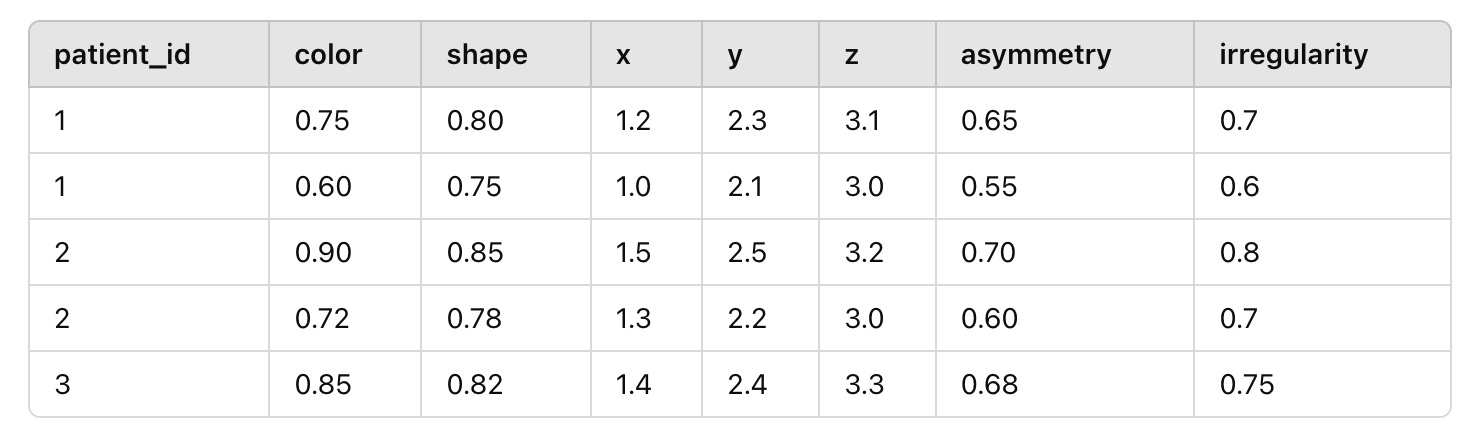
Logo, company name

Description automatically generated

**Introducción a la Ciencia de Datos**

Solucion 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?**



Un enfoque eficaz de feature engineering en este caso sería cuantificar cuánto se desvía cada lunar de los demás lunares del mismo paciente. Específicamente, se pueden crear features que capturen cuán diferente es cada lunar del resto de los lunares, lo que se alinea directamente con el concepto del "patito feo". Algunas características posibles son:

1. Desviación del promedio del paciente (función basada en la distancia): Para cada lunar, calcule el valor promedio de las características clave (por ejemplo, tamaño, forma, intensidad del color, textura) en todos los lunares del paciente. Luego, calcule la desviación de cada lunar de este promedio, quizás usando una distancia euclidiana.

2. Comparación por pares (característica relativa): Para cada lunar, compárelo con todos los demás lunares del mismo paciente, calculando puntuaciones de similitud basadas en características como el color, la asimetría o la irregularidad de los bordes. Identificar al lunar con mayor puntuación de disimilitud como potencial candidato a ser "patito feo".

3. Feature basada en clusters: Agrupar todos los lunares de un paciente determinado mediante un algoritmo de aprendizaje no supervisado (por ejemplo, k-means). Para cada lunar, calcule su distancia desde el centro del grupo. Los valores atípicos en estos grupos probablemente serían candidatos potenciales a melanoma.

Estas features aprovechan la comparación entre lunares del mismo paciente, lo cual es fundamental para identificar valores atípicos que pueden indicar melanoma.

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.**

Este escenario describe un problema de clasificación, específicamente una clasificación binaria donde el objetivo es distinguir entre transacciones fraudulentas y no fraudulentas. En este caso, la clase positiva suele representar fraude, mientras que la clase negativa representa transacciones legítimas.

