

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

Perfil Ciencia de Datos

PREGUNTAS COMUNES PARA AMBOS PERFILES

1. Considera 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.

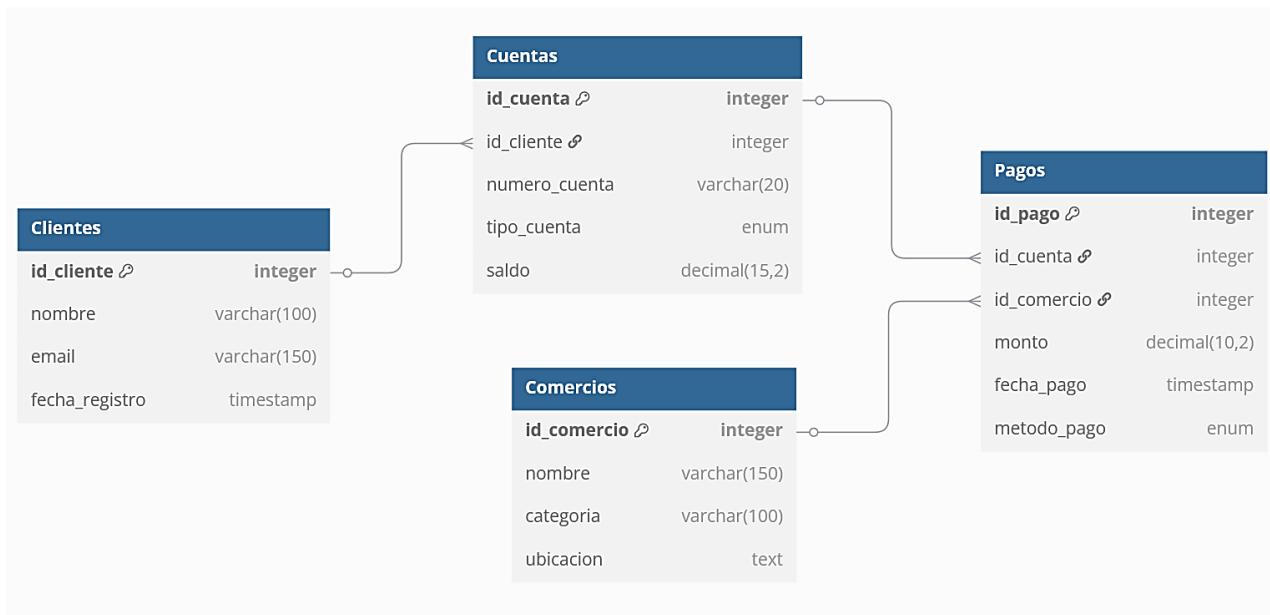
- 3. Se plantea un problema de clasificación binaria en el que la relación de clases sigue una proporción de 99% para la clase mayoritaria y 1% para la clase minoritaria. A la hora de diseñar un estudio de diferentes modelos supervisados de clasificación y teniendo en cuenta que ambas clases son importantes 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.

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.

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]
```

Indique 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()
```

- 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 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())
```

Considera 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 Kingdom",  
      "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)
```

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:

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

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 $\left\| w \right\|_2^2$ 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)

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

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

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

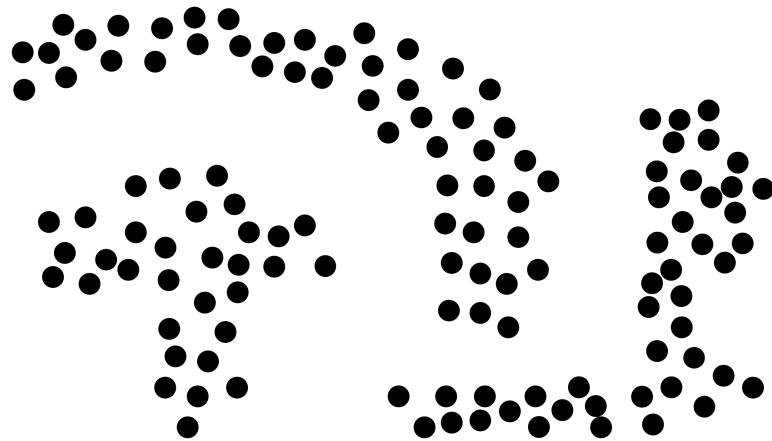
d)

