Contenido

[General 1](#_Toc202787439)

[Machine Learning 1](#_Toc202787440)

[Deep Learning 2](#_Toc202787441)

[Supervisado 2](#_Toc202787442)

[Regresión 2](#_Toc202787443)

[Clasificación 2](#_Toc202787444)

[No pupervisado 5](#_Toc202787445)

[Reducción de la dimensión con análisis de componente principales 5](#_Toc202787446)

[K medias 5](#_Toc202787447)

[Deep learning 6](#_Toc202787448)

[RNA 6](#_Toc202787449)

[RNC 7](#_Toc202787450)

[RNR 7](#_Toc202787451)

# General

## **Machine Learning**

1. **¿Cuál es la diferencia entre aprendizaje supervisado y no supervisado? Da ejemplos de algoritmos para cada uno.**

**El aprendizaje supervisado es aquel tipo de aprendizaje en la cual tenemos las instancias de salida conocidas y por lo tanto podemos medir directamente el rendimiento con diferentes métricas. Es decir, dividimos nuestro conjunto total de datos en dos subconjuntos y uno de ellos se utiliza para entrenar al modelo y el restante se utiliza para evaluarlo. Usualmente la proporción de los datos de entrenamiento es mayor respecto a la de prueba**

El aprendizaje no supervisado es cuando no tenemos las instancias de salida y por lo tanto el objetivo de este tipo de aprendizaje es encontrar estructuras intrínsecas en los datos para extraer información.

Ejemplos de aprendizaje supervisado es la regresión y clasificación. Ejemplos de aprendizaje no supervisado es la segmentación

1. **¿Qué es el overfitting y cómo se puede prevenir?**

**Intuitivamente sobre ajuste es cuando el modelo “memoriza” los datos, pero no encuentra patrones dentro de la estructura del modelo. Una de las causas del sobreajuste es tener una gran cantidad de características y, por lo tanto y en general, de parámetros. Una de las posibles soluciones es hacer selección de características ya sea manualmente o con algún tipo de regularización. Otra alternativa es utilizar un modelo más robusto para los datos correspondientes.**

**El caso contrario al sobre ajuste, es el subajuste. El su ajuste se refiere a que el modelo no es capaz de ajustarse o generalizar a los datos y por lo tanto tiene un error muy grande. Una de las posibles causas es la cantidad pequeña de datos. Una de las posibles soluciones es hacer ingeniería de características, o bien, aumentar la complejidad del modelo**

1. **Explica el sesgo y la varianza en modelos de ML. ¿Qué es el trade-off entre ambos?**

**El objetivo es conseguir un sesgo y varianza bajos para obtener buenos modelos. Aquí el sesgo se refiere al promedio del error que comete un modelo. Y la varianza significa cuánto varía Un modelo respecto a otro entrenado un subconjunto diferente**

**El concepto de Trade of se refiere a la compensación entre una métrica y otra. Por ejemplo la compensación entre precisión y sensibilidad en modelos de clasificación**

1. **¿Qué métricas usarías para evaluar un modelo de clasificación? ¿Y uno de regresión?**

**Para evaluar un modelo de clasificación utilizaría la matriz de confusión y elegir las métricas adecuadas según el objetivo del problema. Ya que no es lo mismo escoger una métrica para un problema médico que para un problema financiero. La forma que usaría para optimizar un clasificador depende de los parámetros e hiperparametros del modelo, pero por mencionar algunos; para regresión logística utilizaría el umbral de decisión con la curva roc o el gradiente descendente**

**Para un modelo de regresión utilizaría métricas como el error cuadrático medio o la raíz del error cuadrático medio. La forma que usaría para optimizar el regreso depende del modelo particular, Pero por mencionar algunos utilizaría el gradiente descendente, la eliminación hacia atrás a partir de los p valores (en regresión lineal múltiple).**

