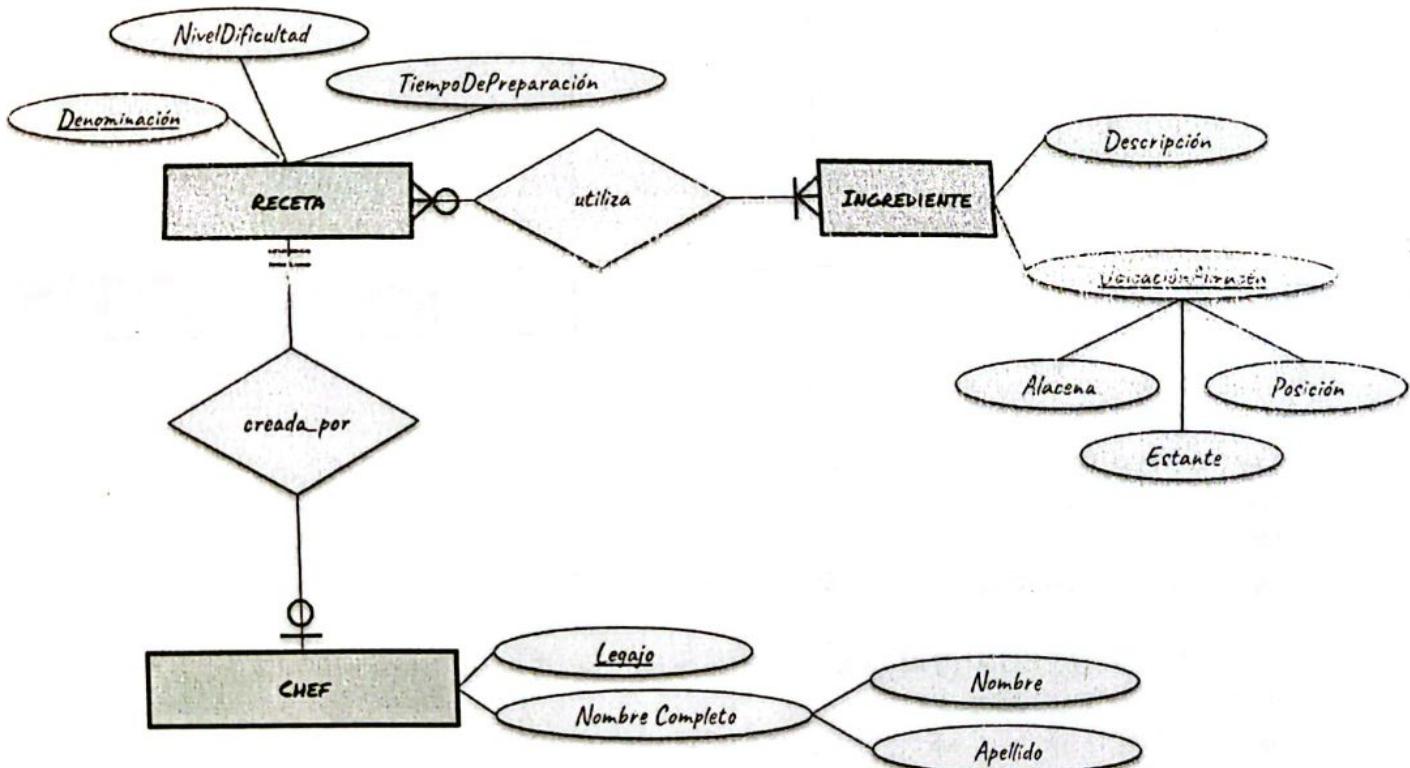


Grilla de puntuajes. No completar.

Ej 1	Ej 2	Ej 3	Ej 4	Ej 5	Ej 6	Ej 7	Ej 8	Total	Condición
15	15	15	9	8,5	8	10	14	94,5	(A)

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 (las FK con flechas). En el caso de existir más de una opción para implementar una relación, elegir aquella que evite los Nulls a través del diseño.