Dado el alto costo del fraude no detectado y el bajo costo de la detección de fraudes falsos, el objetivo es minimizar los falsos negativos (no detectar el fraude) en lugar de centrarse en los falsos positivos (marcar incorrectamente transacciones legítimas). Por lo tanto, necesita una métrica que capture eficazmente este requisito. Aquí hay métricas adecuadas a considerar:

**Recall** (Sensitivity/True Positive Rate): esta métrica mide la proporción de casos de fraude reales que se identifican correctamente. En este caso, un alto nivel de recuerdo es crucial, ya que pasar por alto un caso de fraude es muy costoso para el banco. La atención debería centrarse en capturar tantos casos de fraude como sea posible.

En resumen, la métrica principal para este problema debe ser Recall, ya que minimizar los falsos negativos es crucial para reducir los costos del banco asociados con el fraude no detectado.

1. **Explique al menos 3 razones por las cuales es necesaria la reducción de dimensionalidad.**
2. **Mitigar la maldición de la dimensionalidad (Curse of dimensionality)**: A medida que aumentan las dimensiones (features) de un dataset, los datos tienden a dispersarse en el espacio, lo que puede hacer que los modelos pierdan capacidad predictiva o encuentren clusters significativos. Esto se conoce como la "maldición de la dimensionalidad". Reducir las dimensiones ayuda a concentrar la información relevante y mejora la capacidad de los modelos para generalizar.
3. **Visualizacion:** con fines de visualizar y entender nuestros conjuntos de datos podemos utilizar tecnicas de reduccion de dimensionalidad para proyectar nuestro espacio a 2d o 3d. Luego en este espacio podemos hacer diversas visualizaciones. Un algoritmo para tal proposito es t-SNE.
4. **Eliminar el ruido y mejorar la interpretabilidad**: En datasets de alta dimensionalidad, no todas las características son relevantes para el problema en cuestión, y algunas pueden introducir ruido o redundancia. La reducción de dimensionalidad, a través de técnicas como PCA (Análisis de Componentes Principales), ayuda a eliminar este ruido, haciendo los datos más fáciles de interpretar y permitiendo visualizar las relaciones subyacentes en menos dimensiones. Asimismo, eliminamos features correlacionadas deshaciendonos de informacion duplicada.
5. **Qué es un outlier? Cómo trataría un outlier de un equipo de maquinaria en una planta industrial de producción?**

En una planta de producción, los equipos de maquinaria pueden generar datos de operación como temperatura, vibraciones, presión, consumo de energía, etc. Un outlier en este contexto podría indicar una anomalía en el funcionamiento de la máquina o un fallo inminente. Sería importante verificar si proviene de un fallo en el sensor, un error de calibración o una lectura errónea. Podriamos verificar si el outlier ocurrió durante condiciones inusuales (como picos de producción o mantenimiento). Esto ayudaría a determinar si es un comportamiento esperado en situaciones especiales. Si el outlier sugiere un problema real con el equipo (por ejemplo, picos de temperatura anómalos o vibraciones excesivas), sería prudente realizar una inspección del equipo o aplicar un mantenimiento preventivo para evitar daños mayores, interrupciones en la producción o incluso fatalidades humanas. En resumen, el tratamiento de outliers en un equipo de maquinaria debe ser cuidadoso, ya que puede indicar tanto un fallo como un error de medición. La acción dependerá de si el outlier es parte de un patrón más amplio o un evento aislado.

1. **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.**

Hipótesis nula (H₀): El volumen medio de shampoo en los frascos es igual a 500 ml.

Hipótesis alternativa (H₁): El volumen medio de shampoo en las botellas es diferente a 500 ml.

Calcular estadístico t:

Como la letra no establece una direccionalidad hacemos un test de dos colas. En la tabla provista esta el lado positivo (diferencia positiva) pero es simetrico debido a la simetría de la distribución t.

Como el tamaño de muestra es 30 nuestro grado de libertad es n-1 = 29. Vamos a la tabla y vemos que el estadístico es mayor que el valor critico de 2.045. El p-value asociado se calcula haciendo una interpolación lineal:

=

Como el p-value es menor que el nivel de significancia alfa decidimos rechazar la hipótesis nula en favor de la hipótesis alternativa y detenemos la planta de producción.