**En ambos casos podría utilizar un método de validación cruzada para hacer más robusta la evaluación**

1. **¿Qué es la validación cruzada (cross-validation) y por qué es importante?**

La validación cruzada es un método de evaluación que consiste en:

* Tenemos nuestro conjunto de datos
* Ese conjunto de datos lo dividimos en k folds
* Entrenamos cada fold con el modelo con una proporción definida, pero variando los subconjuntos de prueba y entrenamiento en cada fold
* Evaluamos con una misma métrica preferentemente para todos los resultados de los k fold
* Hacemos el promedio de las métricas y ese es el resultado del rendimiento de nuestro modelo

Esta técnica es importante para la evaluación de modelos ya que permite robustez frente a sesgos. También es importante porque no sólo se utiliza para la evaluación de modelo sino también para la búsqueda de hiperparámetros óptimos

1. **¿Cuál es la diferencia entre regresión lineal y regresión logística?**

**La diferencia entre estos dos modelos de aprendizaje supervisado es que la regresión lineal predice un valor continuo a partir de un conjunto de características. Y la regresión logística a partir de un mismo conjunto de características predice la probabilidad de que la observación pertenezca a una categoría u otra (ya sea binaria o multiclase). La diferencia matemática es que la regresión lineal utiliza los mínimos cuadrados ordinarios y la logística la máxima verosimilitud.**

**La regresión logística utiliza la función sigmoide y umbrales de decisión y la regresión lineal no. Ambos puedes optimizarse con el gradiente descendente.**

1. **¿Qué es la regularización? Explica L1 (Lasso) y L2 (Ridge).**

**La regularización en realidad son modelos modificados para reducir la complejidad del modelo:**

* **En el caso de la regresion lasso se eliminan completamente las características menos relevantes (con amyor multicolinealidad)**
* **En el cado de ridge se reducen considerablemente los coeficientes de las características menos relevantes, pero sin llegar a ser cero**

**En términos generales, las regularizaciones hacen la selección de características por nosotros.**

1. **¿Cómo funciona un árbol de decisión? ¿Qué criterios se usan para dividir los nodos?**
2. **¿Qué es un Random Forest y cómo mejora sobre un solo árbol de decisión?**
3. **¿Qué es el método del k-vecinos más cercanos (KNN)? ¿Cuáles son sus ventajas y desventajas?**
4. **Explica cómo funciona el algoritmo de Support Vector Machines (SVM). ¿Qué es el margen?**
5. **¿Qué es el feature engineering y por qué es importante?**
6. **¿Cómo se puede manejar el desbalance de clases en un conjunto de datos?**
7. **¿Qué es la matriz de confusión y qué información nos da?**
8. **¿Qué es un pipeline de Machine Learning y por qué es útil?**
9. **¿Qué es la selección de características (feature selection) y cómo se puede hacer?**
10. **¿Qué es el clustering? Menciona y compara K-Means y DBSCAN.**
11. **¿Qué significa que un modelo sea interpretable? ¿Qué técnicas conoces para interpretar modelos?**
12. **¿Qué son los learning curves y cómo ayudan a diagnosticar problemas del modelo?**
13. **¿Cuál es la diferencia entre Bagging y Boosting? ¿Conoces algoritmos de Boosting?**

## **Deep Learning**

1. **¿Qué es una red neuronal artificial y cómo se estructura una red feedforward básica?**
2. **¿Qué función cumple la función de activación en una red neuronal? Menciona algunas comunes.**
3. **¿Qué es la retropropagación (backpropagation) y cómo funciona el proceso de entrenamiento de una red?**
4. **¿Cuál es la diferencia entre una red neuronal tradicional y una red neuronal convolucional (CNN)?**
5. **¿Qué es una red recurrente (RNN) y en qué tipo de tareas se utiliza comúnmente?**
6. **¿Qué problemas enfrentan las RNN y cómo los resuelven las LSTM o GRU?**
7. **¿Qué es el dropout y cómo ayuda a evitar el overfitting en redes neuronales?**
8. **¿Qué es la normalización por lotes (batch normalization) y qué ventajas ofrece?**
9. **¿Qué papel juegan los optimizers como SGD, Adam, RMSProp en el entrenamiento de redes neuronales?**
10. **¿Qué son los embeddings y en qué contexto se utilizan en Deep Learning?**