2. (15 p) Dado el siguiente esquema, correspondiente a la manera en que una biblioteca almacena los datos de los préstamos de sus libros, decir si está en 2FN y/o en 3FN (no olvide justificar). En caso de no estarlo proponer una descomposición que se encuentre en 3FN, que preserve las dependencias funcionales y sea lossless join (no es necesario aplicar el algoritmo de descomposición para demostrarlo, sólo proponer una posible descomposición en caso de ser necesaria). Marcar las claves primarias (PK) y las dependencias funcionales en los esquemas surgidos por la descomposición.

Esquema

PRESTAMOS_LIBROS(IdEjemplar, IdSocio, fechaInicioPrestamo,
fechaFinPrestamo, IdLibro, nombreLibro, Editorial,
nombreSocio, apellidoSocio, emailSocio)

Dependencias Funcionales

- DF1: $IdEjemplar, IdSocio, fechaInicioPrestamo \rightarrow fechaFinPrestamo$
- DF2: $IdEjemplar \rightarrow IdLibro$
- DF3: $IdLibro \rightarrow nombreLibro, Editorial$
- DF4: $IdSocio \rightarrow nombreSocio, apellidoSocio, emailSocio$

3. (15 p) Dadas las tablas **COLECTIVO** y **TERMINAL**, 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.

COLECTIVO

Línea	Ramal	CantUnidades
42	A	30
42	B	40
107	A	28
107	B	22
10	A	20

TERMINAL

Línea	Ramal	Cabecera
42	A	River
42	B	Ciudad Univ.
107	A	Ciudad Univ.
107	B	Ramsay
10	A	Palermo
10	B	Aeroparque

i) **SELECT c.Línea, c.Ramal, t.Cabecera
FROM COLECTIVO AS c
LEFT OUTER JOIN TERMINAL AS t
ON c.Línea=t.Línea
WHERE c.Línea < 100
ORDER BY c.Línea ASC, c.Ramal ASC, t.Cabecera DESC**

ii) **SELECT c.Línea, COUNT(') AS TotalRamales, SUM(c.CantUnidades) AS TotalUnidades
FROM COLECTIVO AS c
INNER JOIN TERMINAL AS t
ON c.Línea=t.Línea AND c.Ramal=t.Ramal
GROUP BY c.Línea
HAVING TotalUnidades < 60 } (le agrega la where en manual)
ORDER BY c.Línea ASC**

4. (10 p) Para evaluar el problema de calidad de datos en la materia Laboratorio de Datos, se desea analizar la Completitud del dato email asociado a los alumnos inscriptos a la materia. Explicar brevemente qué representa el atributo de Completitud. Proponer un modelo GQM para encarar el problema (debe figurar la meta, la pregunta y la métrica a utilizar).

- ✓ 1) CHEF (LEGATO, NUMERO, ~~APPELLIDO~~, VENOMINACION - RECETA)
RECETAS (LEGATO, VENOMINACION)
- ✓ RECETA (VENOMINACION, NIVEL_VIP(CULTAV), TIEMPO_PREPARACION)
- ✓ UTILIZACION (VENOMINACION - RECETA, ALACENA, POSICION, ESTANTE)
- ✓ INGREDIENTE (ALACENA, ESTANTE, POSICION, DESCRIPCION)

~~Introduciendo las relaciones entre creaciones y utilización, logre tener la ejecución de Nivel en un tablero correspondiente. (de que una receta puede no tener creador, o un ingrediente no ser utilizado.)~~

2)

PREST_LIBROS (IV_EJ, IV_SOC, FECHA_INIC_PREST, FECHA_FIM_PREST,
IV_LIBRO, NUM_LIBRU, EDITORIAL, NUM_SOCIO, APELLIDOS_SOCIO,
EMAIL_SOCIO)

• 2FN: claramente NO estan en 2FN, ya que todos los atributos -salvo FECHA_FIM_PREST- NO dependen de forma completa de la clave primaria. Por ende tenemos esto en 3FN.

~~FIN~~ Muestra una descomposición y me ni estan en 3FN:

FIN_PRESSTA_MU (IV_EJ, IV_SOC, FECHA_INI_PREST, FECHA_FIM_PREST)
 | | | |
 DF1 DF1 DF1 DF1

EJEMPLAR_LIBRO (IV_EJEMPLAR, IV_LIBRO)
 | |
 DF2 DF2

2s

EVITURIAL (IU-LIANO, NOMBRE-LIANO, EVITURAL)

OF 3

CARACTERISTICAS-SOCIO (IU-SOCIO, NOMBRE-SOCIO, APELL-SOCIO, DNI-SOCIO)

OF 4

La descomposición genera las 4 relaciones funcionales
y todos los atributos.