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

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

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

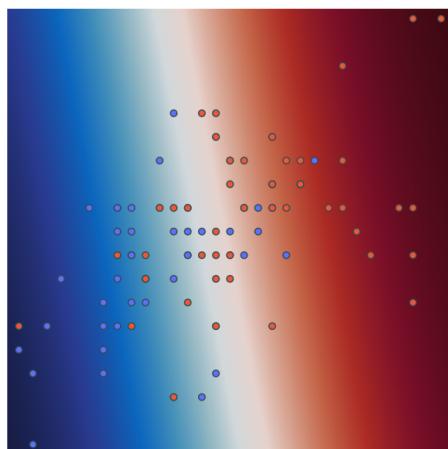
16. Está trabajando en un análisis no supervisado de datos, para el que espera encontrar una serie de clusters bien diferenciados. Observa que al realizar proyecciones de baja dimensión de los datos aparecen estructuras como las siguientes:



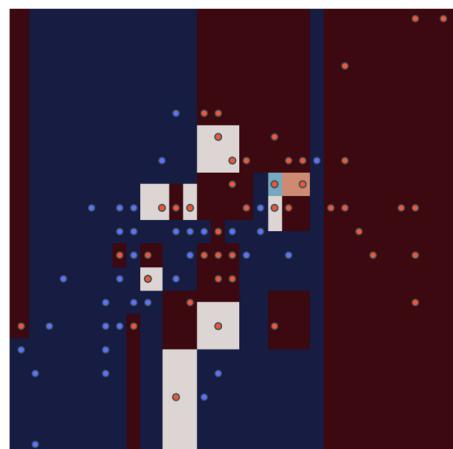
Teniendo en cuenta esta información, ¿qué algoritmo de clustering sería más adecuado?

- a) K-means.
- b) K-means modificado con distancias de Mahalanobis.
- c) Modelo de mezcla de gaussianas.
- d) DBSCAN.

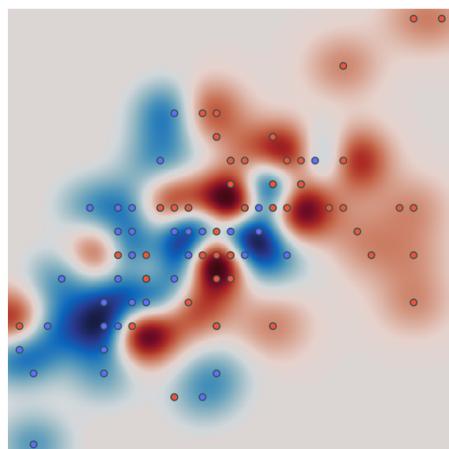
17. Se han entrenado 3 modelos de clasificación supervisada sobre el mismo conjunto de datos, obteniendo las siguientes visualizaciones de sus fronteras de decisión:



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”.

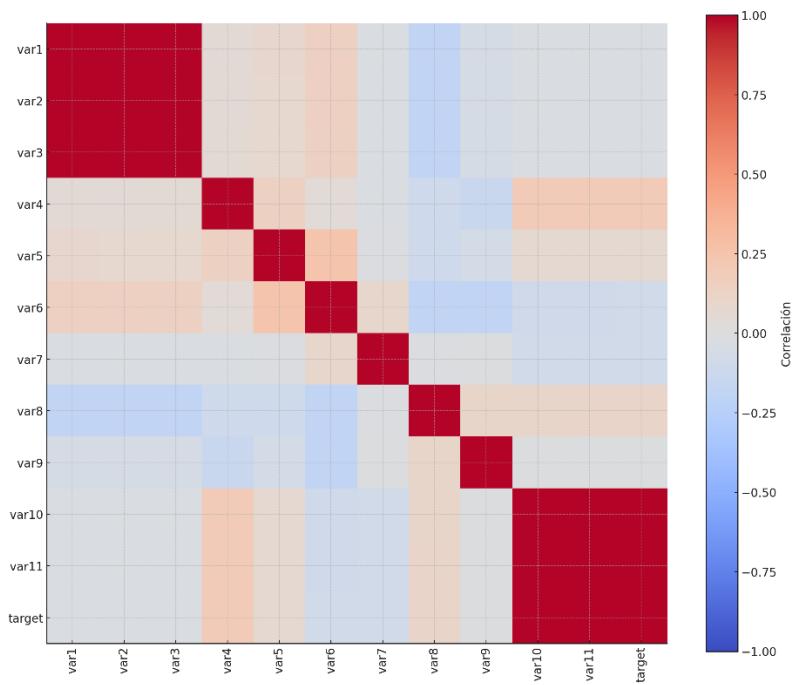
19. Tras entrenar un modelo de regresión lineal sobre un gran conjunto de datos de entrenamiento, se realiza un análisis de errores en un conjunto de test que se ha obtenido un año después en el proceso de recogida de datos, y se observa que el modelo tiene baja varianza pero alto sesgo, a pesar de que 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` → $(\sin(\text{mes} \cdot 2\pi/12), \cos(\text{mes} \cdot 2\pi/12))$, `dia_del_mes` → $(\sin(\text{dia_del_mes} \cdot 2\pi/31), \cos(\text{dia_del_mes} \cdot 2\pi/31))$, `hora_del_dia` → $(\sin(\text{hora_del_dia} \cdot 2\pi/24), \cos(\text{hora_del_dia} \cdot 2\pi/24))$
- b) `mes` → $(\sin(\text{mes} \cdot 12\pi/2), \cos(\text{mes} \cdot 12\pi/2))$, `dia_del_mes` → $(\sin(\text{dia_del_mes} \cdot 31\pi/2), \cos(\text{dia_del_mes} \cdot 31\pi/2))$, `hora_del_dia` → $(\sin(\text{hora_del_dia} \cdot 24\pi/2), \cos(\text{hora_del_dia} \cdot 24\pi/2))$
- c) $(\text{mes}, \text{dia_del_mes}, \text{hora_del_dia}) \rightarrow \text{mes} \cdot 31 \cdot 24 + \text{dia_del_mes} \cdot 24 + \text{hora_del_dia}$.
- d) $(\text{mes}, \text{dia_del_mes}, \text{hora_del_dia}) \rightarrow \text{mes} + \text{dia_del_mes} \cdot 12 + \text{hora_del_dia} \cdot 31$.

