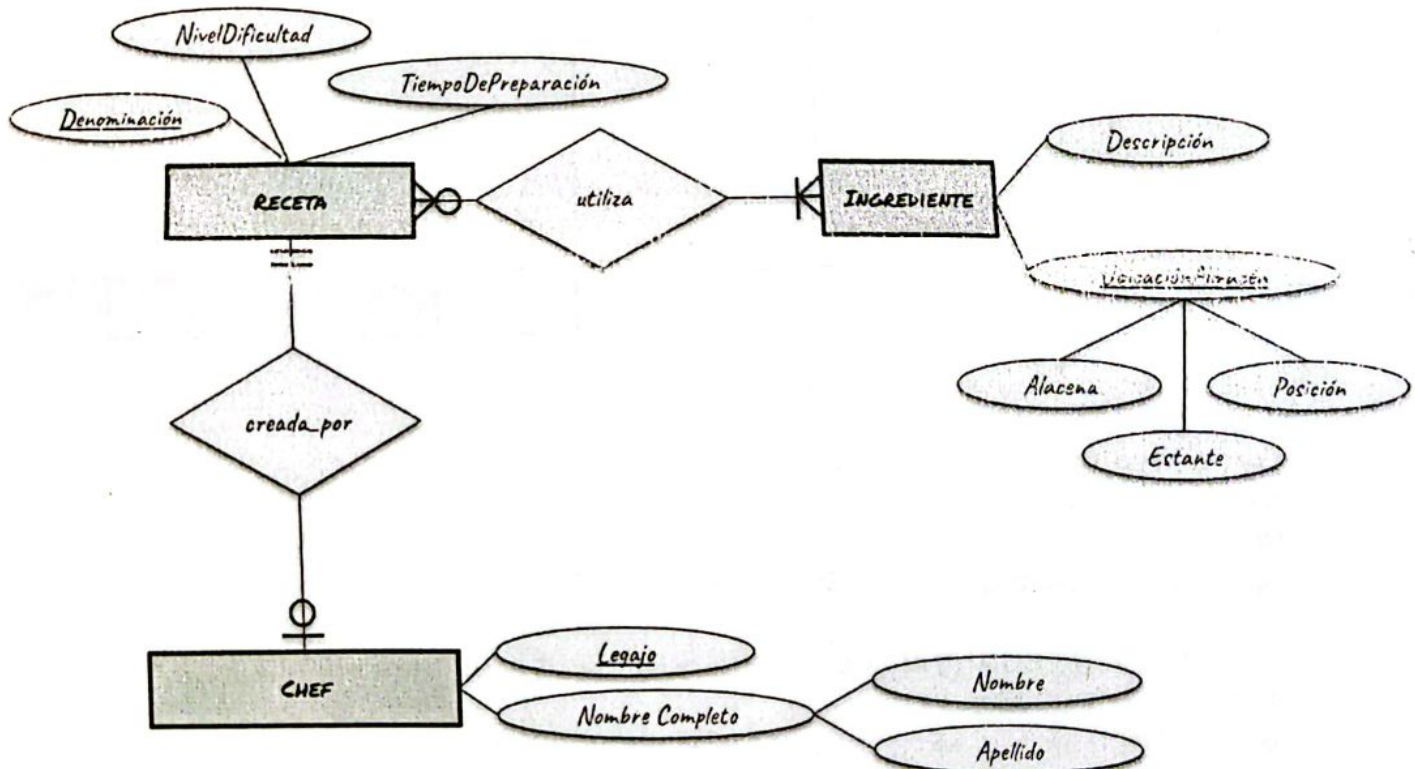


Grilla de puntajes. No completar.

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.

- (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.