- 2FN: Cada atributo no-primo (en este caso el atributo que no es la clave primaria), depende de forma completa
de la clave primaria. Tiene esto en 2FN
- 3FN: No hay dependencias transitivas de una clave a
un atributo no-primo, por lo que está en 3FN.

Además se lossness JOIN que puede ir luego un
JOIN de:

FIN-PROGRAMA con CARACT-SOCIO mediante IU-SOCIO,
y luego un JOIN de esa tabla con ~~EVITURAL~~
~~EVITURAL~~ ^{EJEMPLO} IUD-ESTADISTICAL, y luego un JOIN final entre
con ~~EVITURAL~~ ^{EJEMPLO} IUD-LIANO y ~~EVITURAL~~
~~EVITURAL~~

Recupero la tabla original. (no tenemos las claves)

3)

~~EVITURAL~~ ~~LAMAN~~ ~~CARACTER~~

3)

	<u>LÍNEA</u>	<u>RAMA\</u>	<u>CABECERA</u>
i)	✓ 10	A	PALEAMO
	10	A	AEROPARQUE
	42	A	RIVER
	42	A	ANCONCIUAV UNIV.
	42	B	RIVER
	42	B	CIUDAD UNIV.

ii)

	<u>LÍNEA</u>	<u>TOTAL - RAMAS\</u>	<u>CANT - UNI\UNIV\</u>
	10	1	20 ✓
	107	2	50

4) El problema de completitud de un dato consiste en la falta de los mismos. En este caso podríamos tener que en la base de datos muchos maestros de estudiantes a profesores aparezcan como Nulos o, ~~sin~~ sin más datos, o alguna variable. ~~de la que no puedes saber si el mail es real.~~

Un modelo GAM sería:

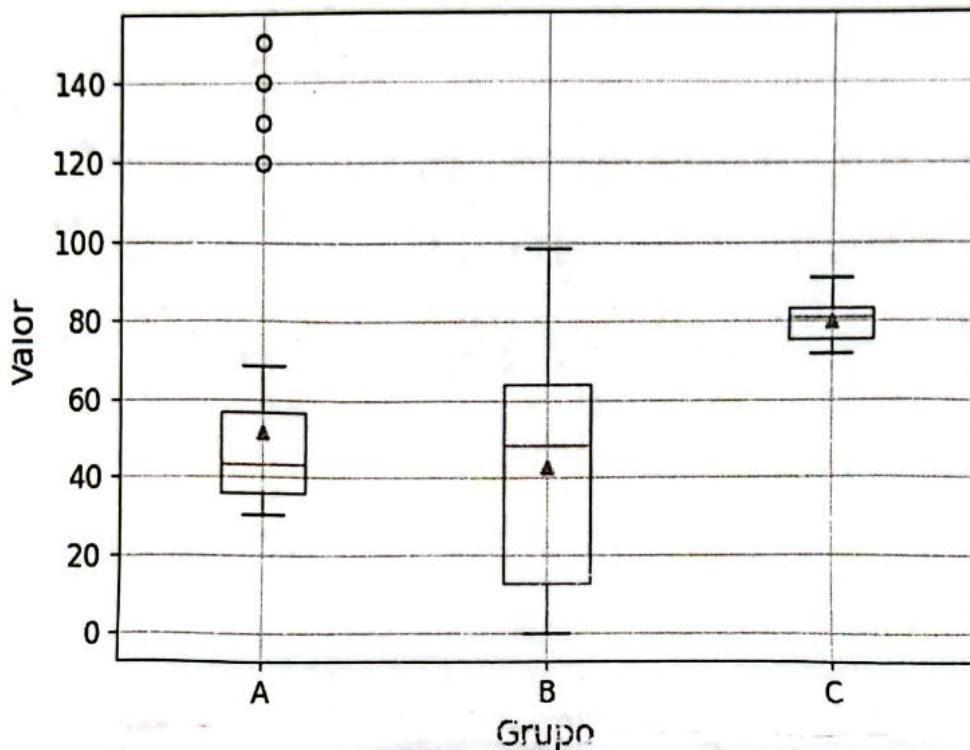
G: No ~~quiero~~ tienes datos faltantes en la columna mail.