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)
}
```

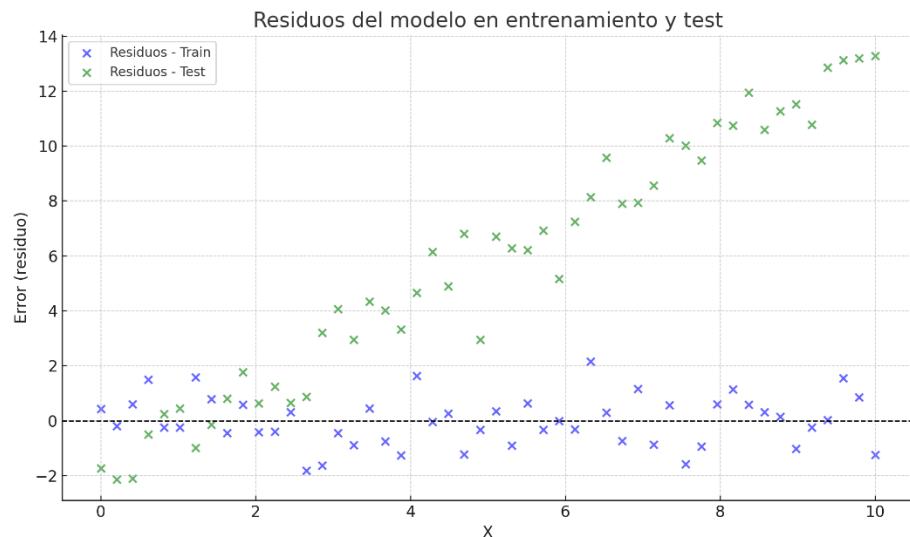
d)

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

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).

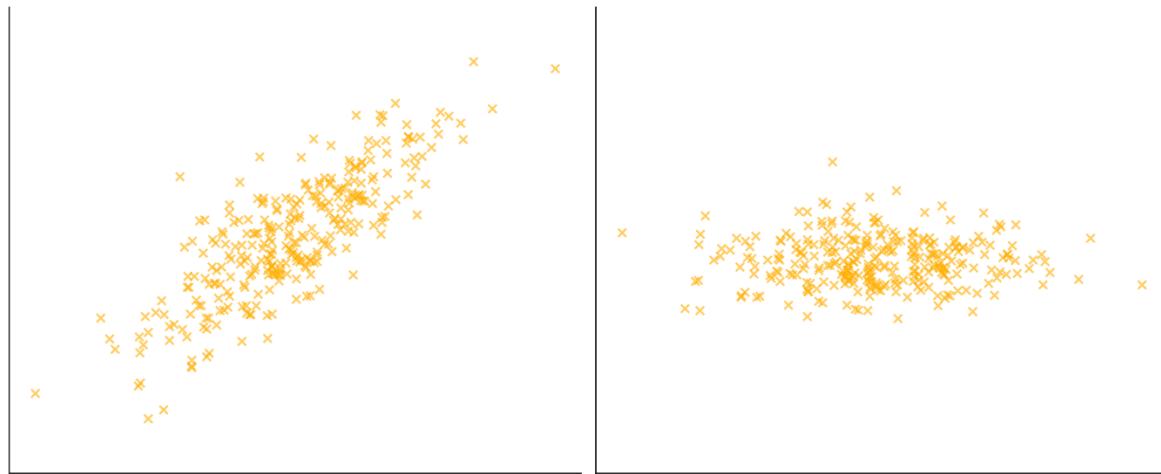
26. Tras entrenar un modelo de regresión lineal sobre un gran conjunto de datos de entrenamiento se ha realizado el siguiente análisis de sus errores, tanto sobre el conjunto de entrenamiento original como sobre un conjunto de test recogido un año después.



¿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.

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).

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]
```

Considerese 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.