- (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

PRETAMOS_LIBROS(idEjemplar, idSocio, fechaInicioPrestamo,
fechaFinPrestamo, idLibro, nombreLibro, Editorial,
nombreSocio, apellidoSocio, emailSocio)

Dependencias Funcionales

DF1: {idEjemplar, idSocio, fechaInicioPrestamo} -> {fechaFinPrestamo}

DF2: {idEjemplar} -> {idLibro}

DF3: {idLibro} -> {nombreLibro, Editorial}

DF4: {idSocio} -> {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

<u>Línea</u>	<u>Ramal</u>	CantUnidades
42	A	30
42	B	40
107	A	28
107	B	22
10	A	20

TERMINAL

<u>Línea</u>	<u>Ramal</u>	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 agregas la cabecera misma)*

ORDER BY c.Línea **ASC** *Misma línea 42.*

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

1) ✓ CHEF (LEGATO, NOMBRE, ~~CONJUNTO~~ APELLIDOS, DENOMINACION-RECETA)

~~CREACIONES~~ ~~LEGATO~~ ~~APELLIDOS~~ ~~DENOMINACION~~

✓ RECETA (DENOMINACION, NIVEL-DIFICULTAD, TIEMPO-PREPARACION)

✓ UTILIZACION (DENOMINACION-RECETA, ALACENA, POSICION, ESTANTE)

✓ INGREDIENTE (ALACENA, ESTANTE, POSICION, DESCRIPCION)

~~Introduciendo las relaciones entre capacidades y utilización, hay que estar la opinión de nous en las tablas correspondientes (de que una receta puede no tener creador, o un ingrediente no ser utilizado.)~~

2)

PREST- LIBROS (IV-ET, IV-SOC, FECHA-INIC-PRST, FECHA-FIN-PRST,
IV-LIBRO, NOM-LIBRO, EDITORIAL, NOM-SOCIO, APELL ^{SOCIO} ~~LIBRO~~,
EMAIL-SOCIO)

• 2FN: Claramente NO esta en 2FN, ya que todas las atribuciones salvo FECHA-FIN-PRST NO dependen de forma completa de la clave primaria. Por ende tampoco esta en 3FN.

~~Hay~~ Hay una descomposición y meo si esta en 3FN:

FIN-PRSTAMO (IV-ET, IV-SOC, FECHA-IN-PRST, FECHA-FIN-PRST) ✓

DF1

EJEMPLAR-LIBRO (IV-EJEMPLAR, IV-LIBRO) ✓

DF2

2)

EDITORIAL (ID-LIBRO, NOMBRE-LIBRO, EDITORIAL)

OP3

CARACTERISTICAS-SOCIO (ID-SOCIO, NOMBRE-SOCIO, APELL-SOCIO, EMAIL-SOCIO)

OP4

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

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

Además en LOSSLESS JOIN ya que puede ir luego un JOIN de:

FIN - PRISTAMO con CARACT-SOCIO mediante ID-SOCIO, y luego un JOIN de esa tabla con ~~EDITORIAL-LIBRO~~ según ID-~~EDITORIAL-LIBRO~~, y luego un JOIN final con ~~EDITORIAL-LIBRO~~ según ID-LIBRO.

Reemplazo la tabla original. (Una herramienta las bases)

37

~~LIBRO~~ ~~RAMAL~~ ~~CATEGORIA~~

3)

LINEA	RAMAL	CABECERA
i) 10	A	DALEAMO
10	A	AEROPARQUE
42	A	RIVER
42	A	CIUDAD UNIV.
42	B	RIVER
42	B	CIUDAD UNIV.

ii)

LINEA	TOTAL - RAMALES	CANT - UNIDADES
10	1	20
107	2	50

4) El problema de completo de un dato consiste en la falta de los mismos. En este caso podríamos ser que con la lista de datos muchas mails de estudiantes o profesores aparezcan como NULL o, ~~como strings vacíos~~, o alguna variante. ~~de la que no puede ser el mail real.~~

Un modelo GQM sería:

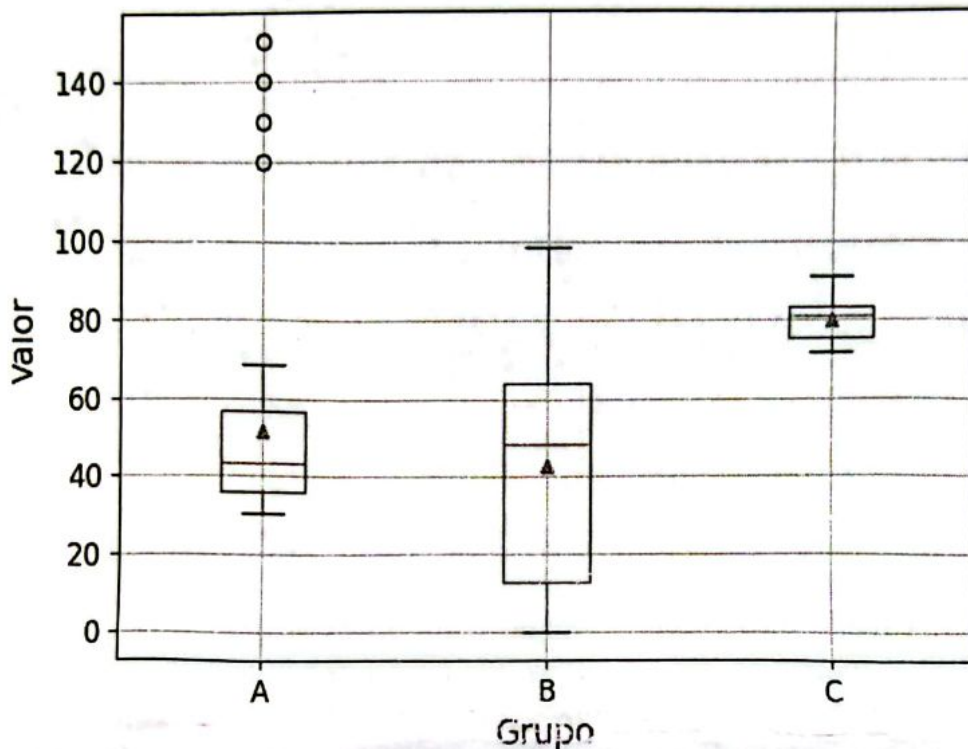
G: No ~~quiero~~ tener datos faltantes en la columna mails.

Q: ¿Qué proporción de filas no tienen mail (están incompletas)?

M: $\# \text{FILAS SIN MAIL} / \# \text{FILAS TOTALES}$.

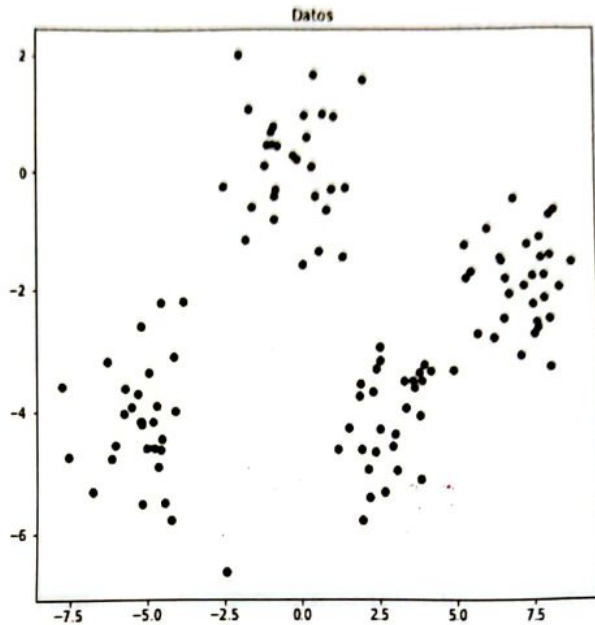
por 100?

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



- Ordene los grupos de manera creciente según: i) su media, ii) su mediana.
 - ¿Qué medida de tendencia central preferiría utilizar para comparar a los grupos A y B? Justifique.
 - ¿Cuál de los grupos muestra mayor rango intercuartílico (IQR)? ¿Y mayor rango? Justifique.
6. (10 p) Un equipo de científicos 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 científicos 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.
- ¿Qué métrica de performance se está utilizando?
 - ¿Cuál es el valor de la métrica en el sistema original? ¿Y en el nuevo?
 - ¿Se podría utilizar otra métrica? ¿Por qué?

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



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 ventajas y 1 desventaja del método elegido.

