Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

19 de Febrero, 2024

1. Explique brevemente 5 características de un dataset que analizaría a la hora de realizar un análisis exploratorio.
2. Explique las métricas MAE y RMSE utilizando un diagrama. En que escenarios debe utilizarse cada una? Por qué?
3. Clasifique las gráficas de la figura 1 en las categorías Just Right, Underfitting y Overfitting. Explique las distintas categorías, por qué suceden estos fenómenos y cómo pueden remediarse.
4. Una compañía de marketing le solicita que cree un gráfico impactante sobre los riesgos de fumar. Usted sabe que estadísticamente el 12% de los fumadores tienen un mayor riesgo de cáncer de pulmón. Qué gráfico de los vistos en clase utilizaría? Explique como diseñaría el gráfico dando al menos 3 razones de su elección.
5. En qué tipos de problema de Machine Learning se utiliza la función sigmoidea? Explique por qué se utiliza esta función y cuál es su output.
6. Usted ha entrenado un modelo predictivo para predecir las ganancias de una estación de energía solar en los Emiratos Árabes. Tras haber tenido éxito, una empresa de energía solar en Inglaterra lo contrata para que le realice un trabajo similar. ¿Qué es esperable que suceda con la precisión de su modelo con los datos ingleses y por qué? Justifique su respuesta.
7. Una empresa quiere segmentar sus clientes para ofrecerles ofertas más personalizadas. Su cliente le ofrece un dataset con los patrones de compra de sus clientes y le advierte que existe un pequeño porcentaje de clientes con patrones de compra muy únicos o anormales. A qué tipo de problema cree que se esta enfrentando? Qué algoritmo utilizaría para resolver este problema?
8. Explique 3 aplicaciones prácticas de procesamiento de lenguaje natural no vistas en clase, una vinculada al e-commerce, otra a librerías y la última a medicina.
9. Qué es un outlier? Cómo trataría un outlier de un equipo de maquinaria en una planta de producción?
10. Si poseemos un dataset extremadamente grande, es razonable que el conjunto de entrenamiento tenga un porcentaje mucho más grande o mucho más chico del usual 80/20? Por qué?

A diagram of different types of dots

Description automatically generated

Figura 1