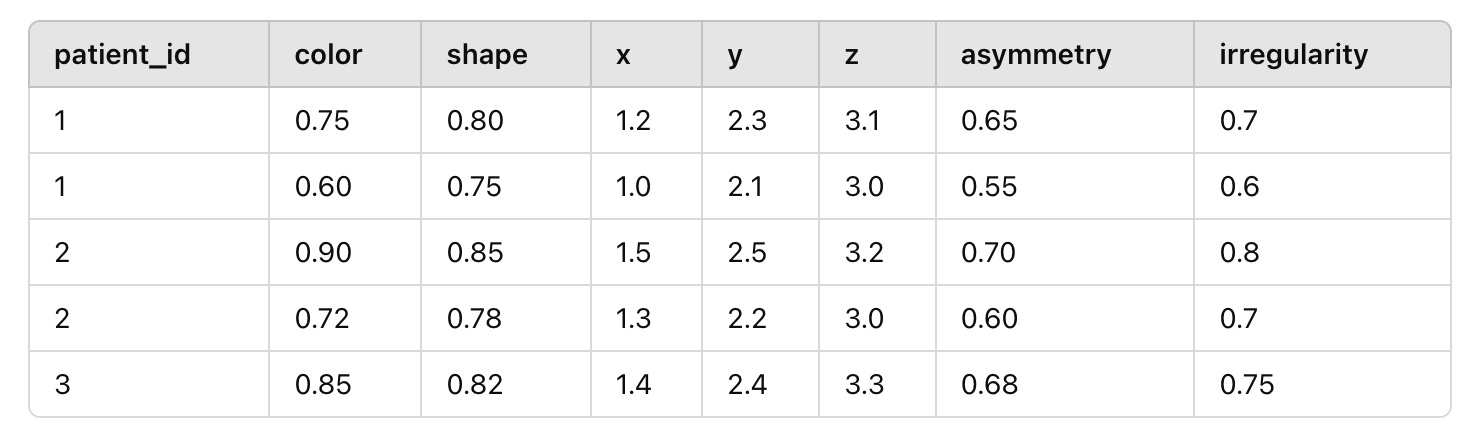
Logo, company name

Description automatically generated

**Introducción a la Ciencia de Datos**

Parcial 07 de Octubre 2024

1. Usted es contratado por un hospital para detectar lesiones cutáneas con el objetivo de clasificarlas como benignas o malignas. En dermatología, un buen indicador de la presencia de un melanoma es mediante la técnica del “patito feo”. Se basa en la idea de que el melanoma a menudo se ve diferente de otros lunares o lesiones en la piel de la misma persona, de manera muy similar a como un "patito feo" se destaca entre patos similares. Usted posee un dataset donde cada fila corresponde al análisis de cada lunar de un paciente, por tanto, cada paciente puede tener varios análisis. Que feature engineering cree que puede ser eficaz para la detección de melanomas?



1. Una empresa financiera lo contrata para intentar detectar fraudes. Usted es informado que el banco tiene costos muy altos por cada fraude. Adicionalmente, el banco no tiene prácticamente costos al detectar erróneamente un fraude, simplemente toma medidas preventivas e informa al cliente que su cuenta puede estar comprometida, congelando movimientos por un corto periodo de tiempo hasta resolver su caso. Qué tipo de problema de Machine Learning esta usted resolviendo? Qué tipo de métrica debe elegir dada la información prevista por el banco? Explique su respuesta.
2. Explique al menos 3 razones por las cuales es necesaria la reducción de dimensionalidad.
3. Qué es un outlier? Cómo trataría un outlier de un equipo de maquinaria en una planta industrial de producción?
4. Una planta industrial produce botellas de shampoo. La etiqueta del envase establece un contenido de 500 ml. El control de calidad toma muestras aleatorias de 30 botellas de un lote de producción reciente y mide los volúmenes, obteniendo una media muestral de 495 ml y una desviación estándar de 8 ml.

Utilizando un nivel de significancia de 0,05, determinar si hay suficiente evidencia para concluir que el volumen promedio de shampoo en las botellas es menor de lo especificado. ¿Debería la planta detener la producción para solucionar este problema?

Plantee las hipótesis necesarias, calcule el estadístico asociado, el p-value e

indique su conclusión sobre si se debe detener la producción, justificando su respuesta.

1. Suponga que usted tiene un conjunto de datos (X, y) para resolver un problema de regresión. A su vez, escoge una metrica que mide el error promedio entre sus predicciones y lo real para evaluar la performance del modelo. Como determinaría los hiperparametros optimos de un modelo de Machine Learning? Asuma que el modelo tiene 2 hiperparametros que debe tunear.
2. 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.
3. Calcule la ganancia de información de la columna "Age Group" con respecto a la variable objetivo "Purchased" utilizando la entropía como medida de impureza. Muestra todo tu trabajo, incluidos los cálculos intermedios. Redondea tu respuesta final a tres decimales.
4. Describa 3 mecanismos de imputación de valores faltantes, ilustrando con ejemplos e indicando las ventajas y desventajas de cada mecanismo.
5. Eres un administrador de base de datos para una librería online. La base de datos de la librería contiene las siguientes tablas:

* Books (BookID, Title, Author, Price, PublisherID)
* Publishers (PublisherID, Name, Address)
* Orders (OrderID, CustomerID, OrderDate)
* OrderDetails (OrderID, BookID, Quantity)
* Customers (CustomerID, Name, Email)

El equipo de marketing quiere realizar una promoción para la editorial más popular según la cantidad de libros vendidos. Escriba una query de SQL que:

1. Calcule el número total de libros vendidos por cada editorial.
2. Identificar la editorial con el mayor número de libros vendidos
3. Devuelve el nombre del editor y el número total de libros vendidos.

A screenshot of a group

Description automatically generated