# Supervisado

## Regresión

Regresión líneal

📌 **Preguntas teóricas sobre Regresión Lineal**

1. **¿Qué es la regresión lineal y cuál es su objetivo principal?**
2. **¿Cuál es la diferencia entre una regresión lineal simple y una múltiple?**
3. **¿Qué supuestos fundamentales deben cumplirse para que una regresión lineal sea válida?**
4. **¿Qué representa el coeficiente de determinación R2R^2? ¿Qué significa un valor cercano a 1?**
5. **¿Qué es el p-valor asociado a un coeficiente en regresión lineal? ¿Cómo lo interpretas?**
6. **¿Qué problemas pueden surgir si hay multicolinealidad entre las variables independientes? ¿Cómo se puede detectar?**
7. **¿Qué es el error cuadrático medio (MSE) y cómo se relaciona con la calidad del modelo?**
8. **¿Qué son los residuos (residuals) en regresión y por qué es importante analizarlos?**
9. **¿Cuál es la diferencia entre underfitting y overfitting en el contexto de regresión lineal?**
10. **¿Qué métodos puedes usar para seleccionar las variables más relevantes en un modelo de regresión múltiple?**

## Clasificación

Regresión logística

📌 **Preguntas teóricas sobre Regresión Logística**

1. **¿Qué es la regresión logística y en qué casos se utiliza?**
2. **¿Cuál es la principal diferencia entre regresión lineal y regresión logística?**
3. **¿Qué tipo de función se utiliza en la regresión logística para mapear la salida? ¿Cuál es su fórmula?**
4. **¿Qué representa el output de una regresión logística antes y después de aplicar el umbral de clasificación?**
5. **¿Qué es la función de pérdida usada en regresión logística y por qué no se usa el error cuadrático medio?**
6. **¿Qué es el log-odds y cómo se interpreta en el contexto de regresión logística?**
7. **¿Cómo se interpretan los coeficientes de una regresión logística?**
8. **¿Qué técnicas puedes usar para manejar clases desbalanceadas en regresión logística?**
9. **¿Cómo evalúas el desempeño de un modelo de regresión logística? Menciona al menos 3 métricas comunes.**
10. **¿Qué supuestos tiene la regresión logística y cuáles son sus limitaciones?**

Arboles de decisión

🌳 **Preguntas teóricas sobre Árboles de Decisión**

1. **¿Qué es un árbol de decisión y para qué tipo de problemas se puede usar?**
2. **¿Cómo decide un árbol de decisión en qué variable dividir en cada nodo?**
3. **¿Qué es la impureza de un nodo y cuáles son las medidas más comunes para calcularla?**
4. **¿Cuál es la diferencia entre Gini y Entropía como criterios de división?**
5. **¿Qué significa que un árbol esté sobreajustado (overfitting)? ¿Cómo se puede evitar?**
6. **¿Qué son la profundidad del árbol (depth), el número mínimo de muestras por nodo, y el número mínimo de muestras para dividir? ¿Cómo afectan al modelo?**
7. **¿Qué es el pruning (poda) y en qué se diferencia el pre-pruning del post-pruning?**
8. **¿Cómo se interpreta un árbol de decisión entrenado? ¿Qué representa cada nodo y cada hoja?**
9. **¿Qué ventajas y desventajas tienen los árboles de decisión en comparación con modelos lineales?**
10. **¿Qué problemas pueden surgir si hay muchas variables categóricas o numéricas altamente correlacionadas?**