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

- Si un algoritmo de clasificación tiene 98% de exactitud, entonces tiene un alto valor de recall. **F**
- 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**
- El error cuadrático medio (en algoritmos de regresión) depende de la escala a la que se encuentran los datos. **V**
- De los algoritmos de clustering vistos en clase, DBSCAN es el mejor manejando datos atípicos. **V**
- Antes de entrenar un modelo de clasificación con árboles de decisión es conveniente reescalar los datos. **F**

8,5/10

5)

a) Medias: $B < A < C$ ✓
Mediana: $A < B < C$ ✓

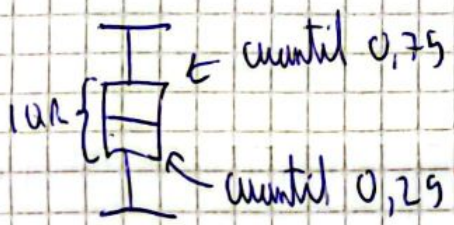
* Normales o gaussianas

b) El hecho de que el grupo A presente muchas outliers me indica que la mejor medida es la mediana. ✓
Esto se debe a que la media es el elemento central que minimiza el error cuadrático medio (matemáticamente o intuitivamente). No elegimos la media porque es influenciada por valores extremos (por como se calcula).

La media minimiza el ECM para distribuciones normales

Con el dibujo *

c) $IQR = C < A < B$ ✓



No hay mucho que explicar, perfecto. ✓
Observa el gráfico y comprueba los dibujos entre el primer cuartil y el tercero. ✓

NO CONTESTA SOBRE RANGO. $RANGO = MAX - MIN$ → (A) es el grupo de mayor RANGO

b) SIST. ORIG = 1 de 10 ran lat
SIST. NUEVO = 2 de 10 ran lat

a) se está viendo la precisión ya que se evalúan los TRUE POSITIVES respecto a las clasificaciones positivas (clasi como lat)

(b y el resto en la pag 4)

b) $2/10$ en la ~~vent~~ el sistema mayo y $2/10$ en el sistema nuevo.

c) En este caso es muy importante tener en cuenta el RECALL también. ~~pero~~ de otro ~~pero~~ más tiene una buena idea del desempeño ^{global} ~~concreto~~ del modelo.

El modelo ~~seguir~~^{nuevo} puede limpiar los con menos seguidores,
pero eso no implica que los diente de forma constante.
Puede ser que solo diente 1 de cada 100 los, pero
~~que es~~ e igual tener ~~0,2~~ de ~~presencia~~ mayor presencia
que un modelo que diente 99 de c/100, pero sea
ligeramente menor precio.

7

a) ~~some clay plaster lines decorated~~

como la cantidad de clusters es clara ($K=4$), y el cluster tiene tamaño y densidad similares, y estos son aproximadamente iguales. Esto indica que K-MEANS es un algoritmo ideal para este caso.

b) la cantidad de centros de Argentina (k)

c) $k=4$ al ser claro la cantidad de aulas.

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

Disadvantages: - En may result a outliers

8)

a) \boxed{F} : puede 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.

b) \boxed{V} PCA Se aplica PCA idealmente en los casos donde existe un subespacio de dimensión menor ~~que~~ ~~aproximado~~ ~~hacia~~ al que aproximadamente pertenecen los datos.

✓ De cualquier forma el enunciado es delatador, ya que PCA no asume ninguna hipótesis sobre las variables aleatorias para que el método funcione. Simplemente resuelve un problema de minimización.

$$c) \text{ECM}(X, X_{\text{PREV}}) = \sum (X_i - X_i^{\text{PREV}})^2$$

Si se recala los datos por una c :

$$E(M(X, X_0, \mu_0)) = c^2 \sum (x_i - x_0^{\mu_0})$$

Por lo que efectivamente depende de la traducción.

~~the causal factors, the relations of the variables to the response~~
~~linear or independent of the variables~~

5) VASCAN es el único algoritmo que tiene en cuenta la estructura de outliers. K-means y ~~los~~ el método principal ^{solo} se basan en ^{comparar} distancias ~~simplemente~~ entre puntos. ~~esto~~ Esto los hace susceptibles a outliers.

¿Por qué?

