I. Prueba conocimientos básicos. Convocatoria 2025A08. Personal Experto en Inteligencia Artificial y Ciencia de Datos

Perfiles Inteligencia Artificial y Ciencia de Datos

PREGUNTAS COMUNES PARA AMBOS PERFILES

1. Considere la siguiente ecuación e indique a qué distribución de probabilidad conocida corresponde, así como sus parámetros:

$$\frac{1}{\sqrt{8\pi}} e^{-\frac{(x-1)^2}{8}}$$

a) Normal con media -1 y desviación típica 4.

b) Normal con media 1 y desviación típica 2.

c) Lognormal con media -1 y desviación típica 4.

d) Lognormal con media 1 y desviación típica 2.

2. Al realizar una descomposición en autovectores de la matriz de covarianza de los datos observa que existe un autovalor cercano a 0. ¿Qué implicaciones tiene esto?

a) Hay una alta varianza en los datos en la dirección del autovector correspondiente.

b) Existe colinealidad entre las variables.

c) Las variables son independientes entre sí.

d) El modelo que se ha desarrollado se está comportando de manera no lineal.

3. Se plantea un problema de clasificación binaria en el que la relación de clases sigue una proporción de 99% para la clase mayoritaria y 1% para la clase minoritaria. A la hora de diseñar un estudio de diferentes modelos supervisados de clasificación y teniendo en cuenta que ambas clases son importantes para el problema a resolver, ¿qué métrica sería más adecuada para validar qué modelo produce mejores resultados?:

a) Accuracy (porcentaje de aciertos).

b) Cobertura de la clase mayoritaria.

c) Cobertura de la clase minoritaria.

d) Área bajo la curva ROC (Receiver Operating Characteristics).

4. Está usted realizando ingeniería de características para un problema de clasificación de documentos, y ha decidido emplear un método del tipo bolsa de palabras (bag of words) para transformar los textos a vectores. Suponiendo que ha configurado este método para contar bigramas de palabras, ¿cuál de los siguientes vectores sería una representación posible del texto "El reloj de la pared de la cocina no funciona"?

a) [1, 1, 1, 1, 1, 1, 1, 1, 1]

b) [1, 1, 2, 1, 1, 1, 1, 1]

c) [1, 1, 1, 1, 1, 2, 1]

d) [3, 1, 1, 1, 1, 1, 1]

5. Participa usted en un proyecto donde se plantea instalar en una plaza pública una serie de cámaras que, mediante algoritmos de visión artificial, identifiquen en tiempo real a las personas que cometan actividades delictivas, como pudiera ser la realización de pintadas ilegales. Bajo el reglamento europeo de la IA, ¿en qué nivel de riesgo debería catalogarse este tipo de sistema?

a) Nivel de riesgo mínimo.

b) Nivel de riesgo de transparencia.

c) Nivel de riesgo alto.

d) Nivel prohibido.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

6. Indique el resultado que se generaría al ejecutar el siguiente programa en R:

```
ff <- function(n) {
  f <- c()
  i <- 2
  while (i < n) {
      if (n %% i == 0) {
          f <- c(f, i)
          n <- n / i
      }
      else i <- i + 1
  }
  if (n > 2) f <- c(f, n)
  return(f)
}
print(ff(70))
```

a) 7 2 5.

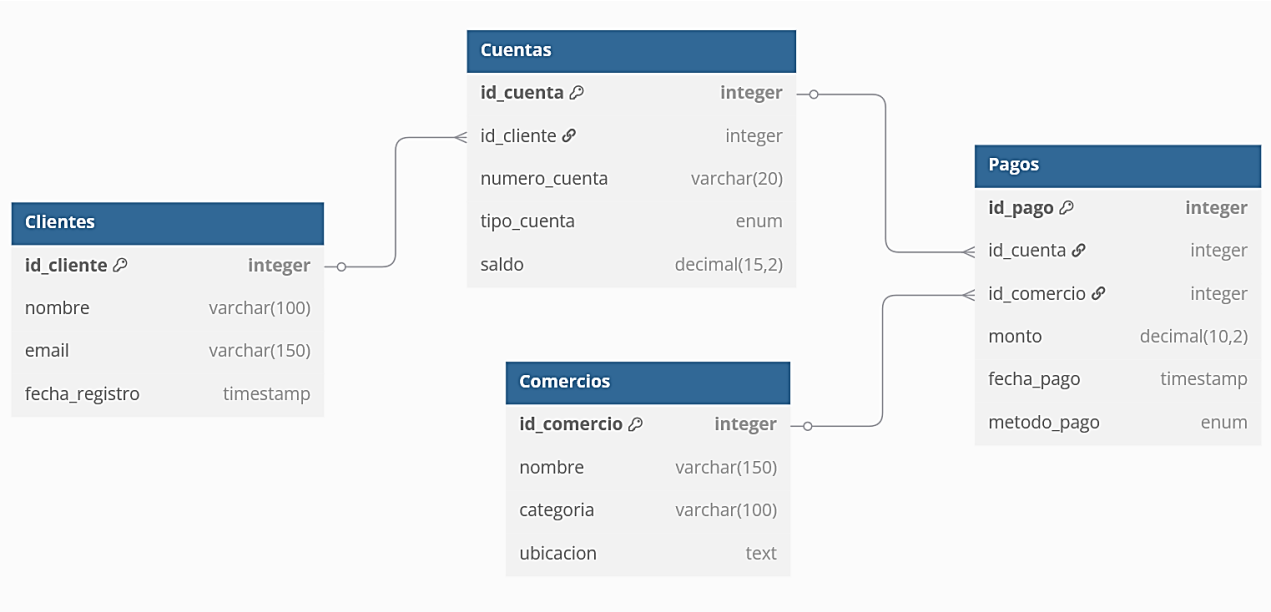
b) 7 10.

c) 2 5 7.

d) 70.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

7. Analice la siguiente estructura de base de datos:



¿Cuál sería la consulta SQL adecuada para obtener los nombres de los 10 clientes que han realizado pagos por mayor importe total de todas sus compras en el comercio "Tienda ABC" durante el mes de marzo de 2025?

a)

```
SELECT c.nombre, SUM(p.monto) AS tgas
FROM Pagos p
JOIN Cuentas cu ON p.id_cuenta = cu.id_cuenta
JOIN Clientes c ON cu.id_cliente = c.id_cliente
JOIN Comercios co ON p.id_comercio = co.id_comercio
WHERE co.nombre = 'Tienda ABC'
AND strftime('%Y-%m', c.fecha_registro) = '2025-03'
GROUP BY c.nombre
ORDER BY tgas DESC
LIMIT 10;
```

b)

```
SELECT c.nombre, SUM(p.monto) AS total_gastado
FROM Pagos p
JOIN Cuentas cu ON p.id_cuenta = cu.id_cuenta
JOIN Clientes c ON cu.id_cliente = c.id_cliente
JOIN Comercios co ON p.id_comercio = co.id_comercio
WHERE co.nombre = 'Tienda ABC'
AND strftime('%Y-%m', p.fecha_pago) = '2025-03'
GROUP BY c.id_cliente
ORDER BY p.monto DESC
LIMIT 10;
```

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

c)

```
SELECT c.nombre, MAX(p.monto) AS total_gastado
FROM Pagos p
JOIN Cuentas cu ON p.id_cuenta = cu.id_cuenta
JOIN Clientes c ON cu.id_cliente = c.id_cliente
JOIN Comercios co ON p.id_comercio = co.id_comercio
WHERE co.nombre = 'Tienda ABC'
AND strftime('%Y-%m', p.fecha_pago) = '2025-03'
GROUP BY c.id_cliente
ORDER BY total_gastado DESC
LIMIT 10;
```

d)

```
SELECT c.nombre, SUM(p.monto) AS tgas
FROM Pagos p
JOIN Cuentas cu ON p.id_cuenta = cu.id_cuenta
JOIN Clientes c ON cu.id_cliente = c.id_cliente
JOIN Comercios co ON p.id_comercio = co.id_comercio
WHERE co.nombre = 'Tienda ABC'
AND strftime('%Y-%m', p.fecha_pago) = '2025-03'
GROUP BY c.id_cliente
ORDER BY tgas DESC
LIMIT 10;
```

8. Se encuentra usted desarrollando un modelo de aprendizaje supervisado para predecir impagos en préstamos bancarios. Analice la siguiente muestra de datos que tiene disponible, e indique si debe tener en cuenta alguna consideración legal o ética respecto al uso de estos datos:

Nombre	Apellido	Sexo	Edad	Importe en cuenta	Hipoteca	Número de impagos
Juan	Pérez	M	34	12,500.75	True	2
María	Gómez	F	28	8,200.00	False	0
Carlos	López	M	45	20,300.50	True	3
Laura	Fernández	F	39	15,750.90	False	1
Pedro	Ramírez	M	50	5,600.30	True	5
Ana	Torres	F	31	9,850.60	False	0
Miguel	Soto	M	42	18,200.10	True	2
Daniela	Ruiz	F	25	7,900.75	False	0

a) Según el reglamento europeo de la IA, no es legal implementar un sistema de IA que realice perfilado de personas, y dado que este sistema está orientado a detectar perfiles de riesgo de impago, no sería legal.

b) Los campos Sexo y Edad deberían eliminarse de la tabla de datos a la hora de entrenar el modelo, para mitigar así que el modelo pueda tomar decisiones en base a estos datos personales.

c) La tabla no debería contener información personal como el Nombre y el Apellido, ya que el Reglamento General de Protección de Datos prohíbe mantener esta información.

d) No se dispone de suficientes variables explicativas como para realizar un modelo con suficiente precisión, lo cual resultaría en un perjuicio injusto para los clientes del banco.