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)

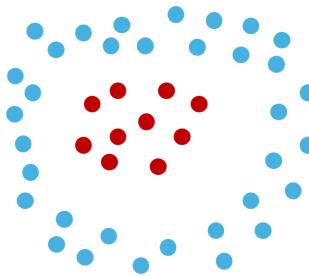
```
CREATE TABLE Medicos (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE NOT NULL,
    Especialidad VARCHAR(100) NOT NULL
);
CREATE TABLE Pacientes (
    ID INT AUTO_INCREMENT PRIMARY KEY,
    Nombre VARCHAR(100) NOT NULL,
    DNI VARCHAR(20) UNIQUE 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 CIENCIA DE DATOS

31. 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

32. 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.

33. 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

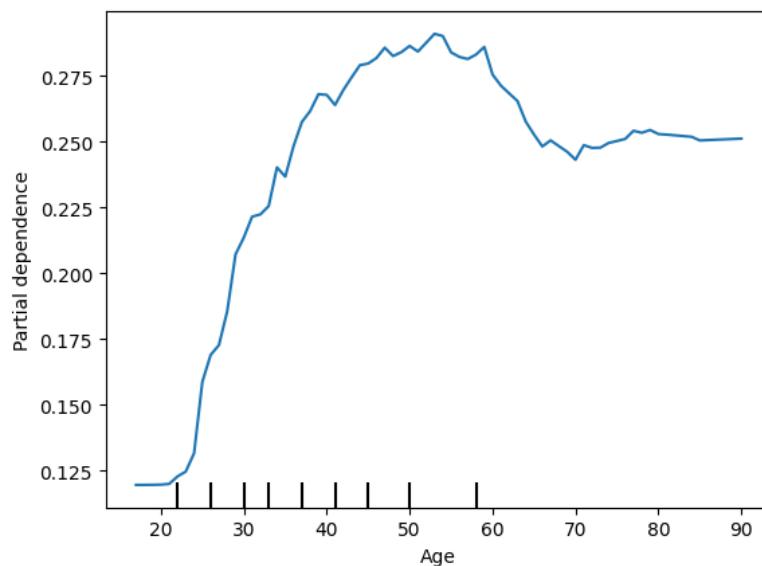
34. El algoritmo Proximal Policy Optimization (PPO) es uno de los métodos más utilizados en sistemas de aprendizaje por refuerzo. Suponga la siguiente notación relacionada con este algoritmo: π_θ la función de política que está siendo entrenada, $\pi_{\theta_{old}}$ una versión anterior de esa política, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ 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)
$$\min_{\theta} \mathbb{E}_t \left[\min(r_t(\theta)A_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)) - c_1 (V_\theta(s_t) - V_{target}(s_t))^2 + c_2 entropy[\pi_\theta](s_t) \right]$$
- b)
$$\max_{\theta} \mathbb{E}_t \left[\min(r_t(\theta)A_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t) - c_1 (V_\theta(s_t) - V_{target}(s_t))^2 + c_2 entropy[\pi_\theta](s_t) \right]$$
- c)
$$\min_{\theta} \mathbb{E}_t \left[\max(r_t(\theta)A_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t) - c_1 (V_\theta(s_t) - V_{target}(s_t))^2 + c_2 entropy[\pi_\theta](s_t) \right]$$
- d)
$$\max_{\theta} \mathbb{E}_t \left[\min(r_t(\theta)A_t, clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t) - c_1 (V_\theta(s_t) - V_{target}(s_t))^2 - c_2 entropy[\pi_{ref}](s_t) \right]$$

35. 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.

36. Se ha entrenado un modelo usando Extreme Gradient 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.

37. 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.

38. 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.

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

a)

$$\begin{aligned}
 f_t &= \tanh(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned}$$

b)

$$\begin{aligned}
 f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned}$$

c)

$$\begin{aligned}
 f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
 i_t &= \tanh(W_i \cdot [h_{t-1}, x_t] + b_i) \\