Bosques aleatorios

🌲 **Preguntas teóricas sobre Bosques Aleatorios**

1. **¿Qué es un bosque aleatorio y cómo se construye?**
2. **¿Cómo se diferencia un bosque aleatorio de un solo árbol de decisión?**
3. **¿Qué es el bagging (Bootstrap Aggregating) y cómo se utiliza en un bosque aleatorio?**
4. **¿Qué significa el parámetro "n\_estimators" en un modelo de bosque aleatorio y cómo afecta al rendimiento?**
5. **¿Qué es la selección aleatoria de características en cada división de un árbol y por qué se usa en bosques aleatorios?**
6. **¿Cómo se combina la salida de los múltiples árboles en un bosque aleatorio para hacer una predicción final?**
7. **¿Cuáles son las ventajas de usar un bosque aleatorio frente a un solo árbol de decisión?**
8. **¿Cómo evalúas el desempeño de un modelo de bosque aleatorio? Menciona al menos 2 métricas comunes.**
9. **¿Qué es la importancia de las características (feature importance) en un bosque aleatorio y cómo se calcula?**
10. **¿Cuáles son los principales problemas de los bosques aleatorios y cómo se pueden mitigar (por ejemplo, sobreajuste o complejidad del modelo)?**

K vecinos más cercanos

👥 **Preguntas teóricas sobre k-Vecinos Más Cercanos (k-NN)**

1. **¿Qué es el algoritmo k-vecinos más cercanos (k-NN) y en qué tipo de problemas se utiliza comúnmente?**
2. **¿Cómo decide el modelo k-NN la clase o el valor de un punto de datos?**
3. **¿Qué papel juega el parámetro 'k' en k-NN y cómo afecta al rendimiento del modelo?**
4. **¿Qué significa que el valor de k sea demasiado pequeño o demasiado grande? ¿Cómo afecta a la precisión del modelo?**
5. **¿Qué distancia se utiliza habitualmente para medir la cercanía entre los puntos en k-NN? Menciona al menos dos métricas de distancia.**
6. **¿Cómo se maneja la clasificación cuando hay un empate entre los vecinos más cercanos (por ejemplo, si la mitad son de una clase y la otra mitad de otra)?**
7. **¿Qué es el problema de la maldición de la dimensionalidad y cómo afecta a k-NN?**
8. **¿Cuáles son las principales ventajas y desventajas del algoritmo k-NN?**
9. **¿Cómo se puede optimizar el rendimiento de k-NN cuando se tiene un conjunto de datos con muchas características?**
10. **¿Qué técnicas puedes usar para mejorar la eficiencia computacional del modelo k-NN en conjuntos de datos grandes?**

Máquinas de vectores de soporte

🚀 **Preguntas teóricas sobre Máquinas de Vectores de Soporte (SVM)**

1. **¿Qué es una Máquina de Vectores de Soporte (SVM) y para qué tipo de problemas se utiliza comúnmente?**
2. **¿Cómo funciona el algoritmo SVM para encontrar el hiperplano óptimo que separa las clases?**
3. **¿Qué es el margen en SVM y por qué es importante maximizarlo?**
4. **¿Qué son los vectores de soporte en un modelo SVM y cuál es su función?**
5. **¿Cómo se manejan los casos en los que los datos no son linealmente separables en SVM?**
6. **¿Qué es el truco del kernel (kernel trick) en SVM y por qué se utiliza?**
7. **Explica las diferencias entre los principales tipos de kernels en SVM: lineal, polinómico y RBF.**
8. **¿Qué es el parámetro C en SVM y cómo afecta al modelo?**
9. **¿Qué es el overfitting en SVM y cómo se controla el balance entre el margen máximo y los errores de clasificación?**
10. **¿Qué métricas utilizarías para evaluar un modelo SVM en un problema de clasificación?**