9. Se dispone a realizar un análisis sobre un fichero en formato CSV, del cual se muestran a continuación sus primeras filas:

```
Restaurant,City,User,Score,VisitDate,Type
La Trattoria,New York,Alice,4.5,2024-03-10,Italian
Sushi Zen,Los Angeles,Bob,3.8,2024-02-25,Japanese
El Taco Loco,Houston,Charlie,5.0,2024-03-15,Mexican
Steakhouse 101,Chicago,David,4.2,2024-01-20,Steakhouse
Vegan Delight,San Francisco,Emily,3.5,2024-02-10,Vegan
Pasta Paradise,Miami,Frank,4.7,2024-03-05,Italian
BBQ Central,Dallas,Grace,4.0,2024-01-28,Barbecue
Curry House,Seattle,Henry,3.9,2024-02-18,Indian
Ocean Bites,Boston,Irene,4.8,2024-03-12,Seafood
Burger Haven,Denver,Jack,3.6,2024-01-30,American
```

Debe encontrar la puntuación media que los usuarios han otorgado a cada tipo de restaurante en cada ciudad. Para evitar utilizar datos antiguos, deberá filtrar aquellos para los que la puntuación se haya asignado antes del 1 de febrero de 2024. Así mismo, no deberá tener en cuenta en sus cálculos aquellos restaurantes que hayan recibido menos de 5 críticas. Para todo ello ha elaborado el siguiente script en Python:

```
import pandas as pd
df = pd.read_csv("reviews.csv")
df["VisitDate"] = pd.to_datetime(df["VisitDate"])
df = df[df["VisitDate"] >= "2024-02-01"]
restaurant_counts = df["Restaurant"].value_counts()
[FALTANTE]
```

Indiqué qué código debería introducirse en las líneas faltantes del script:

a)

```
df = df[df["Restaurant"].isin(restaurant_counts[restaurant_counts > 5].index)]
df.groupby(["City", "Type"]).mean()["Score"]
```

b)

```
df = df[df.isin(restaurant_counts[restaurant_counts >= 5])]
df.groupby(["Type", "City"]).mean()["Score"]
```

c)

```
df = df[df.isin(restaurant_counts[restaurant_counts > 5].index)]
df.groupby(["Type", "City"])["Score"].mean()
```

d)

```
df = df[df["Restaurant"].isin(restaurant_counts[restaurant_counts >= 5].index)]
df.groupby(["City", "Type"])["Score"].mean()
```

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

10. Se le plantea el diseño de una aplicación de analítica que necesita poder ejecutar consultas sobre una gran cantidad de información de compras realizadas por usuarios en una web de comercio online, con el fin de detectar redes de clientes que compran los mismos productos, y así hacer poder recomendaciones del tipo "otros clientes como tú también compraron…". El sistema en producción emplea una base de datos de tipo SQL para su operativa transaccional habitual; sin embargo, es necesario diseñar un sistema paralelo que permita realizar estas analíticas, para no entorpecer la actividad de producción del sistema actual. Teniendo en cuenta estos requisitos, ¿cuál de las siguientes tecnologías de base de datos sería más adecuada para este fin?

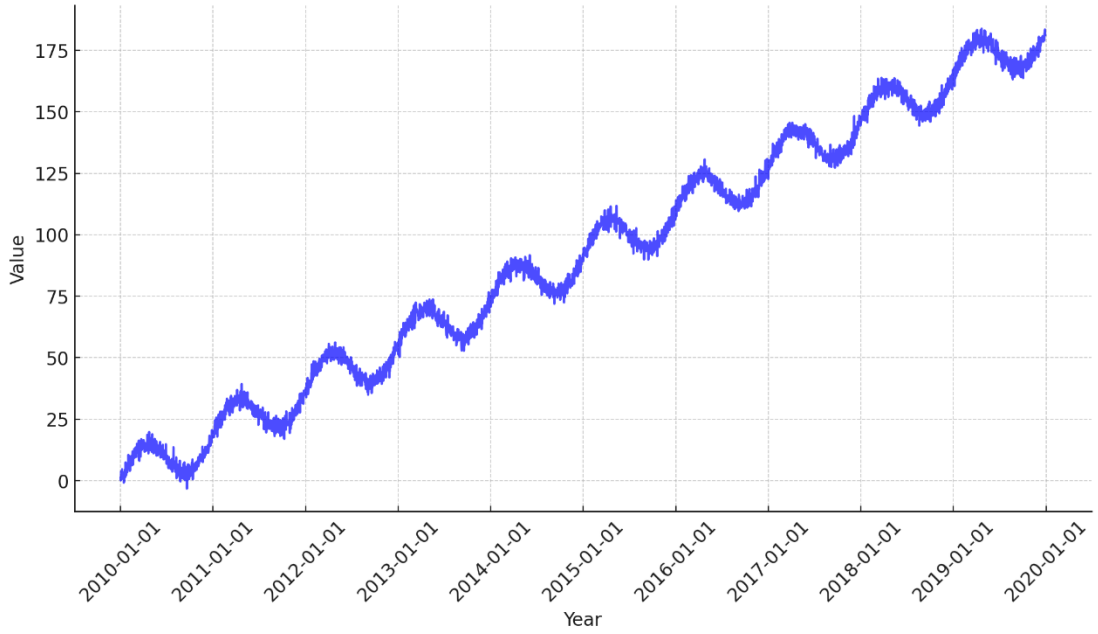
a) Neo4j.

b) MongoDB.

c) PostgreSQL.

d) Apache Cassandra.

11. Se encuentra usted afrontando un problema de predicción de series temporales con resolución diaria. Al visualizar la serie histórica obtiene la siguiente gráfica:



Indique cuál sería la forma más adecuada de preprocesar los valores de esta serie antes de usarlos como entrada para un modelo predictivo:

a) Restar de cada valor su valor anterior en la serie, para después calcular el seno y coseno de los valores resultantes.

b) Restar de cada valor su valor posterior en la serie, para después restar una media de los valores procesados que tienen lugar el mismo día en meses anteriores.

c) Descartar todos los valores que no correspondan a los últimos 3 años.

d) Restar de cada valor su valor anterior en la serie, para después restar una media de los valores procesados que tienen lugar el mismo día en años anteriores.

12. Analice el siguiente código, para el que ya se han realizado los imports necesarios y se ha creado un contexto de Spark en la variable sc:

```
stopwords = sc.textFile("stopwords.txt").map(lambda x: x.strip().lower()).collect()
doc_rdd = sc.textFile("doc.txt")

words = doc_rdd \
    .flatMap(lambda line: line.lower().split()) \
    .filter(lambda word: word and word not in stopwords)

word_counts = words.map(lambda word: (word, 1)) \
                   .reduceByKey(lambda a, b: a + b) \
                   .sortBy(lambda pair: -pair[1])

print(word_counts.collect())
```

Considere también los siguientes ficheros de datos:

“stopwords.txt”

```
en
un
la
el
a
de
```

“doc.txt”

```
En un lugar de la Mancha, de cuyo nombre no quiero acordarme
En nombre de todos los presentes en este lugar
```

¿Cuál sería el resultado de la ejecución del programa?

a) `[('lugar', 2), ('nombre', 2), ('no', 1), ('todos', 1), ('los', 1), ('presentes', 1), ('mancha', 1), ('cuyo', 1), ('quiero', 1), ('acordarme', 1), ('este', 1)]`

b) `[('lugar', 2), ('nombre', 2), ('no', 1), ('todos', 1), ('los', 1), ('presentes', 1), ('Mancha', 1), ('cuyo', 1), ('quiero', 1), ('acordarme', 1), ('este', 1)]`

c) `[('en', 3), ('de', 3), ('lugar', 2), ('nombre', 2), ('un', 1), ('la', 1), ('no', 1), ('todos', 1), ('los', 1), ('presentes', 1), ('mancha', 1), ('cuyo', 1), ('quiero', 1), ('acordarme', 1), ('este', 1)]`

d) `[('lugar', 2), ('nombre', 2), ('todos', 1), ('los', 1), ('presentes', 1), ('mancha', 1), ('cuyo', 1), ('quiero', 1), ('acordarme', 1), ('este', 1)]`

13. Analice el siguiente fichero "companies.json", del cual se muestran las primeras líneas:

```
{
  "companies": [
    {
      "name": "TechNova Solutions",
      "industry": "Information Technology",
      "country": "USA",
      "number_of_employees": 500,
      "founded": 2012,
      "revenues": {
        "2022": 85000000,
        "2023": 91000000
      }
    },
    {
      "name": "GreenLeaf Organics",
      "industry": "Agriculture",
      "country": "United Kindom",
      "number_of_employees": 120,
      "founded": 2008,
      "revenues": {
        "2022": 18500000,
        "2023": 19700000
      }
    },
     …
```

¿Cuál de los siguientes scripts en R calcula correctamente la media de ganancias de las compañías de cada país?

a)

```
library(jsonlite)
library(dplyr)

json_data <- fromJSON("companies.json")
companies_df <- json_data$companies

revenues_2023 <- companies_df %>%
  mutate(
    country = country,
    revenue_2023 = revenues$`2023`,
    .keep = "none"
  )

average_revenue_per_country <- revenues_2023 %>%
  group_by(country) %>%
  mutate(avg_revenue_2023 = mean(revenue_2023))

print(average_revenue_per_country)
```

b)

```
library(jsonlite)
library(dplyr)

json_data <- fromJSON("companies.json")
companies_df <- json_data$companies

revenues_2023 <- companies_df %>%
  mutate(
    country = country,
    revenue_2023 = revenues.2023,
  )

average_revenue_per_country <- revenues_2023 %>%
  group_by(country) %>%
  summarise(avg_revenue_2023 = mean(revenue_2023))

print(average_revenue_per_country)
```

c)

```
library(jsonlite)
library(dplyr)

json_data <- fromJSON("companies.json")
companies_df <- json_data$companies

revenues_2023 <- companies_df %>%
  mutate(
    country = country,
    revenue_2023 = revenues$`2023`,
    .keep = "none"
  )

average_revenue_per_country <- revenues_2023 %>%
  group_by(name) %>%
  summarise(avg_revenue_2023 = mean(revenue_2023))

print(average_revenue_per_country)
```

d)

```
library(jsonlite)
library(dplyr)

json_data <- fromJSON("companies.json")
companies_df <- json_data$companies

revenues_2023 <- companies_df %>%
  mutate(
    country = country,
    revenue_2023 = revenues$`2023`,
    .keep = "none"
  )

average_revenue_per_country <- revenues_2023 %>%
  group_by(country) %>%
  summarise(avg_revenue_2023 = mean(revenue_2023))

print(average_revenue_per_country)
```

