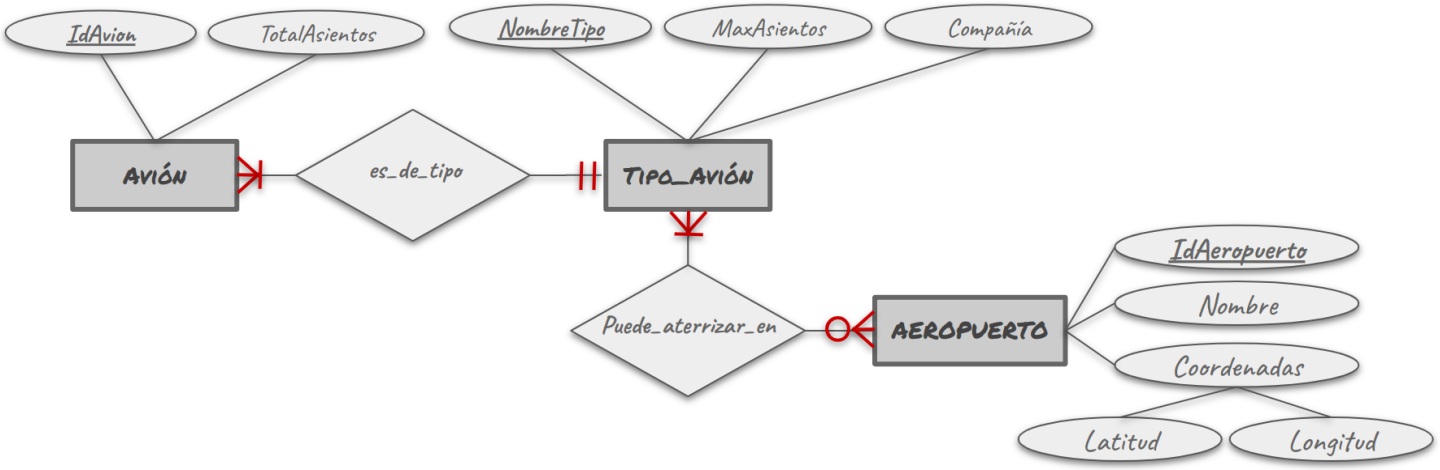
Nombre y apellido:
LU:

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 datos de películas que se proyectan en un cine, 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
PROYECCION(**Título**, **Formato**, Director, Nacionalidad_Director,
Precio, Duración, Puntuación)

Dependencias Funcionales
Título -> Director
Director -> Nacionalidad_Director
Título + Formato -> Precio
Título -> Duración
Titulo -> Puntuación

A modo de ejemplo, a continuación se muestran una tabla con algunos de los datos.

Título	Formato	Director	Nacionalidad_Director	Precio	Duración	Puntuación
Asteroid City	2D	Wes Anderson	Estadounidense	$ 1.300,00	104 minutos	B
Asteroid City	3D	Wes Anderson	Estadounidense	$ 1.800,00	104 minutos	B
Viedma, la Capital que no fue	2D	Jorge Leandro Colás	Argentino	$ 1.300,00	78 minutos	A
Viedma, la Capital que no fue	3D	Jorge Leandro Colás	Argentino	$ 1.800,00	78 minutos	A

3) (10 p) Dados las siguientes tablas TURISMO y UBICACION 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.

TURISMO

Atractivo	País
Cataratas	Argentina
La Quiaca	Argentina
Camino de Santiago	España

UBICACION

País	Provincia
Argentina	Misiones
Argentina	Jujuy
España	La Rioja

i)
```
SELECT a.Atractivo, u.País, u.Provincia
FROM TURISMO AS a
INNER JOIN UBICACION AS u
ON a.País=u.País
```

ii)
```
SELECT a.País, COUNT(*) AS total
FROM TURISMO AS a
WHERE País LIKE '%_a'
GROUP BY País
HAVING total >= 2
```

4) (10 p) En el área de calidad de datos ¿a qué nos referimos cuando mencionamos que hay un problema de calidad asociado a una instancia? Dar un ejemplo aclarando el atributo de calidad (o dimensión) asociado al problema.

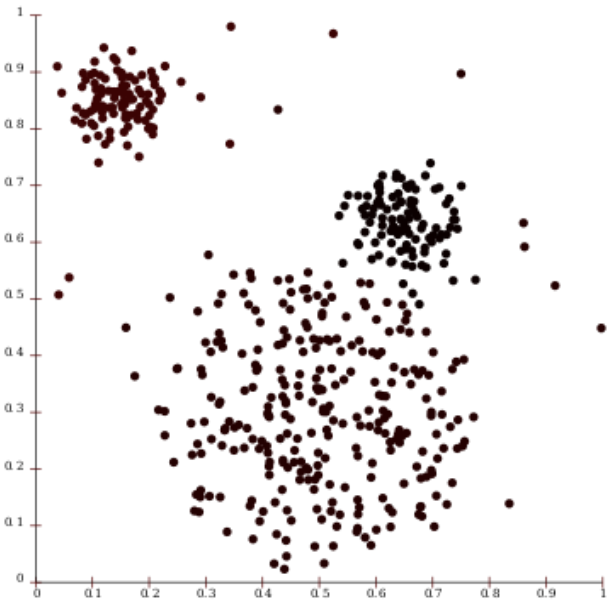
5) (10 p) Se tienen los siguientes datos de alturas. Se quiere predecir la altura de adulta de una mujer cuya madre mide 1.61 m ¿Qué predicción arrojaría un modelo de knn con k = 3?

sexo	altura madre (cm)	altura (cm)
M	155.20	167
F	158.60	158
F	162.30	162
F	158.90	155
F	162.40	160
F	160.70	158
M	159.50	174
M	160.20	171
F	154.50	153
M	158.80	176

6) (15 p) En un jardín de infantes, durante el último mes, ocurrió que a 5 niñxs les indicaron quedarse en sus casas debido a la suposición de que tenían la enfermedad PMB, que es muy contagiosa. Sin embargo, luego de algunas consultas médicas, en 4 de los 5 casos fue descartada la enfermedad, es decir que sólo 1 de esos niños tuvo efectivamente PMB.
Sean a = TP/(TP+FN), b = TP/(TP+FP).

¿A cuál de estas métricas se le dio más peso al indicarles no asistir al jardín?

7) (10 p) Se quiere realizar un clustering con estos datos. De las técnicas vistas en clase, ¿cuál recomienda? ¿Por qué?



8) (15 p) Decidir V o F y justificar.

a. Los árboles de decisión son buenos porque no suelen sobreajustar (overfitting).
b. K-means es un método de clasificación supervisada basada en distancias.
c. Antes de ajustar un modelo de knn conviene reescalar los datos.
d. Para saber si hay overfitting (sobreajuste) es suficiente evaluar la performance de un modelo con los datos de entrenamiento.
e. En una tarea de clasificación binaria se obtiene un accuracy del 95%. A partir de estos datos se puede afirmar que el resultado del método fue exitoso.
f. Siempre es conveniente elegir aquel algoritmo que tenga mejores métricas de clasificación por sobre los demás. F
g. Si se tiene un error bajo en el conjunto de entrenamiento se puede deducir que el modelo utilizado es bueno. F
h. K-means es una buena técnica de clustering si no se sabe de antemano la cantidad de clusters. F