Q: ¿Que proporción de filas no tienen mail? (estás incompletas)?

M: #filas sin mail / #filas totales.

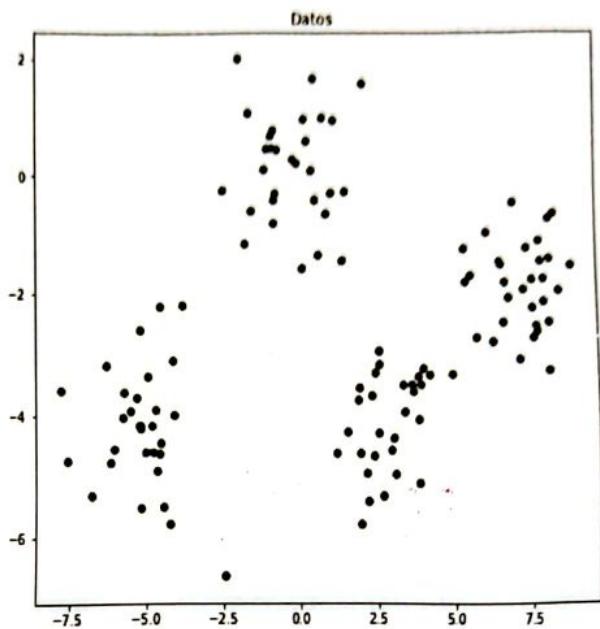
per %?

5. (10 p) Dado el gráfico de caja correspondiente a los grupos A, B y C que se muestran a continuación:



- a. Ordene los grupos de manera creciente según: i) su media, ii) su mediana.
- b. ¿Qué medida de tendencia central preferiría utilizar para comparar a los grupos A y B? Justifique.
- c. ¿Cuál de los grupos muestra mayor rango intercuartílico (IQR)? ¿Y mayor rango? Justifique.
6. (10 p) Un equipo de científicxs es contactado por una consultora para un proyecto de detección de bots en una red social. El equipo de la consultora ya tiene un algoritmo que predice si una cuenta proviene de un bot o no, pero las predicciones no son suficientemente buenas. Las predicciones positivas (bot) del algoritmo son posteriormente revisadas por personal humano para constatar si realmente se trata de un bot. El equipo de científicxs desarrolló un nuevo algoritmo de clasificación bot/no bot basado en aprendizaje automático. Luego de utilizar el algoritmo y de la posterior revisión de las cuentas, el personal que hizo las verificaciones, reportó que:
- Con el sistema original, de las cuentas revisadas, sólo 1 de cada 10 eran bots, el resto eran cuentas asociadas a personas.
 - Con el sistema nuevo, 2 de cada 10 cuentas revisadas resultaron ser bots.
- a. ¿Qué métrica de performance se está utilizando?
- b. ¿Cuál es el valor de la métrica en el sistema original? ¿Y en el nuevo?
- c. ¿Se podría utilizar otra métrica? ¿Por qué?

7. (10 p) Considerar el siguiente conjunto de puntos en el plano.



V

Indicar:

- Cuál método de clustering recomendaría, de los vistos en clase.
- Para el método elegido, indicar qué hiperparámetros tiene.
- Para cada uno de los hiperparámetros indicados, estimar el valor que debería asignarse para realizar un buen agrupamiento.
- Indique al menos 1 ventaja y 1 desventaja del método elegido.

8. (15 p) Decidir V o F y justificar de manera concisa en ambos casos.

- (1) • Si un algoritmo de clasificación tiene 98% de exactitud, entonces tiene un alto valor de recall. F
- (2) • PCA es una técnica de reducción dimensional que se basa en la hipótesis de que los datos, aunque estén en un espacio de dimensión alta, están cerca de un subespacio de dimensión menor. V
- (3) • El error cuadrático medio (en algoritmos de regresión) depende de la escala a la que se encuentran los datos. V
- (4) • De los algoritmos de clustering vistos en clase, DBSCAN es el mejor manejando datos atípicos. V
- (5) • Antes de entrenar un modelo de clasificación con árboles de decisión es conveniente reescalar los datos. F

8,5/
10 (5)