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

14. Analice el siguiente código en Python, teniendo en cuenta que ya cuenta con un conjunto de datos cargado en las variables “X” e “y”:

```
01. X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    random_state=42)
02. scaler = StandardScaler()
03. X_train_scaled = scaler.fit_transform(X_train)
04. X_test_scaled = scaler.fit_transform(X_test)
05. lr = LogisticRegression()
06. param_grid = {'C': [0.1, 1.0, 10.0]}
07.
08. grid_search = GridSearchCV(lr, param_grid)
09. grid_search.fit(X_train_scaled, y_train)
10.
11. y_pred = grid_search.predict(X_test_scaled)
12. print("Accuracy:", accuracy_score(y_test, y_pred))
13. print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Indique en qué línea se ha producido un error metodológico:

a) Línea 01.

b) Línea 04.

c) Línea 08.

d) Línea 09.

15. Para una aplicación de aprendizaje supervisado se ha diseñado un modelo de clasificación cuya función de pérdida toma la siguiente forma:

$$f(x,y) = \frac{1}{2}\sum_{i=1}^{N} ReLU(y_i(w \cdot x_i)) + \lambda\left|\left|w\right|\right|_2^2$$

donde x es una lista con cada uno de los vectores de variables explicativas de los datos de entrenamiento, y es un vector con las correspondientes etiquetas de clasificación (0 o 1), y $\left|\left|w\right|\right|_2^2$ denota el cuadrado de la norma L2 o euclídea del vector w.

Suponiendo que se va a realizar el aprendizaje de este modelo mediante un algoritmo de descenso por gradiente con momento, denotando la tasa de aprendizaje como δ y el factor de momento como μ, ¿cuál de las siguientes opciones representa fielmente el método de actualización del vector de parámetros w del modelo?

a)

$$\nabla^{(t)} = \sum_{i=1}^{N} y_i x_i + 2\lambda w^{(t)}$$

$$m^{(t+1)} = \mu m^{(t)} - \delta\nabla^{(t)}$$

$$w^{(t+1)} = w^{(t)} + m^{(t+1)}$$

b)

$$\nabla^{(t)} = \sum_{i=1}^{N} y_i x_i + 2\lambda w^{(t)}$$

$$m^{(t+1)} = \delta m^{(t)} - \mu\nabla^{(t)}$$

$$w^{(t+1)} = w^{(t)} - m^{(t+1)}$$

c)

$$\nabla^{(t)} = \sum_{\substack{i=1 \\ y_i(wx_i)>0}}^{N} y_i x_i + 2\lambda w^{(t)}$$

$$m^{(t+1)} = \mu m^{(t)} - \delta\nabla^{(t)}$$

$$w^{(t+1)} = w^{(t)} + m^{(t+1)}$$

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

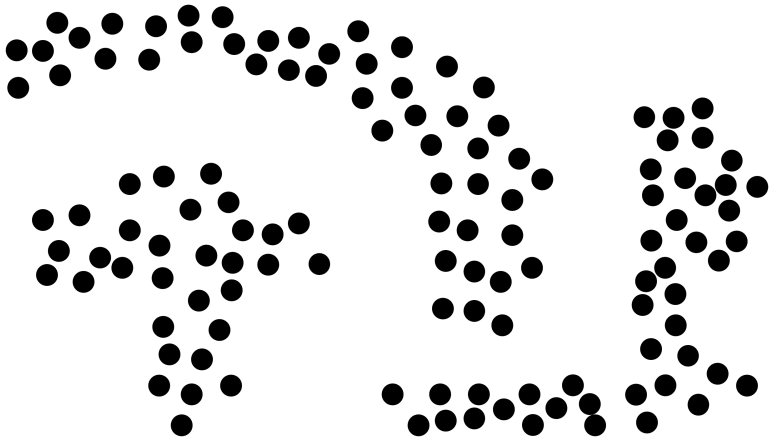
d)

$$\nabla^{(t)} = \sum_{\substack{i=1 \\ y_i(wx_i)>0}}^{N} y_i x_i + 2\lambda w^{(t)}$$

$$m^{(t+1)} = \delta m^{(t)} - \mu \nabla^{(t)}$$

$$w^{(t+1)} = w^{(t)} - m^{(t+1)}$$

16. Está trabajando en un análisis no supervisado de datos, para el que espera encontrar una serie de clusters bien diferenciados. Observa que al realizar proyecciones de baja dimensión de los datos aparecen estructuras como las siguientes:



Teniendo en cuenta esta información, ¿qué algoritmo de clustering sería más adecuado?

a) K-means.

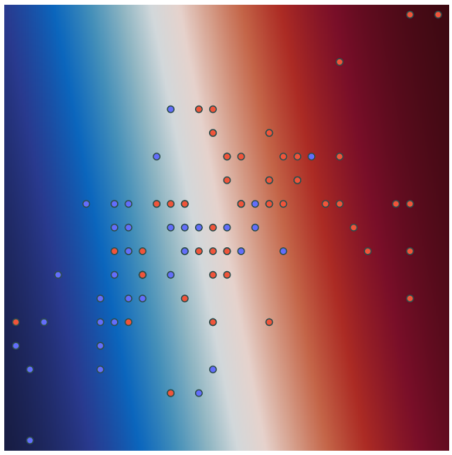
b) K-means modificado con distancias de Mahalanobis.

c) Modelo de mezcla de gaussianas.

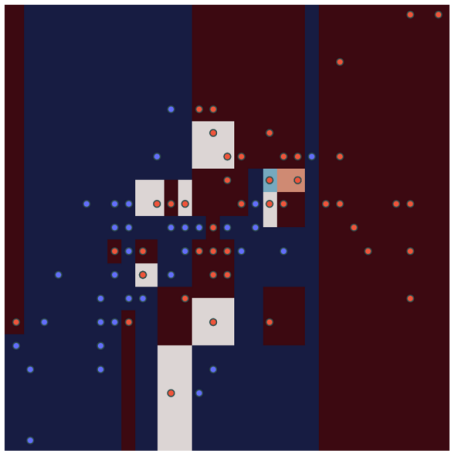
d) DBSCAN.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

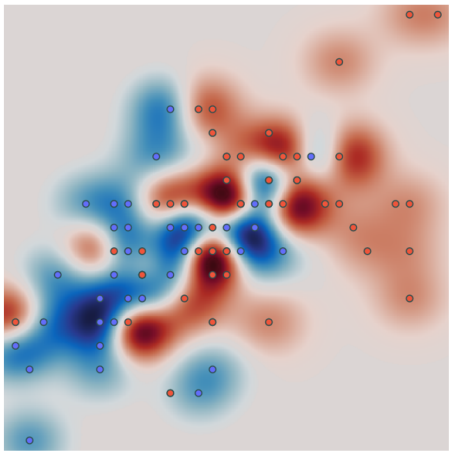
17. Se han entrenado 3 modelos de clasificación supervisada sobre el mismo conjunto de datos, obteniendo las siguientes visualizaciones de sus fronteras de decisión:



1



2



3

Indique, de entre las siguientes hipótesis, cuál de ellas es la más plausible sobre los modelos empleados para generar estas fronteras:

a) 1. Regresión Lineal, 2. Árbol de decisión, 3. SVM con núcleo gaussiano.

b) 1. Regresión Lasso, 2. Vecinos próximos, 3. MLP con activaciones sigmoidales.

c) 1. Perceptrón, 2. Random Forest, 3. MLP con activaciones ReLU.

d) 1. SVM con kernel polinómico de segundo grado, 2. Regresión isotónica, 3. Análisis discriminante lineal.

18. Al trabajar con un conjunto de datos del sector financiero descubre que existen valores NA en las variables "Nacionalidad" y "Tipo de cliente". Al indagar por el origen de estos valores le informan que se marcan con NA aquellas nacionalidades que no se han recogido durante el onboarding de los clientes, mientras que los NA en el tipo de cliente se emplean para marcar aquellos clientes de tipos no recogidos en los datos. ¿Cuál sería la forma más adecuada de tratar estas dos variables?

a) Sustituir los NA por la moda en ambas variables.

b) Crear una categoría nueva para los NA en ambas variables.

c) Crear una categoría nueva para los NA en la variable "Nacionalidad", y sustituir los NA por la moda en la variable "Tipo de cliente".

d) Sustituir los NA por la moda en la variable "Nacionalidad", y crear una categoría nueva para los NA en la variable "Tipo de cliente".

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

19. Tras entrenar un modelo de regresión lineal sobre un gran conjunto de datos de entrenamiento, se realiza un análisis de errores en un conjunto de test que se ha obtenido un año después en el proceso de recogida de datos, y se observa que el modelo tiene baja varianza pero alto sesgo, a pesar de que sobre el conjunto de entrenamiento obtenía buenos resultados. Suponiendo que la selección de las muestras de entrenamiento y de test se ha realizado correctamente, ¿cuál de las siguientes explicaciones de la situación podría ser válida?

a) La naturaleza de los datos es no lineal, por lo que no se ha escogido el tipo de modelo adecuado.

b) El modelo está infrajustado.

c) Existe una tendencia en los datos que es creciente en el tiempo.

d) Se ha empleado demasiada regularización.

20. Se encuentra usted involucrado en un proyecto de ciencia de datos donde se pretende modelar el tiempo que un cliente mantiene contratado un servicio de streaming de películas. A lo largo del proyecto se han realizado dos campañas de recogida de datos, en momentos del tiempo diferentes, y necesita garantizar que en ambos periodos el tiempo de permanencia de los clientes sigue una distribución similar. ¿Qué test estadístico podría utilizar?

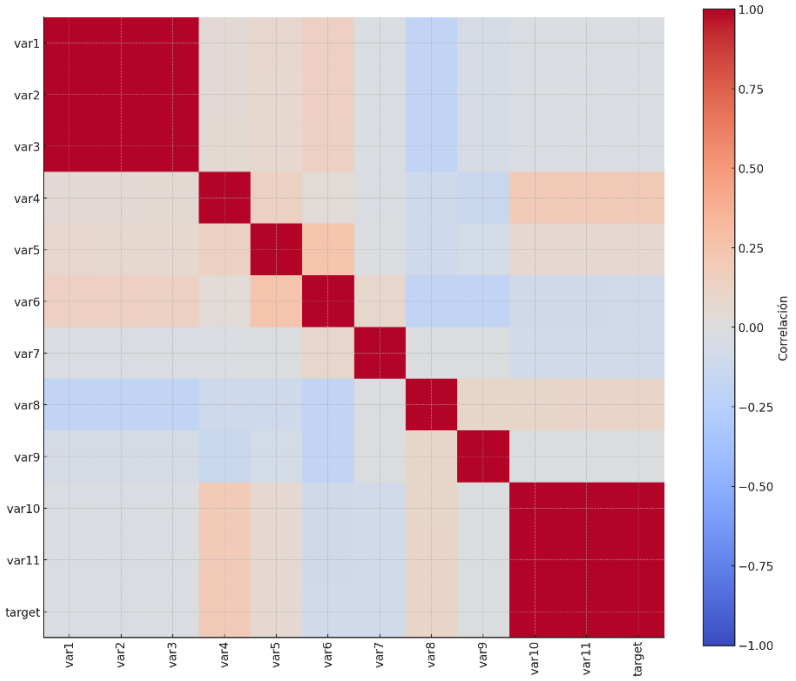
a) ANOVA.

b) t de Student.

c) Chi cuadrado.

d) Kolmogorov-Smirnov de dos muestras.

21. Analice la siguiente matriz de correlaciones:



Suponiendo que va a construir un modelo que prediga el target y que solo puede seleccionar 4 variables explicativas, ¿cuál sería la mejor selección?

a) (var4, var5, var8, var11)

b) (var2, var3, var10, var11)

c) (var2, var4, var9, var10)

d) (var1, var2, var3, var9)

22. El siguiente código incompleto en R realiza un trabajo de modelización sobre un conjunto de datos, asumiendo que ya se han importado todas las librerías necesarias y los datos ya se encuentran disponibles en train_data y test_data:

```
modelo_rf <- ranger(formula = target ~ ., data = train_data, num.trees = 500)
[CÓDIGO FALTANTE]
results <- accuracy_vec(truth = test_data$target, estimate = preds)
print(results)
```

Indique cuál es la opción más adecuada para la línea de código faltante:

a) `preds <- predict(modelo_rf, data = test_data)`

b) `preds <- predict(modelo_rf, data = test_data)$predictions`

c) `preds <- predict(modelo_rf, data = train_data)`

d) `estimate <- score(modelo_rf, data = train_data)$predictions`

23. Se le presenta el fichero de datos "ventas.csv", del cual se muestran las 10 primeras filas:

```
datetime,mes,dia_del_mes,hora_del_dia,temperatura,humedad,evento,ventas
2023-01-01 00:00:00,1,1,0,28.82,42.38,soleado,13
2023-01-01 01:00:00,1,1,1,22.0,39.65,tormenta,20
2023-01-01 02:00:00,1,1,2,24.89,84.27,lluvia,23
2023-01-01 03:00:00,1,1,3,31.2,47.22,nublado,26
2023-01-01 04:00:00,1,1,4,29.34,40.77,tormenta,22
2023-01-01 05:00:00,1,1,5,15.11,90.0,nublado,21
2023-01-01 06:00:00,1,1,6,24.75,48.55,tormenta,22
2023-01-01 07:00:00,1,1,7,19.24,37.48,soleado,26
2023-01-01 08:00:00,1,1,8,19.48,57.71,tormenta,17
```

Teniendo en cuenta que la variable objetivo "ventas" muestra un comportamiento estacional a varias escalas temporales, ¿qué proceso de ingeniería de características sería más útil?

a) mes → (sin(mes*2*pi/12), cos(mes*2*pi/12)), dia_del_mes → (sin(dia_del_mes *2*pi/31), cos(dia_del_mes *2*pi/31)), hora_del_dia → (sin(hora_del_dia *2*pi/24), cos(hora_del_dia *2*pi/24))

b) mes → (sin(mes*12*pi/2), cos(mes*12*pi/2)), dia_del_mes → (sin(dia_del_mes *31*pi/2), cos(dia_del_mes *31*pi/2)), hora_del_dia → (sin(hora_del_dia *24*pi/2), cos(hora_del_dia *24*pi/2))

c) (mes, dia_del_mes, hora_del_dia) → mes * 31 * 24 + dia_del_mes * 24 + hora_del_dia.

d) (mes, dia_del_mes, hora_del_dia) → mes + dia_del_mes * 12 + hora_del_dia * 31.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

24. Debe implementar en R una función "count_duplicates" que devuelva cuántos caracteres diferentes aparecen más de 1 vez en una cadena de entrada, sin distinción entre mayúsculas y minúsculas. Por ejemplo:

```
count_duplicates("abcde") → 0
count_duplicates("aabbcde") → 2
count_duplicates("aaAbBcde") → 2
```

¿Cuál de las siguientes implementaciones sería correcta?

a)

```
count_duplicates <- function(text) {
  text <- tolower(text)
  seen <- character()
  duplicates <- character()
  for (char in strsplit(text, "")[[1]]) {
    if (char %in% seen) {
      duplicates <- c(duplicates, char)
    } else {
      seen <- c(seen, char)
    }
  }
  return(length(duplicates))
}
```

b)

```
count_duplicates <- function(text) {
  chars <- unlist(strsplit(text, split = ""))
  chars <- tolower(chars)
  freq_table <- table(chars)
  duplicate_count <- sum(freq_table > 2)
  return(duplicate_count)
}
```

c)

```
count_duplicates <- function(text) {
  chars <- unlist(strsplit(text, split = ""))
  chars <- tolower(chars)
  duplicate_count <- sum(chars > 1)
  return(duplicate_count)
}
```

d)

```
count_duplicates <- function(text) {
  text <- tolower(text)
  chars <- strsplit(text, split = "")
  freq_table <- table(chars)
  duplicate_count <- sum(freq_table > 1)
  return(duplicate_count)
}
```

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

25. Se necesita construir un modelo de aprendizaje automático que sea capaz de identificar el idioma en el que está escrito un texto, entre los siguientes idiomas: español, inglés, portugués, ruso, griego, chino, árabe y japonés. ¿Cuál de las siguientes aproximaciones sería más práctica en términos de acierto y uso de recursos computacionales?

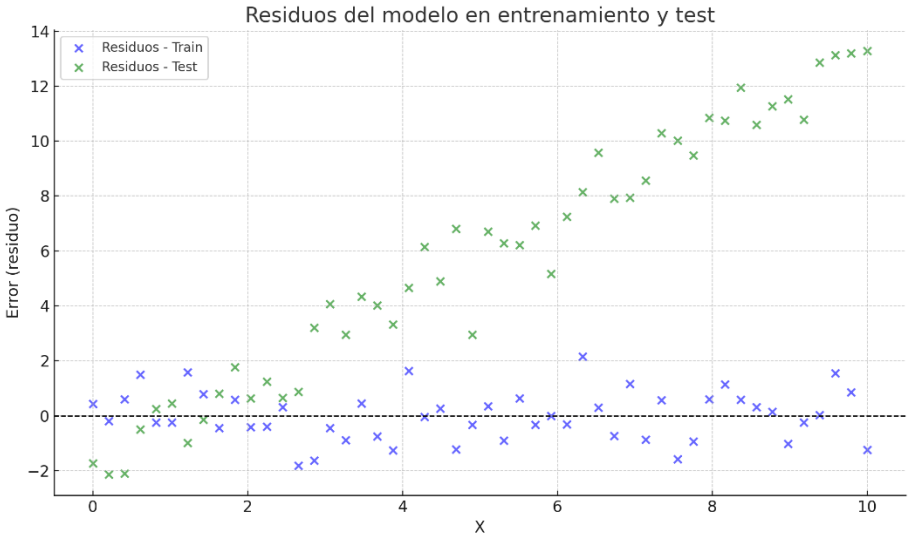
a) Frecuencias de bigramas de caracteres + Support Vector Machine (SVM) lineal.

b) Frecuencias de unigramas de caracteres + Regresión logística.

c) Latent Dirichlet Allocation.

d) Tokenización + Bidirectional Encoder Representations from Transformers (BERT).

26. Tras entrenar un modelo de regresión lineal sobre un gran conjunto de datos de entrenamiento se ha realizado el siguiente análisis de sus errores, tanto sobre el conjunto de entrenamiento original como sobre un conjunto de test recogido un año después.



¿Cuál de las siguientes afirmaciones sería una explicación factible de la situación?

a) El modelo empleado está infrajustado, porque no es capaz de capturar una tendencia lineal simple como la que se observa en el test.

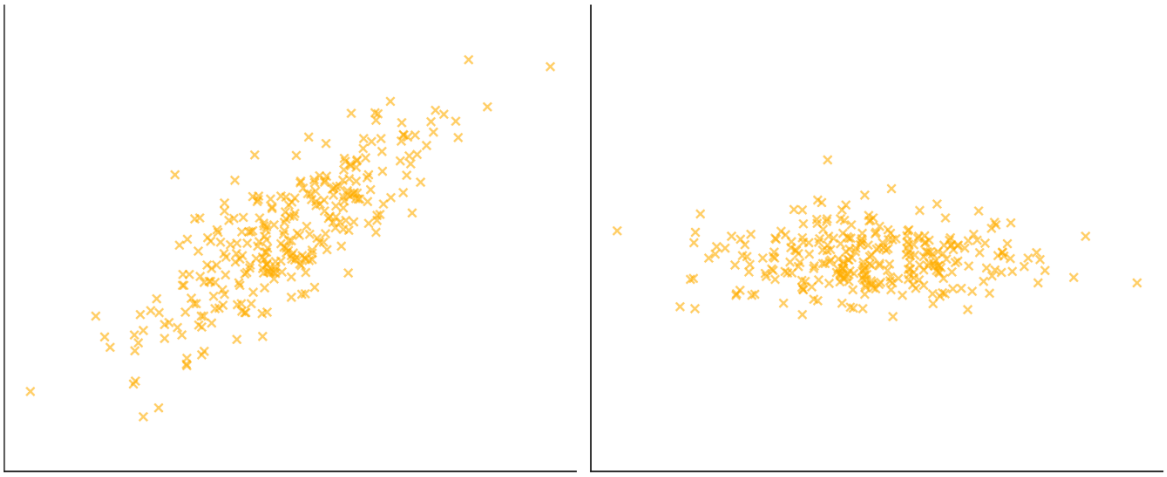
b) El modelo presenta bajo sesgo tanto en entrenamiento como en test.

c) El modelo no está haciendo uso de la variable X, y por eso comete un error mayor a medida que esta aumenta su valor.

d) En la distribución de datos subyacente la relación entre la variable X y el target ha cambiado en el último año.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

27. Analice la siguiente transformación de datos, donde la gráfica de la izquierda presenta los datos originales y la gráfica de la derecha los datos transformados. ¿Qué técnica se ha utilizado?



a) MDS (Multidimensional Scaling).

b) LOF (Local Outlier Factor).

c) ICA (Independent Component Analysis).

d) PCA (Principal Component Analysis).

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

28. Se desea implementar un proceso en Spark para unificar una serie de ficheros texto. El proceso de unificación consiste en crear un único fichero que contenga todas las líneas de los ficheros originales, pero eliminando líneas duplicadas, esto es, aquellas líneas que aparezcan en algún otro lugar, ya sea en el mismo o en diferentes ficheros, de forma que solo quede una única línea de entre todas las copias. Para ello se ha preparado el siguiente script incompleto:

```
from pyspark.sql import SparkSession
import shutil
spark = SparkSession.builder.appName("EliminarDuplicados").getOrCreate()
input_fold = "/content/entrada_txt"
output_fold = "/content/resultado"
[FALTANTE]
```

Considere también que los ficheros de entrada a cargar tienen el siguiente aspecto (se muestran las 5 primeras líneas de uno de los ficheros)

```
Asg akag1
12481 8751
Xxxxa AAA nn
Lorem ipsum
En un lugar de la Mancha, de cuyo nombre no quiero acordarme
```

¿Cuál sería la línea de código faltante?

a)

```
df = spark.read.text(input_fold).drop().write.text(output_fold)
```

b)

```
df = spark.read.csv(input_fold).distinct().write.text(output_fold)
```

c)

```
df = spark.read.text(input_fold).distinct().coalesce(1).write.text(output_fold)
```

d)