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.**

Divida el conjunto de datos (X,y) en dos subconjuntos:

* **Conjunto de entrenamiento**: Para entrenar el modelo.
* **Conjunto de validación**: Para evaluar la performance del modelo con diferentes combinaciones de hiperparámetros.

Como el problema establece que la metrica es el error promedio debemos minimizar esta metrica. Para determinar la combinacion optima podemos hacer distintas iteraciones con combinaciones distintas de hiperparametros. Esto puede ser mediante un for loop y en cada iteracion evaluar la performance del modelo mediante la metrica comparando:

Luego seleccionamos la combinacion de hiperparametros que tenga el menor error asociado. De manera anecdotica, en Python existen clases para determinar los hiperparametros optimos como GridSearch, RandomSearch o optimizacion bayesiana (optuna).

1. **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.**

Se precisa un grafico impactante que resalte el mensaje. Para esto en clase vimos graficos como “Big Number” o “Icon Array” que son ideales para este propósito. Algunas técnicas de diseño pueden ser:

* Seleccionar colores que contrasten el numero o icono respecto a los otros iconos o números. Por ejemplo, podría ser un color negativo como el rojo ya que queremos enfatizar el peligro de fumar o seleccionar un color mas oscuro en el numero respecto a las otras letras.
* Los iconos pueden ser humanos para demostrar cercania con el interlocutor.
* Uno puede hacer un array de 100 personas y pintar 12 o incluso un array de 8 personas y pintar 1 (12/100 ~ 8). Elegir un color que resalte a los fumadores.

1. **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.**

Ver anexo con solución.

1. **Describa 3 mecanismos de imputación de valores faltantes, ilustrando con ejemplos e indicando las ventajas y desventajas de cada mecanismo.**

Mecanismo: imputacion por la media.

Ejemplo: llenar la altura promedio de un conjunto de personas.

Ventajas: Fácil de implementar y computacionalmente eficiente. Adecuado cuando los valores faltantes son pocos y aleatorios.

Desventajas: Puede distorsionar la variabilidad de los datos, subestimando la varianza. No captura la relación entre variables, lo que puede llevar a sesgos si los datos faltantes no son aleatorios. Susceptible a outliers.

Mecanismo: Imputacion por la media.

Ejemplo: imputar el color preferido.

Ventajas: Fácil de aplicar y rápida de computar. Mantiene la coherencia con el valor más común en la columna.

Desventajas: Puede introducir sesgos si hay una gran cantidad de valores faltantes. Ignora la relación entre la variable imputada y otras variables, lo que podría reducir la precisión del modelo. Problematico en el caso de que sea multimodal.

Mecanismo: KNN-imputer o hot-deck imputation.

Ejemplo: Llenar el peso de una persona teniendo en consideracion la altura de la persona.

Ventajas: Utiliza la estructura del conjunto de datos, lo que puede producir imputaciones más precisas, especialmente cuando existe una relación entre las variables. Adecuado tanto para datos categóricos como numéricos.

Desventajas: Computacionalmente costoso, especialmente para grandes conjuntos de datos. Sensible a la elección del número de vecinos (K) y la métrica de distancia utilizada.

1. **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 computer program

Description automatically generated

Empezamos uniendo las tablas necesarias:

1. Publishers (p) se une a Books (b) mediante PublisherID
2. Luego, Books (b) se unen con OrderDetails (od) usando BookID
3. Usamos SUM(od.Quantity) para calcular el número total de libros vendidos por cada editorial.
4. La cláusula GROUP BY agrupa los resultados por PublisherID y Name, asegurando que obtengamos totales para cada editor.
5. ORDER BY TotalBooksSold DESC ordena los resultados en orden descendente, de modo que la editorial con más libros vendidos aparezca primero.
6. LÍMITE 1 garantiza que solo obtengamos el mejor resultado, que es el editor con la mayor cantidad de libros vendidos.

Esta query devolverá una sola fila que contiene el nombre de la editorial con la mayor cantidad de libros vendidos y la cantidad total de libros vendidos por esa editorial.

A paper with writing on it

Description automatically generatedA screenshot of a group

Description automatically generated