Naive bayes

🧠 **Preguntas teóricas sobre Naive Bayes**

1. **¿Qué es el clasificador Naive Bayes y en qué tipo de problemas se utiliza comúnmente?**
2. **¿En qué se basa el clasificador Naive Bayes para realizar las predicciones?**
3. **¿Qué significa que Naive Bayes sea "naive" (ingenuo) y cómo impacta esto en su funcionamiento?**
4. **Explica el teorema de Bayes y cómo se aplica en Naive Bayes.**
5. **¿Cómo maneja Naive Bayes la dependencia entre características?**
6. **¿Qué tipo de distribuciones se asumen para las variables en Naive Bayes (en especial para las características continuas)?**
7. **¿Cuáles son las principales ventajas y desventajas de usar un clasificador Naive Bayes?**
8. **¿Qué es la suavización de Laplace y por qué es importante en Naive Bayes?**
9. **¿Cómo se calcula la probabilidad posterior en Naive Bayes para una clase dada?**
10. **¿Cómo evaluarías el rendimiento de un clasificador Naive Bayes en un problema de clasificación? Menciona al menos 2 métricas comunes.**

# No pupervisado

## Reducción de la dimensión con análisis de componente principales

📉 **Preguntas teóricas sobre Reducción de Dimensionalidad**

1. **¿Qué es la reducción de dimensionalidad y por qué es útil en el análisis de datos?**
2. **¿Cuáles son las principales técnicas de reducción de dimensionalidad y en qué casos se utilizan cada una?**
3. **Explica qué es el Análisis de Componentes Principales (PCA) y cómo se utiliza para reducir la dimensionalidad de un conjunto de datos.**
4. **¿Qué es la varianza explicada en PCA y cómo se determina cuántos componentes principales mantener?**
5. **¿Cuál es la diferencia entre PCA y el Análisis de Discriminante Lineal (LDA)?**
6. **¿Cómo puede la reducción de dimensionalidad ayudar a mejorar el rendimiento de los modelos de machine learning?**
7. **¿Qué son los "eigenvectores" y "eigenvalores" en el contexto de PCA y qué rol juegan en la reducción de dimensionalidad?**
8. **¿Qué es la "maldición de la dimensionalidad" y cómo se relaciona con la reducción de dimensionalidad?**
9. **¿Cuáles son las principales limitaciones de la reducción de dimensionalidad y cómo se pueden mitigar?**
10. **Explica el concepto de t-SNE y en qué situaciones es más útil que PCA.**

## K medias

🔵 **Preguntas teóricas sobre k-medias (k-means)**

1. **¿Qué es el algoritmo k-medias y en qué tipo de problemas se utiliza comúnmente?**
2. **¿Cómo funciona el algoritmo k-medias para agrupar los datos? Explica brevemente el proceso.**
3. **¿Qué significa el parámetro 'k' en k-medias y cómo se determina el número óptimo de clústeres?**
4. **¿Cómo se inicializan los centroides en el algoritmo k-medias y por qué es importante esta inicialización?**
5. **¿Qué pasa si se elige un valor incorrecto de 'k'? ¿Cómo afecta al rendimiento del modelo?**
6. **¿Cómo se mide la calidad de un clúster en k-medias? Menciona al menos una métrica común.**
7. **¿Qué es el "inercia" o "suma de los errores cuadráticos" en k-medias y cómo se usa para evaluar el modelo?**
8. **¿Qué significa que el algoritmo k-medias pueda quedar atrapado en un mínimo local? ¿Cómo se puede mitigar este problema?**
9. **¿Cuáles son las principales ventajas y desventajas del algoritmo k-medias en comparación con otros métodos de clustering?**
10. **¿Qué son los algoritmos de inicialización, como k-means++ y cómo mejoran el rendimiento de k-medias?**