```
df = spark.read.text(input_fold).distinct().write.text(output_fold)
```

29. Una empresa situada en Estados Unidos ha creado una app para móviles, abierta a todo el mundo, que permite a sus usuarios mejorar sus hábitos alimenticios. Para ello, los usuarios introducen información sobre lo que van comiendo y la app les ofrece recomendaciones de dieta basadas en modelos de aprendizaje automático. ¿Debe esta app someterse al Reglamento Europeo sobre Inteligencia Artificial (AI Act)?

a) Sí, puesto que la app está abierta a todo el mundo, y eso incluye potenciales usuarios dentro de la Unión Europea.

b) No, dado que la empresa tiene su sede en Estados Unidos, y por tanto no le aplica el reglamento.

c) No, dado que el reglamento solo aplica a sistemas que usan IA Generativa.

d) Sí, puesto que la app va a tratar con datos relacionados con la salud, los cuales se consideran datos sensibles.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

30. Se necesita crear una base de datos para gestionar un centro médico, de forma que permita las siguientes operaciones:

- **Listar los nombres, DNIs y especialidad de todos los médicos del centro.**
- **Identificar a los pacientes por nombre y DNI.**
- **Recuperar la fecha de primer registro en el centro de un paciente.**
- **Consultar para un paciente dado el diagnóstico de las consultas que ha realizado, pudiendo filtrar por médico o tipo de consulta.**
- **Almacenar un listado de medicamentos empleados en el centro, identificados por nombre y código CIE.**

Indique cuál de los siguientes grupos de sentencias SQL construiría una serie de tablas capaz de satisfacer todos estos requisitos, evitando incluir características innecesarias.

a)

```
CREATE TABLE Medicos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL
);
CREATE TABLE Pacientes (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    FechaRegistro DATE NOT NULL
);
CREATE TABLE Consultas (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    TipoConsulta VARCHAR(100) NOT NULL,
    Diagnostico TEXT,
    MedicoID INT NOT NULL,
    PacienteID INT NOT NULL,
    FOREIGN KEY (MedicoID) REFERENCES Medicos(ID),
    FOREIGN KEY (PacienteID) REFERENCES Pacientes(ID)
);
CREATE TABLE Medicamentos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    CodigoCIE VARCHAR(20) UNIQUE NOT NULL
);
```

b)

```
CREATE TABLE Medicos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    Especialidad VARCHAR(100) NOT NULL
);
CREATE TABLE Pacientes (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    FechaRegistro DATE NOT NULL
);
CREATE TABLE Consultas (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    TipoConsulta VARCHAR(100) NOT NULL,
    Diagnostico TEXT,
    MedicoID VARCHAR(100),
    PacienteID VARCHAR(100)
);
CREATE TABLE Medicamentos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    CodigoCIE VARCHAR(20) UNIQUE NOT NULL
);
```

c)

```
CREATE TABLE Medicos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    Especialidad VARCHAR(100) NOT NULL
);
CREATE TABLE Pacientes (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    FechaRegistro DATE NOT NULL
);
CREATE TABLE Consultas (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    TipoConsulta VARCHAR(100) NOT NULL,
    Diagnostico TEXT,
    MedicoID INT NOT NULL,
    PacienteID INT NOT NULL,
    Prescripcion INT NOT NULL,
    FOREIGN KEY (MedicoID) REFERENCES Medicos(ID),
    FOREIGN KEY (PacienteID) REFERENCES Pacientes(ID),
    FOREIGN KEY (Prescripcion) REFERENCES Medicamentos(ID)
);
CREATE TABLE Medicamentos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    CodigoCIE VARCHAR(20) UNIQUE NOT NULL
);
```

d)

```
CREATE TABLE Medicos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    Especialidad VARCHAR(100) NOT NULL
);
CREATE TABLE Pacientes (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    FechaRegistro DATE NOT NULL
);
CREATE TABLE Consultas (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    TipoConsulta VARCHAR(100) NOT NULL,
    Diagnostico TEXT,
    MedicoID INT NOT NULL,
    PacienteID INT NOT NULL,
    FOREIGN KEY (MedicoID) REFERENCES Medicos(ID),
    FOREIGN KEY (PacienteID) REFERENCES Pacientes(ID)
);
CREATE TABLE Medicamentos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    CodigoCIE VARCHAR(20) UNIQUE NOT NULL
);
```

PREGUNTAS ESPECÍFICAS PARA EL PERFIL DE INTELIGENCIA ARTIFICIAL

31. Indique cuál es la explicación válida sobre la técnica "KV caché" en el contexto de los LLMs:

a) En modelos de tipo generativo, se identifican aquellos "key values" o activaciones más significativas en las capas Transformers de la red, de manera que pueden podarse aquellas secciones de la red sin estos valores, optimizando así el tamaño del modelo.

b) En los modelos con atención bidireccional se procesan los tokens uno a uno, pasándolos por todas las capas transformer antes de pasar a procesar el siguiente. Dado que este proceso es costoso computacionalmente, se almacenan en memoria las matrices de claves (keys) y valores (values) calculadas hasta el momento, reduciendo así el coste de procesar el siguiente token.

c) En modelos de tipo generativo decoder-only, se guardan en memoria los vectores de claves (keys) y valores (values) de los tokens de entrada ya procesados, de forma que si después se continúa la conversación puede ahorrar tiempo de cómputo a la hora de calcular los scores de atención entre los nuevos tokens y los ya procesados.

d) Se trata de un almacenamiento temporal y seguro de claves de acceso y otros valores relevantes como el nombre del usuario, de manera que cuando un proceso necesita descargar un LLM de repositorios populares como Hugging Face puede hacerlo sin solicitar las credenciales al usuario en cada descarga.

32. Debe diseñar un sistema que analice automáticamente una base de datos de contratos en español, marcando/clasificando aquellos párrafos que contengan cláusulas de propiedad intelectual. Cuenta usted con 1000 contratos ya anotados por expertos legales. ¿Qué tipo de modelo LLM sería el más óptimo para la clasificación?

a) T5 (Encoder-Decoder model).

b) RoBERTa-XLM (Autoencoder model).

c) GPT-4º (Autoregressive model).

d) Qwen2.5-72B-Instruct (Decoder-only model).

33. En una arquitectura del tipo Retrieval-augmented Generation (RAG), ¿cuál es el papel que cumple el modelo de embeddings?

a) Codificar cada token de entrada como un vector denso, de manera que sirva como entrada para las capas Transformer del LLM.

b) Representar cada documento o segmento de documento de la base de datos como un vector denso de números, de manera que puedan recuperarse de manera eficiente los documentos más relacionados con una query dada.

c) Permitir comparar de manera muy eficiente los diferentes documentos que se indexan en la base de datos, lo cual habilita la detección de clusters y la eliminación de documentos duplicados en grandes volúmenes documentales.

d) Cachear en un vector denso las consultas realizadas previamente por los usuarios del sistema RAG, de manera que pueda ahorrarse tiempo de cálculo si otro usuario realiza una pregunta que ya se ha planteado previamente.

34. Usted ha participado en el desarrollo de un automatismo que hace uso de un LLM comercial para estructurar documentos. El automatismo recibe documentos en forma de texto sin un formato específico, y se encarga de generar un fichero en formato JSON con una serie de campos con valores binarios que indican si el documento de entrada cumple con un conjunto de criterios. Para ello, el LLM recibe una serie de instrucciones muy detalladas (aprox. 2000 palabras) de los criterios que deben usarse para rellenar los campos del JSON, seguido del documento concreto a analizar (500-1000 palabras). El automatismo se encuentra funcionando en producción con un nivel de acierto adecuado, pero se observa que supone unos costes económicos elevados en llamadas a la API del LLM. ¿Cuál de las siguientes técnicas podría emplearse para reducir estos costes, garantizando que el sistema siga produciendo las mismas respuestas?

a) Cuantización.

b) Destilación.

c) Prompt caching.

d) Low-Rank Adapters (LoRA).

35. Debe configurar un modelo de lenguaje (LLM) para que realice tareas de traducción del inglés al español de textos de pequeño tamaño. Indique cuál de las siguientes configuraciones de generación sería más adecuada.

a)

- System prompt: "Suppose you are an expert english to spanish translator. Translate any prompt received into the spanish language."
- Model: "meta-llama/Llama-3.1-8B-Instruct "
- Temperature = 0.6
- Min new tokens = 20
- Max new tokens = 200
- Top p = 0.9

b)

- System prompt: "You are an expert english to spanish translator. Translate any text received into the spanish language."
- Model: "tiiuae/Falcon3-7B-Instruct"
- Temperature = 0.6
- Min new tokens = 200
- Max new tokens = 1000
- Top p = 0.9

c)

- System prompt: "Translate any text received in what follows into the spanish language."
- Model: "Qwen/Qwen2.5-7B-Instruct"
- Temperature = 0.6
- Min new tokens = 20
- Max new tokens = 200
- Top p = 0.001

d)

- System prompt: "Translate the following texts into the spanish language."
- Model: "mistralai/Mistral-7B-Instruct-v0.3"
- Temperature = 0.6
- Min new tokens = 20
- Max new tokens = 10
- Top p = 0.9

36. Un hiperparámetro de bastante relevancia a la hora de realizar el ajuste fino de un LLM es la tasa de aprendizaje o learning rate. Suponiendo que durante el proceso de entrenamiento no se modifica ningún otro hiperparámetro, ¿cuál de las siguientes sería la aproximación correcta para que el proceso de entrenamiento sea convergente y eficiente?

a) Mantener una tasa de aprendizaje constante, del orden de 10^{-1}.

b) Comenzar con una tasa de aprendizaje de 1, reduciéndola de manera lineal de manera que tome el valor 0 al final del entrenamiento.

c) Empezar con una tasa del orden de 10^{-5}, haciendo crecer su valor linealmente hasta 10^{-4} durante los primeros pasos del entrenamiento, para luego hacerla decrecer linealmente para que llegue a 0 al final del entrenamiento.

d) Iniciar el entrenamiento con una tasa del orden de 10^{-5}, y duplicarla tras cada época.

37. Está usted trabajando con un LLM de tamaño 24B, cuyos parámetros se encuentran en formato bfloat16. Dispone también de una versión cuantizada del mismo a formato NormalFloat4. Suponiendo que ambos modelos se cargan en la misma GPU, ¿cuál sería la mejor estimación sobre la cantidad de memoria RAM de GPU que se ocuparía tras esta carga?

a) 36 GB.

b) 60 GB.

c) 72 GB.

d) 480 GB.

38. Haciendo uso de un modelo de difusión para generación de imágenes observa que de manera reiterada las imágenes generadas son de poca calidad, aunque el proceso de generación en sí es muy rápido y se intuye que el prompt se está siguiendo correctamente. ¿Cuál sería el parámetro más apropiado a ajustar para conseguir una mejor calidad, incluso si esto aumenta los tiempos de generación?

a) Sampling steps.

b) Resolution.

c) Random seed.

d) CFG (Classifier-Free Guidance).

39. Se quiere diseñar un Agente que permita a un usuario realizar reservas de pistas en un polideportivo. Se dispone ya un servicio por API que permite realizar operaciones como consultar la disponibilidad de pistas, así como realizar reservas para una franja horaria que esté disponible. Asumiendo que toda la interacción con el usuario deberá hacerse en lenguaje natural, ¿qué arquitectura del sistema sería la más adecuada?

a) Retrieval-augmented Generation (RAG).

b) Chain-of-Thought (CoT).

c) Mixture of Experts (MoE).

d) Model Context Protocol (MCP).

40. En un sistema RAG, (Retrieval Augmented Generation), ¿cuál es el propósito de emplear un reranker dentro de dicha arquitectura de solución?

a) Reducir el tiempo de inferencia del generador al disminuir el número de documentos recuperados.

b) Mejorar la calidad de las respuestas aplicando un reordenamiento semántico más preciso sobre los documentos recuperados, generalmente, de una base de datos vectorial.

c) Aumentar la diversidad temática de los documentos entregados al generador, para así maximizar la cobertura.

d) Validar la veracidad factual de los documentos antes de generar la respuesta.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

41. Un problema recurrente en los LLMs son sus alucinaciones. ¿Qué estrategia garantiza su eliminación completa?

a) Realizar un pre-entrenamiento únicamente con datos factuales y de calidad contrastada.

b) Montar el LLM en una arquitectura RAG.

c) Realizar un prompting adecuado y emplear características avanzadas como el razonamiento.

d) No existe ninguna estrategia que garantice la eliminación de las alucinaciones.

42. Está realizando un ajuste fino de un LLM para una tarea de NER (Named Entity Recognition). Los textos de entrenamiento a procesar son razonablemente largos, por lo que debe plantear una estrategia eficiente para tratarlos, sin por ello perder efectividad en el modelado. ¿Cuál de las siguientes aproximaciones garantizaría un entrenamiento adecuado del modelo, a la par que optimizaría al máximo los tiempos de entrenamiento?

a) Procesar todos los textos usando el tokenizador del modelo, y añadir padding hasta llegar al tamaño de contexto máximo del modelo, recortando los últimos tokens de los textos con longitud mayor a ese tamaño máximo. De esta forma se consigue sacar el máximo partido al LLM.

b) Procesar todos los textos usando el tokenizador del modelo, y agrupar los textos en batches de tamaño 1 para así poder pasar cada texto por el LLM sin tener que usar padding, optimizando el consumo de memoria.

c) Preprocesar todos los textos usando el tokenizador del modelo, e implementar un cargador de datos en batches, de forma que para cada batch mida cuál es el texto más largo en número de tokens, y añada a los demás textos del batch una serie de tokens de padding hasta alcanzar ese mismo tamaño. De esta forma se consigue cuadrar todos los textos del batch al mismo tamaño, optimizando el procesado en GPU.

d) Emplear un tokenizador BPE entrenado sobre los propios datos de entrenamiento, para así maximizar la cantidad de información por token en los textos que se van a trabajar.

43. Ha desarrollado un modelo de clasificación para un banco, con el fin de clasificar si se concede una hipoteca a un potencial cliente. Debido a la regulación del sector, en caso de no conceder la hipoteca, su modelo debe proporcionar una explicación de los motivos por los que no se ha concedido. Dado que el modelo es un ensemble complejo de diferentes modelos base, propone aplicar una técnica que toma las variables explicativas del cliente y busca la modificación más pequeña de las mismas que provoca que el modelo apruebe la hipoteca, obteniendo así una explicación sobre qué factores han producido la denegación. ¿Cuál es la técnica que ha aplicado?

a) Principio de las explicaciones múltiples.

b) LIME.

c) Explicación por antagonismo.

d) Contrafactuales.

44. Analice el siguiente texto para el entrenamiento de un LLM de conversación general:

```
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can
has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis
balls. 5 + 6 = 11. The answer is 11.
```

¿Qué técnica de prompting se está empleando aquí?

a) CoT (Chain of Thought).

b) ReAct (Reason + Act).

c) RAG (Retrieval Augmented Generation).

d) Few-Shot.

45. El siguiente código hace uso de LangChain para resolver un caso de uso concreto con LLMs:

```
from langchain.chat_models import init_chat_model
from langchain_community.tools.tavily_search import TavilySearchResults
from langchain_core.messages import HumanMessage
from langgraph.prebuilt import create_react_agent

model = init_chat_model("gpt-4o-mini", model_provider="openai")
agent_executor = create_react_agent(model, [TavilySearchResults(max_results=2)])
response = agent_executor.invoke({"messages": [HumanMessage(content=USER_QUERY)]})
for message in response["messages"]:
    message.pretty_print()
```

¿Cuál de las siguientes descripciones define mejor la solución se ha implementado aquí?

a) Un agente que incluye una herramienta de búsqueda para contestar las preguntas del usuario.

b) Un sistema de RAG que permite contestar al usuario consultando información en una base de datos vectorial.

c) Una interfaz contra un modelo de OpenAI para contestar a preguntas del usuario.

d) Un sistema que genera una cadena de razonamiento en base a la consulta del usuario, antes de contestarle.

46. En el contexto de los modelos de difusión, ¿cuál de las siguientes arquitecturas de red neuronal es más frecuentemente utilizada?

a) Redes recurrentes (RNN), porque son capaces de modelar secuencias de ruido en el dominio temporal, y esencialmente un proceso de difusión es un proceso temporal de inyección de ruido.

b) Autoencoders variacionales (VAE), porque permiten una codificación latente compacta de las imágenes ruidosas.

c) Redes convolucionales U-Net con atención, dado su éxito en aplicaciones como la segmentación de imágenes.

d) Transformers puros porque pueden modelar eficientemente la estructura espacial de imágenes sin necesidad de convoluciones.

47. ¿Qué framework se utiliza para gestionar el ciclo de vida completo de los modelos?

a) MLFlow.

b) HuggingFace.

c) Scikit-learn.

d) LangChain.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

48. Analice el siguiente código Python incompleto que prepara una llamada a la API de OpenAI para continuar con una conversación entre el chatbot y el usuario:

```
from openai import OpenAI
import os
client = OpenAI(api_key=os.getenv("OPENAI_API_KEY"))
response = client.chat.completions.create(
    [CÓDIGO FALTANTE]
)
```

¿Cuál sería el bloque de código más adecuado para el bloque faltante?

a)

```
model="gpt-4o-mini",
messages=[
    {"role": "system", "content": "Eres un asistente útil que siempre contesta en español."},
    {"role": "user", "content": "Hola"},
    {"role": "assistant", "content": "Hola, ¿en qué puedo ayudarte?"},
    {"role": "user", "content": "¿Qué día es hoy?"}
]
```

b)

```
model="meta-llama/Llama-3.1-8B",
messages=[
    {"role": "system", "content": "Eres un asistente conversacional que siempre contesta en español."},
    {"role": "user", "content": "Buenos días"},
    {"role": "assistant", "content": "¡Buenos días! ¿En qué necesitas que te ayude hoy?"},
    {"role": "user", "content": "¿Qué temperatura tenemos ahora en Barcelona?"}
]
```

c)

```
model="gpt-4o-mini",
prompt="[USER] Que pasa [ASSISTANT] ¡Todo bien por aquí! ¿En qué puedo ayudarte? [USER] Ayudame a entender las ecuaciones trigonometricas porfa"
```

d)