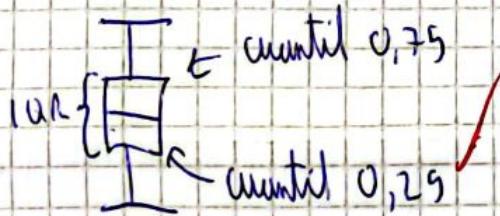
a) i) Medio: $B < A < C$ ✓
Mediano: $A < B < C$ ✓

→ Normales o gaussianas

b) El hecho de que el grupo A presente muchos outliers me indica que la mediana es la mejor medida en la mediana. ✓
entre se debe a que la media es el estimador constante que minimiza el error cuadrático medio, (método de los cuadrados mínimos) el ECM para el criterio de los cuadrados mínimos.

cantiles) No elegimos la media porque es influenciada por valores extremos (por como se calcula). Con el dibujo *

d) $IQR = C(A \sim)$.



Mo muy mucho que quieren perfectos
alguno el que compone las II
diferencias entre el primer cuantil y
el tercero.

NO CONTESTA SOBRE RANGO. $RANGO = MAX - MIN \rightarrow A$ es el grupo

e) SIST-ORIG = 1 de 1/10 non fols ? de mayor RANGO

SIST-NUEVO = 2 de 1/10 non fols

en donde

f) se estan usando la PRECISION para que no erratum los TRUE POSITIVES respecto a las clasificaciones positivas (distinguiendo los fols)

(y el resto en la pág 4)

b) 1/10 en los peores el entorno mejor y 2/10 en el entorno peor.

c) En este caso es muy importante tener en cuenta el LECTUR también. pero de ésta forma no se tiene una buena idea del desempeño ^{global} global del modelo.

El modelo ~~seguir~~ ^{nuevo} puede clasificar lotes con mejores segundas, pero eso no implica que los lotes de forma correcta. Pueden ser que solo obtenga 1 de cada 100 lotes, pero que sea el igual tener ~~0,2 de precision~~ ^{mujer} ~~mujer~~ ^{precision} que un modelo que obtenga 99 de 100, pero sea muy ligeramente menor ^{precision}.

7)

a) ~~cuantos clusters tiene dividido~~

? ^{No entiendo} cuantos clusters tiene el dato ($k=4$), si el cluster tiene tamaño y densidad similar, y estos son aproximadamente iguales. Esto indica que K-MEANS ha sido algoritmo ideal para este caso.

b) La cantidad de centroides del algoritmo (k)

c) $k=4$ al ser clara la cantidad de clusters.

d) Ventajas: - Es muy rápido y simple de implementar.

Deshventajas: - Es muy sensible a outliers

8)

- a) Puedes tener un dataset con un 2% de datos positivos, y que el modelo clasifique todo el dataset como negativo. En este caso el recall es 0 y el accuracy 0,98.

5

- b) PCA se aplica PCA idealmente en los casos donde existen un subconjunto de dimension menor que ~~aproxima~~ dimensiones al que aproximadamente pertenezcan los datos.

- c) De analizar para el enunciado es debatible, ya que PCA no cumple ninguna hipótesis sobre las variables aleatorias para que el método funcione. Simplemente lleva un problema de minimización.

d) $ECM(X, X_{\text{prev}}) = \sum (X_i - \bar{x}_i^{\text{prev}})^2$

Si recalo los datos por uno de c :

$$ECM(X, X_{\text{prev},c}) = c^2 \sum (X_i - \bar{x}_i^{\text{prev}})^2$$

Por lo que efectivamente depende de la trazo

~~los cuales dependen de la rotación del sistema de coordenadas lineales e independientes de los datos~~

- e) URGSCAN es el único algoritmo que tiene en cuenta la existencia de outliers. k-means y ~~deben~~ el método ^{Por} permanecen dentro, se basan en ^{comparar} distancias únicamente entre puntos. ¿Cuál es esto lo hace más eficiente a otros?

2^o

e) Los valores de ~~deben~~ tienen la particularidad de no ~~ser~~ ser un resultado al tener los resultados ~~un~~ continuo. Esto se debe a que la medida de moda y ~~antes~~ se hace en base a ~~una~~ a ~~una~~ medida de impureza, y no distancias (como por ej KNN).