 o_t &= \tanh(W_o \cdot [h_{t-1}, x_t] + b_o) \\
 c_t &= f_t + c_{t-1} + i_t + \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
 h_t &= \sigma(c_t)
 \end{aligned}$$

d)

$$\begin{aligned}
 f_t &= \sigma(W_f \cdot h_{t-1} + x_t + b_f) \\
 i_t &= \sigma(W_i \cdot h_{t-1} + x_t + b_i) \\
 o_t &= \sigma(W_o \cdot h_{t-1} + x_t + b_o) \\
 c_t &= (f_t + i_t) \odot \tanh(W_c \cdot h_{t-1} + x_t + b_c) \\
 h_t &= o_t + c_t
 \end{aligned}$$

40. ¿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.

41. 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 ϵ óptimo obtenido.
- d) Proceso gaussiano.

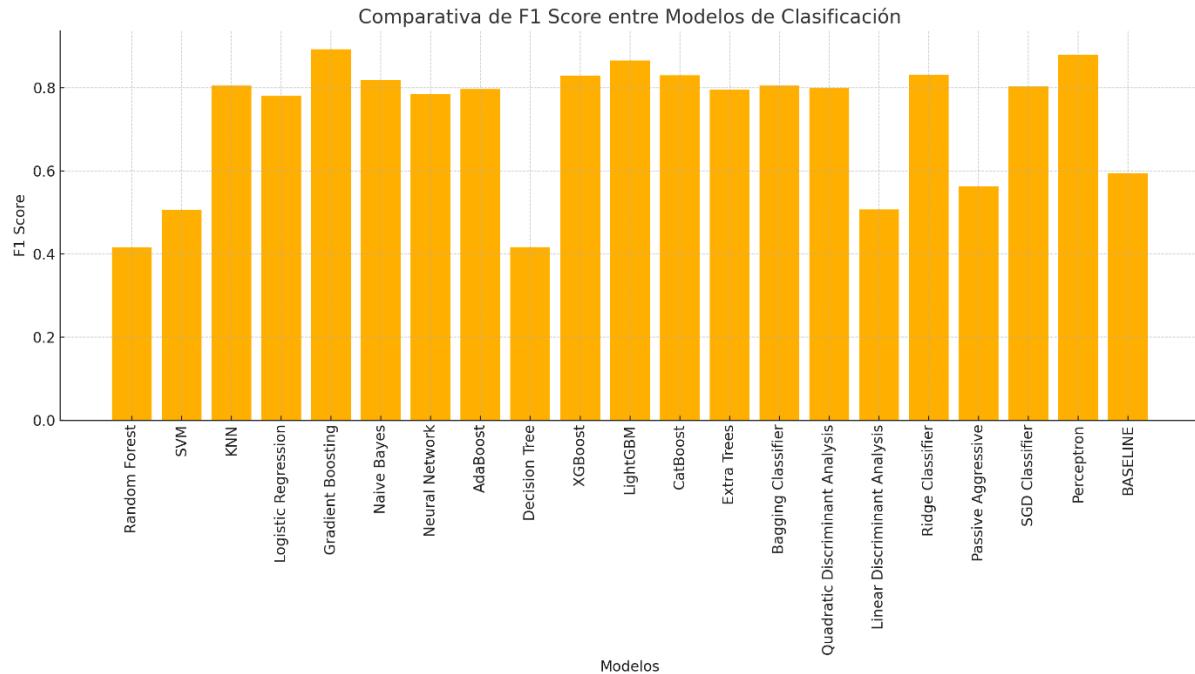
42. Según la teoría de los ensambles de modelos, ¿cuál de las siguientes configuraciones de ensamblado ofrecería una ventaja sobre emplear únicamente uno de los modelos base? Considere que cada modelo del ensemble se entrena sobre ligeras variaciones del mismo conjunto de entrenamiento.

- 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.

43. 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.

44. 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?



- Ordenar las columnas por F1 Score, y reemplazar la columna BASELINE por una línea horizontal.
- Colorear las columnas según el valor de F1 conseguido, siguiendo una escala perceptualmente uniforme.
- Eliminar aquellas columnas que tengan un F1 Score inferior al baseline, ya que son modelos que no tiene sentido desplegar en la práctica.
- Ordenar los modelos alfabéticamente, y hacer que el origen del eje Y coincida con el modelo de menor rendimiento.

45. 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.

46. 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.

47. 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
- 48. 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.

49. 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

50. 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

51. 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 alimenta 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.

52. 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.

53. ¿Cuál de las siguientes fórmulas representa correctamente el cálculo del TF-IDF (Term Frequency – Inverse Document Frequency), 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)

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

54. 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.

55. 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