```
model="gpt-4o-mini",
messages=[
    {"role": "assistant", "content": "¿Cómo se fabrican los lápices?"}
]
```

49. Está diseñando un sistema RAG para indexar todo el histórico de BOEs, tanto del Estado como de las Comunidades Autónomas, particionado por resoluciones. ¿Cuál de las siguientes configuraciones para una base de datos vectorial sería la más adecuada para garantizar una baja latencia sin grandes reducciones en la calidad de la respuesta?

a) Utilizar una base de datos relacional con extensiones para vectores, sin particionamiento ni índices adicionales, primando así la simplicidad de la solución.

b) Emplear una base de datos vectorial, en modo in-memory con flat index y alojada en una única instancia, minimizando así los costes de transferencia de red y de disco a memoria.

c) Hacer uso de una base de datos vectorial distribuida, con uso de índices HNSW (Hierarchical Navigable Small World Graph) y memory-mapping.

d) Búsqueda de similitudes mediante el algoritmo BM25.

50. En el caso de que se quiera desplegar modelos de lenguaje de gran tamaño (LLM) a nivel local, es decir en modo on-premises, ¿qué tecnología se emplearía para realizar la inferencia de dichos modelos?

a) TensorBoard.

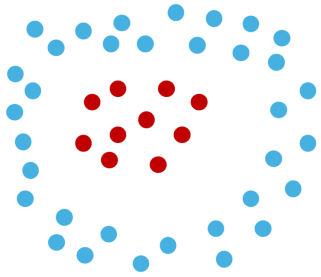
b) Apache Airflow.

c) ONNX Runtime.

d) Qdrant.

PREGUNTAS ESPECÍFICAS PARA EL PERFIL DE CIENCIA DE DATOS

51. Considere el siguiente conjunto de datos donde cada color es una clase, representado mediante un diagrama de dispersión (scatter plot). Considere también que va a entrenarse un modelo de tipo Support Vector Machine (SVM) sobre estos datos, para el que debe ajustarse el peso de regularización C, el tipo de kernel K y los parámetros de este. ¿Qué configuración de parámetros de la SVM proporcionaría mejores resultados para este problema de clasificación binaria?



a) C=0, K=gaussian, gamma=1/2

b) C=10, K=linear, gamma=3

c) C=1, K=gaussian, gamma=1

d) C=2, K=linear, gamma=0

52. Teniendo en cuenta la siguiente tabla de valores del algoritmo Q-learning, ¿qué acción realizaría a continuación un agente que estuviera en el estado s3 para el siguiente paso de explotación?

Q	a1	a2	a3
s1	-1	4	10
s2	3	-10	-3
s3	7	2	3

a) Acción a3 con 100% de probabilidad.

b) Acción a1 con 70% de probabilidad, a2 con 20% de probabilidad, a3 con 30% de probabilidad.

c) Acción a1 con 100% de probabilidad.

d) Acción a1 con 33% de probabilidad, a2 con 33% de probabilidad, a3 con 33% de probabilidad.

53. Suponga que está usted diseñando una red neuronal artificial para abordar un problema de clasificación con 3 clases excluyentes entre sí, y que cuenta con 10 variables explicativas. Además, se desea que la red devuelva una estimación de las probabilidades de pertenencia a cada una de clases. ¿Qué diseño de red sería más adecuado?

a) Input(10) → Dense(128) → ReLU → Dropout(1.0) → Dense(3) → Sigmoid

b) Input(10) → Dense(128) → ReLU → Dropout(0.1) → Dense(3) → Sigmoid

c) Input(10) → Dense(128) → ReLU → Dropout(0.1) → Dense(3) → Softmax

d) Input(10) → Dense(128) → ReLU → Dropout(1.0) → Dense(3) → Softmax

54. El algoritmo Proximal Policy Optimization (PPO) es uno de los métodos más utilizados en sistemas de aprendizaje por refuerzo. Suponga la siguiente notación relacionada con este algoritmo: π_θ la función de política que está siendo entrenada, $\pi_{\theta_{old}}$ una versión anterior de esa política, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ la razón de probabilidades de ambas políticas, A_t la ventaja o "advantage" estimada para el paso t, $clip(x, a, b)$ una función que recorta el valor de x para que esté en el intervalo [a, b], V_θ la función valor, V_{target} una estimación de lo que debería haberse obtenido por la función valor, (ϵ, c_1, c_2) hiperparámetros a ajustar. Indique cuál de las siguientes expresiones representa correctamente este algoritmo:

a) $$\min_\theta \ \mathbb{E}_t\left[\min\left(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\right) - c_1\left(V_\theta(s_t) - V_{target}(s_t)\right)^2 + c_2\, entropy[\pi_\theta](s_t)\right]$$

b) $$\max_\theta \ \mathbb{E}_t\left[\min(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t) - c_1\left(V_\theta(s_t) - V_{target}(s_t)\right)^2 + c_2\, entropy[\pi_\theta](s_t)\right]$$

c) $$\min_\theta \ \mathbb{E}_t\left[\max(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t) - c_1\left(V_\theta(s_t) - V_{target}(s_t)\right)^2 + c_2\, entropy[\pi_\theta](s_t)\right]$$

d) $$\max_\theta \ \mathbb{E}_t\left[\min(r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t) - c_1\left(V_\theta(s_t) - V_{target}(s_t)\right)^2 - c_2\, entropy[\pi_{ref}](s_t)\right]$$

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

55. Para una aplicación en el sector bancario se necesita construir un modelo supervisado que estime la probabilidad de que un potencial cliente de un préstamo incurra en morosidad. Para ello se cuenta con un conjunto de datos que recopila unas 50 variables derivadas de la información del cliente. Debido a la regulación propia del banco, este modelo debe ser representable como una serie de reglas de fácil explicación. Teniendo en cuenta estas limitaciones, ¿qué modelo sería el más adecuado desde un punto de vista de explicación de su resultado?

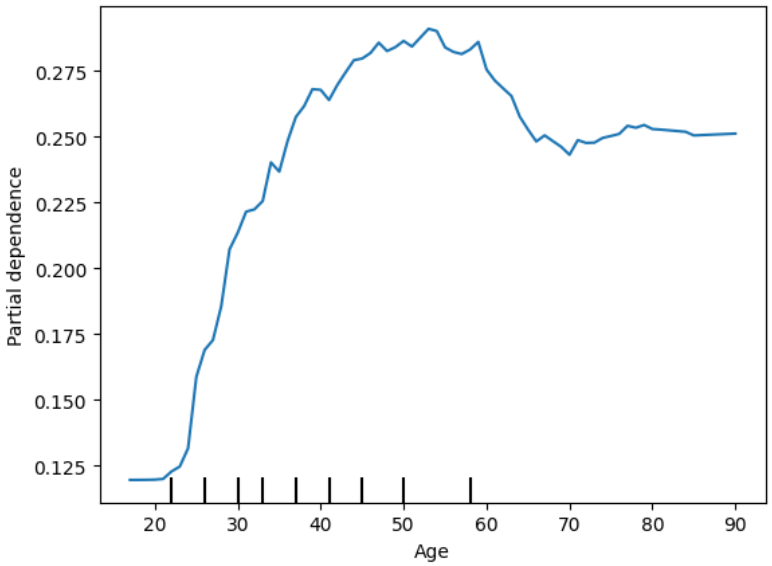
a) Árbol de decisión con profundidad máxima limitada a 3 niveles.

b) Red neuronal con una única capa oculta, formada por un número pequeño de neuronas.

c) Regresión Ridge.

d) SVM con un alto valor del parámetro de regularización.

56. Se ha entrenado un modelo usando Extreme Gradiente Boosting (XGB) sobre un dataset cuyas variables explicativas son información clínica de un paciente, y la variable objetivo es si tal paciente va a sufrir complicaciones tras una intervención quirúrgica (clase positiva = complicaciones, clase negativa = sin complicaciones). Para poder conseguir explicabilidad de este modelo se ha generado la siguiente gráfica de dependencias parciales. ¿Cuál de las siguientes conclusiones sería una interpretación válida de esta gráfica?



a) Según el modelo, la probabilidad de sufrir una complicación aumenta tras los 20 años de edad, llegando a su pico en torno a los 50 años, y reduciéndose un poco antes de los 60 para alcanzar un nivel estable, todo ello considerando el efecto de otras variables explicativas.

b) Según los datos recogidos, la probabilidad de sufrir una complicación aumenta tras los 20 años de edad, llegando a su pico en torno a los 50 años, y reduciéndose un poco antes de los 60 para alcanzar un nivel estable, todo ello considerando el efecto de otras variables explicativas.

c) Según el modelo, la probabilidad de sufrir una complicación aumenta tras los 20 años de edad, llegando a su pico en torno a los 50 años, y reduciéndose un poco antes de los 60 para alcanzar un nivel estable, todo ello sin considerar el efecto de otras variables explicativas.

d) Según los datos recogidos, la probabilidad de sufrir una complicación aumenta tras los 20 años de edad, llegando a su pico en torno a los 50 años, y reduciéndose un poco antes de los 60 para alcanzar un nivel estable, todo ello sin considerar el efecto de otras variables explicativas.

57. Tras entrenar un clasificador de tipo Random Forest sobre una serie de datos de entrenamiento observa que en el conjunto de test el error de clasificación es significativamente más alto que en el conjunto de entrenamiento. ¿Cuál sería una medida que contribuiría a reducir este sobreajuste?

a) Emplear un único árbol de decisión en lugar de un Random Forest, ya que es claro que el modelo está sobreajustando por haber ensamblado más de un árbol.

b) Utilizar en los árboles un criterio de impureza que no provoque sobreajuste, como es el criterio de entropía.

c) Reducir la fuerza α del Cost Complexity Pruning produciendo así una poda más ligera en los árboles.

d) Modificar la configuración de los árboles para aumentar el número mínimo de datos que deben existir en un nodo para que este pueda ser dividido.

58. Se tiene un dataset de 50 variables numéricas, para el cual se desea construir una visualización informativa y de alta densidad sobre las correlaciones cruzadas entre todas ellas. ¿Qué tipo de gráfica sería más adecuada?

a) Tabla formada por 50x50 subgráficas, cada una de ellas siendo una gráfica de dispersión de dos de las variables del dataset.

b) Diagrama de cuerdas (Chord diagram).

c) Mapa de calor con 50x50 celdas, cada una codificando en su color el valor de la correlación de Pearson entre dos variables.

d) Gráfica de violín.

59. ¿Cuál de los siguientes conjuntos de ecuaciones representa correctamente el funcionamiento de una celda LSTM (Long short-term memory) estándar?

a)

$$f_t = tanh\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$h_t = o_t \odot \tanh\,(c_t)$$

b)

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$h_t = o_t \odot \tanh\,(c_t)$$

c)

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big)$$

$$i_t = tanh(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$o_t = tanh(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$c_t = f_t + c_{t-1} + i_t + tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$h_t = \sigma(c_t)$$

d)

$$f_t = \sigma\big(W_f \cdot h_{t-1} + x_t + b_f\big)$$

$$i_t = \sigma(W_i \cdot h_{t-1} + x_t + b_i)$$

$$o_t = \sigma(W_o \cdot h_{t-1} + x_t + b_o)$$

$$c_t = (f_t + i_t) \odot tanh(W_c \cdot h_{t-1} + x_t + b_c)$$

$$h_t = o_t + c_t$$

60. ¿Cuál es el principal aporte de Extreme Gradient Boosting (XGB) sobre los métodos de Gradient Boosting estándar?

a) Añade un mecanismo de regularización que penaliza la complejidad de los nuevos árboles añadidos al ensemble.

b) Permite abordar problemas de regresión múltiple.

c) Aceleración del proceso de entrenamiento, mediante el cálculo de histogramas de las variables continuas, de forma que puedan probarse varios puntos de corte relevantes de forma muy eficiente.

d) Permiten tratar los missing values sin que hayamos tenido que imputarlos nosotros previamente.

61. Se le plantea construir un sistema predictivo para una entidad que realiza leasing de automóviles. El modelo debe ser capaz de predecir el valor de venta que tendrá dentro de 4 años un vehículo comprado a fecha de hoy. Además, la entidad necesita también una estimación del riesgo inherente a cada predicción de manera flexible, dando una distribución completa por predicción, de forma que pueda usar esta información para reservar fondos que puedan cubrir posibles pérdidas causadas por errores en la predicción. ¿Qué modelo de machine learning permitiría cumplir estos requisitos?

a) Regresión lineal.

b) Gradient Boosting.

c) Support Vector Regression no lineal, y análisis posterior del hiperparámetro épsilon óptimo obtenido.

d) Proceso gaussiano.

62. Según la teoría de los ensembles de modelos, ¿cuál de las siguientes configuraciones de ensamblado ofrecería una ventaja sobre emplear únicamente uno de los modelos base? Considere que cada modelo del ensemble se entrena sobre ligeras variaciones del mismo conjunto de entrenamiento.

a) Ensemble de N regresiones lineales del tipo ridge.

b) Ensemble de N árboles de decisión sin poda.

c) Ensemble de N árboles de decisión podados.

d) Ensemble de N perceptrones multicapa con decaimiento de pesos.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

63. Un agente de aprendizaje por refuerzo está ejecutando el algoritmo DQN (Deep Q-Network). Observa que, aunque al inicio del entrenamiento el agente se comporta de manera muy variada, explorando diferentes formas de actuación y obteniendo algunos resultados muy positivos de forma esporádica, a partir de un momento dado se ancla a un comportamiento subóptimo y no modifica más esta política. ¿Cuál sería la estrategia más adecuada para mejorar este proceso de entrenamiento?

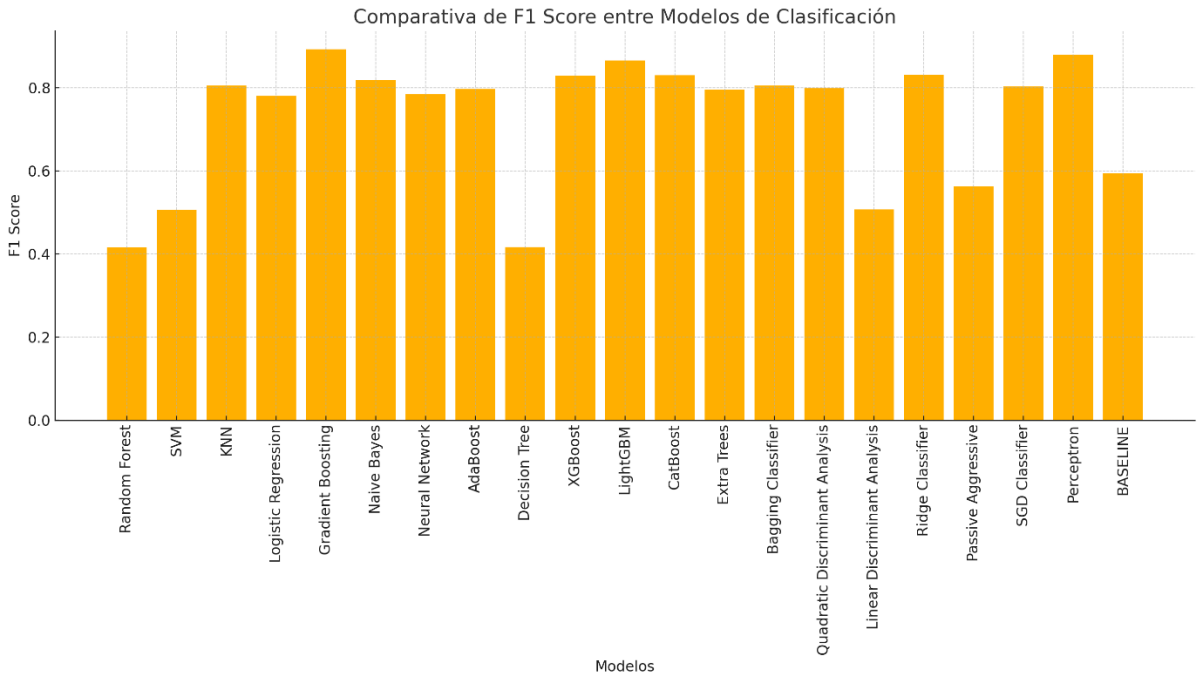
a) Aumentar el tamaño de la memoria de experience replay.

b) Continuar el entrenamiento durante más pasos.

c) Reducir el número de pasos necesarios para actualizar la target network.

d) Aumentar la tasa de aprendizaje.

64. Se ha generado la siguiente gráfica para comparar los resultados de varios modelos de clasificación sobre un problema, incluyendo también una solución básica de referencia (BASELINE). ¿Cómo podría mejorarse esta visualización para presentar la información de manera más clara?



a) Ordenar las columnas por F1 Score, y reemplazar la columna BASELINE por una línea horizontal.

b) Colorear las columnas según el valor de F1 conseguido, siguiendo una escala perceptualmente uniforme.

c) Eliminar aquellas columnas que tengan un F1 Score inferior al baseline, ya que son modelos que no tiene sentido desplegar en la práctica.

d) Ordenar los modelos alfabéticamente, y hacer que el origen del eje Y coincida con el modelo de menor rendimiento.

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

65. Está usted participando en el diseño de un test clínico para una enfermedad pulmonar, el cual se basa en características obtenidas de un análisis de sangre. El test se implementará mediante un modelo de machine learning, el cual deberá determinar la probabilidad de que el paciente analizado esté afectado por la enfermedad, y el objetivo del proyecto es desplegar tal modelo en hospitales de todo el país. Para el entrenamiento del modelo se han recogido muestras de 5 centros diferentes. Un análisis descriptivo de los datos revela que medias y las correlaciones entre las variables son ligeramente diferentes entre los datos recogidos en hospitales diferentes. ¿Cuál sería la estrategia de evaluación más adecuada para el modelo que está creando?

a) Validación cruzada de 5 hojas, cada hoja conteniendo los datos de un hospital. De este modo cada hoja de validación es de una distribución diferente a los datos vistos en entrenamiento.

b) Validación cruzada de 10 hojas, concatenando previamente todos los datos de entrenamiento y reordenándolos aleatoriamente. De esta manera se garantiza la robustez del modelo, ya que en cada hoja de entrenamiento se verá expuesto a una mezcla lo más diversa posible de datos.

c) Creación de un modelo por separado con los datos de cada hospital, y ensamblado de los modelos. En este esquema la evaluación se realizaría con todo el conjunto de entrenamiento, pero empleando para clasificar cada dato solo los modelos que no se han entrenado con el grupo de datos al que pertenece. De esta forma se saca el máximo partido a los datos disponibles.

d) Que la distribución de datos en cada hospital sea diferente revela un fallo grave en el procedimiento de recogida de datos. Este fallo se propagará al modelo si se entrena con estos datos, por lo que no podemos entrenar y validar un modelo con confianza en esta situación.

66. Debe implementar un modelo de clasificación para determinar la probabilidad de que un potencial cliente de una entidad crediticia devuelva un préstamo a tiempo. La tabla de datos disponible para ello es la siguiente (se muestran las 10 primeras filas):

