por la flecha número 1
   ii) la línea horizontal señalada por la flecha número 2
   iii) la línea horizontal señalada por la flecha número 3
   iv) la estrella señalada por la flecha número 4



6. (10 p) Un sistema de reconocimiento de dígitos hablados tiene la siguiente matriz de confusión. Mencionar los 3 dígitos que se confunden con mayor frecuencia por otros (y cuáles son éstos) y cómo sería una matriz de confusión ideal. Expresarlo con palabras, no hace falta hacer las cuentas.



7. (10 p) Dados los datos de entrenamiento mostrados como círculos azules en los gráficos de abajo.
   a. ¿Cuál de los modelos (M0, M1, M2 o M3) tiene mejor ajuste en los datos de entrenamiento?
   b. ¿Considera que ese es el mejor modelo?
   c. ¿Cuál modelo cree que daría mejor con un conjunto de datos nuevo, con el que no se entrenó el modelo?



8. (15 p) Decidir V o F y justificar.
   a. Antes de entrenar un modelo de árboles de decisión no es necesario reescalar los datos.
   b. Para evitar el sobreajuste de un árbol de decisión, es conveniente utilizar el criterio de gini.
   c. DBSCAN es un método útil para clustering en casos en que los datos tienen formas raras y outliers.
   d. El método de regresión de knn puede sobreajustar si se consideran valores de k muy altos.
   e. La validación cruzada es un método para comparar modelos de clasificación.