# Deep learning

## RNA

🧠 **Preguntas teóricas sobre Redes Neuronales Artificiales (RNA)**

1. **¿Qué es una red neuronal artificial (RNA) y cuáles son sus componentes básicos?**
2. **Explica cómo funciona una red neuronal feedforward (avance hacia adelante). ¿Qué ocurre en cada capa?**
3. **¿Qué son las funciones de activación en una red neuronal y por qué son importantes? Menciona al menos dos funciones de activación comunes.**
4. **¿Qué es el proceso de entrenamiento en una red neuronal y cómo se ajustan los pesos durante este proceso?**
5. **¿Qué es la retropropagación (backpropagation) y cómo se usa para entrenar redes neuronales?**
6. **¿Qué es el sobreajuste (overfitting) en redes neuronales y qué técnicas puedes utilizar para evitarlo?**
7. **¿Cuál es la diferencia entre una red neuronal profunda (deep learning) y una red neuronal superficial (shallow learning)?**
8. **¿Qué es el gradiente descendente (gradient descent) y cómo se utiliza en el entrenamiento de redes neuronales?**
9. **¿Qué significa el término "epoch" en el contexto del entrenamiento de redes neuronales? ¿Qué factores influyen en el número de epochs?**
10. **¿Cómo se evalúa el rendimiento de una red neuronal? Menciona algunas métricas y cómo se aplican.**

## RNC

🖼️ **Preguntas teóricas sobre Redes Neuronales Convolucionales (RNC)**

1. **¿Qué es una Red Neuronal Convolucional (CNN) y cómo se diferencia de una red neuronal tradicional (feedforward)?**
2. **¿Cuál es el propósito de las capas convolucionales en una RNC y cómo funcionan?**
3. **Explica qué es un filtro (o kernel) en una CNN y cómo se utiliza durante la convolución.**
4. **¿Qué es el "padding" en una CNN y por qué se utiliza?**
5. **¿Qué es el "pooling" en una red neuronal convolucional? Explica los tipos más comunes de pooling (max pooling y average pooling).**
6. **¿Qué es la función de activación más comúnmente usada en una CNN y por qué se prefiere esta función?**
7. **¿Por qué es importante el proceso de "flattening" en una CNN? ¿Qué sucede después de este paso?**
8. **¿Qué ventajas tienen las CNNs en comparación con las redes neuronales tradicionales, especialmente en el procesamiento de imágenes?**
9. **¿Qué es el sobreajuste (overfitting) en redes neuronales convolucionales y cómo se puede mitigar en el entrenamiento de una CNN?**
10. **¿Cómo evalúas el rendimiento de una CNN en un problema de clasificación de imágenes? Menciona al menos dos métricas que usarías.**

## RNR

🔄 **Preguntas teóricas sobre Redes Neuronales Recurrentes (RNR)**

1. **¿Qué es una Red Neuronal Recurrente (RNR) y en qué se diferencia de una red neuronal feedforward tradicional?**
2. **¿Cómo funciona el mecanismo de retroalimentación en una RNR y por qué es importante para procesar secuencias de datos?**
3. **¿Qué es el problema de desvanecimiento y explosión del gradiente en RNRs? ¿Cómo afecta al entrenamiento y cómo se puede mitigar?**
4. **¿Qué son las celdas LSTM (Long Short-Term Memory) y cómo abordan el problema del desvanecimiento del gradiente en las RNRs?**
5. **¿Qué son las celdas GRU (Gated Recurrent Units) y en qué se diferencian de las LSTM?**
6. **¿En qué casos se utilizan las redes neuronales recurrentes en comparación con las redes neuronales convolucionales (CNN)?**
7. **¿Qué es un "estado oculto" en una RNR y cómo se utiliza para transmitir la información a través del tiempo?**
8. **Explica el concepto de "bidiagonalidad" en RNRs y cómo se implementa en redes como Bidirectional RNNs. ¿En qué situaciones es útil?**
9. **¿Qué es el "backpropagation through time" (BPTT) y cómo se utiliza para entrenar redes neuronales recurrentes?**
10. **¿Cómo evaluarías el rendimiento de una RNR en un problema de predicción de secuencias, como la predicción de series de tiempo? Menciona al menos 2 métricas que usarías.**