```
Cliente previo,Alto importe,Estudios,Nivel salarial,Ocupación,Aval,Devolución
True,True,Primaria,Medio,Empleado,Personal,True
False,False,Primaria,Medio,Empresario,Hipotecario,True
False,False,Universitario,Bajo,Autónomo,Sin aval,False
True,False,Primaria,Medio,Empleado,Personal,True
True,False,Primaria,Medio,Empleado,Sin aval,True
True,False,Universitario,Alto,Empleado,Personal,False
True,False,Secundaria,Alto,Empleado,Hipotecario,True
False,False,Universitario,Bajo,Empleado,Sin aval,True
True,True,Primaria,Medio,Autónomo,Personal,True
```

¿Qué modelo sería más adecuado para esta tarea?

a) Proceso Gaussiano.

b) SVM.

c) Random Forest.

d) Lasso.

67. En un problema de clasificación multiclase ha obtenido la siguiente matriz de confusión:

	Pred A	**Pred B**	**Pred C**
Real A	20	0	0
Real B	0	20	0
Real C	0	20	20

¿Cuál es el F1-score macro averaged?

a) 7/9

b) 3/4

c) 2/3

d) 5/6

68. Indique qué algoritmo de optimización, de tipo gradiente estocástico, está basado en el ajuste dinámico de los momentos estadísticos de distintos órdenes para actualizar los pesos en redes neuronales:

a) Optimizador basado en descenso estocástico sin memoria de pasos anteriores.

b) Algoritmo que incorpora una anticipación del gradiente en su actualización.

c) Método que ajuste de forma adaptativa las contribuciones de los gradientes pasados y actuales.

d) Técnica que suaviza la trayectoria del gradiente mediante un factor de inercia.

69. Se plantea resolver un problema de NER (Named Entity Recognition) sobre documentos de varios cientos de palabras, empleando para ello una arquitectura de red neuronal recurrente. ¿Qué configuración de capas sería más adecuada, considerando que no deben añadirse características innecesarias?

a)

- Embedding
- RNN (Recurrent Neural Network)
- Softmax por palabra

b)

- Embedding
- GRU (Gated Recurrent Unit)
- Softmax por palabra

c)

- Embedding
- LSTM (Long Short-Term Memory)
- Softmax por palabra

d)

- Embedding
- LSTM (Long Short-Term Memory) bidireccional
- Softmax por palabra

70. Necesita crear un modelo para un dataset de gran tamaño, del orden de millones de datos de entrenamiento con decenas de miles de variables. ¿Cuál de los siguientes modelos y algoritmos sería el más eficiente en tiempo de entrenamiento para esta situación?

a) Red neuronal entrenada mediante descenso estocástico por gradiente.

b) Regresión lineal por mínimos cuadrados mediante cálculo de la pseudoinversa.

c) Random Forest, siguiendo su algoritmo estándar de entrenamiento.

d) Support Vector Machine con kernel gaussiano, entrenado mediante LIBSVM.

PREGUNTAS DE RESERVA

71. En una competición de Kaggle ha participado en un equipo con varias personas, cada una de ellas creando un modelo diferente para afrontar el problema. Tras realizar varios experimentos, llegan a la conclusión de que los mejores resultados en validación se alcanzan creando un modelo de tipo XGB que se alimente de las predicciones de cada uno de los modelos individuales. ¿Qué técnica describe mejor el proceso empleado aquí?

a) Meta-learning.

b) Boosting.

c) Mezcla de expertos.

d) Stacking.

72. Suponga que cuenta con datos en el siguiente formato:

```
X = np.array([
    [1, 2],
    [2, 0],
    [3, 1],
    [4, 3],
    [5, 5]
])
y = np.array([2, 1, 2, 3, 5])
lambda = 10
```

¿Qué modelo de machine learning se ajusta más a la implementación de la siguiente línea de código?

```
beta = np.linalg.inv(X.T @ X + lambda * np.eye(2)) @ X.T @ y
```

a) Ridge Regression.

b) Lasso.

c) Regresión por mínimos cuadrados.

d) Perceptrón.

73. ¿Cuál de las siguientes fórmulas representa correctamente el cálculo del TF-IDF (Term Frequency – Inverse Document Frecuency), para un término t en un documento d el cual forma parte de un conjunto de documentos D? Considere que TF(t, d) es la frecuencia del término t dentro del documento d, DF(t) el número de documentos en los que aparece el término t y N el número total de documentos. Considérese un caso clásico donde 0< DF(t)<N.

a)

$$TF(t,d) + \log\left(\frac{N}{DF(t)}\right)$$

b)

$$TF(t,d) \cdot \log\left(\frac{N}{DF(t)}\right)$$

c)

$$TF(t,d) \cdot \log\left(1 + \frac{N}{DF(t)}\right)$$

d)

$$\frac{TF(t,d)}{N} \cdot DF(t)$$

74. En un centro de educación secundaria se está planteando un proyecto para evitar problemas de acoso a estudiantes. Para ello, se va a instalar un sistema de cámaras en las zonas comunes, que mediante técnicas de reconocimiento de imagen analice las emociones de los estudiantes para detectar situaciones problemáticas. En términos legales, ¿qué debe tenerse en cuenta a la hora de desplegar este sistema?

a) La legislación prohíbe desplegar un sistema de estas características.

b) Si no se realiza reconocimiento facial con el fin de identificar a los estudiantes, sino que únicamente se detectan emociones sin asociar la identidad de la persona, el sistema cumpliría con la RGPD.

c) Es necesario el consentimiento de los padres o tutores legales de los estudiantes previo al despliegue del sistema.

d) Dado que los estudiantes son menores y el centro es responsable de los mismos mientras estén en horario lectivo, el centro está cubierto legalmente para desplegar este tipo de sistema.

75. Un casino online emplea una moneda trucada para las tiradas de azar, con el fin de conseguir resultados que parezcan más variables para el visitante. Para ello, la primera tirada de la moneda sigue una distribución uniforme, pero tiradas posteriores sesgan las probabilidades para que el resultado de la tirada anterior solo se repita con una probabilidad del 25%. Teniendo en cuenta esto, ¿cuál es la probabilidad de que al realizar 3 tiradas se obtenga cruz tanto en la primera como en la última tirada?

a) 1/4

b) 3/4

c) 3/16

d) 5/16

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS

CONTESTE A LAS PREGUNTAS EN LA HOJA DE RESPUESTAS