Miguel Ángel López Rojas
MCD Segundo semestre

Métodos estadísticos multivariados

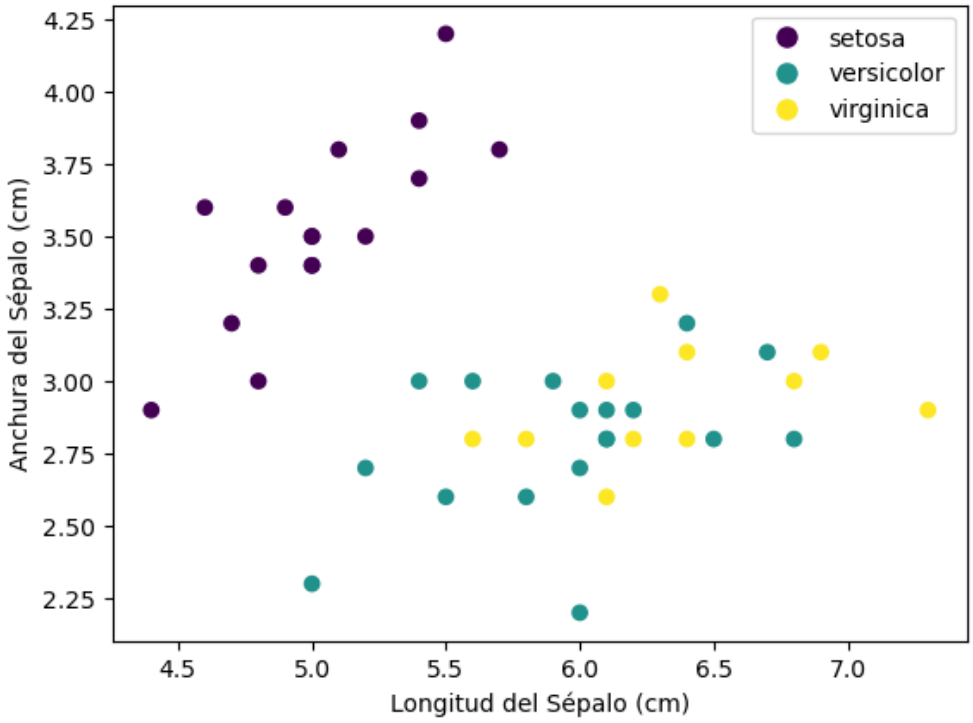
NCR:1684601

Examen 2, 26/ABRIL/2023

PREGUNTAS DE CONCEPTO Y EJEMPLOS ILUSTRATIVOS: (50%)

1. ¿Qué diferencia hay entre los esquemas de aprendizaje-maquina Supervisado, No-supervisado y reforzado?; proporcione un ejemplo descriptivo de cada uno para ilustrarlo.

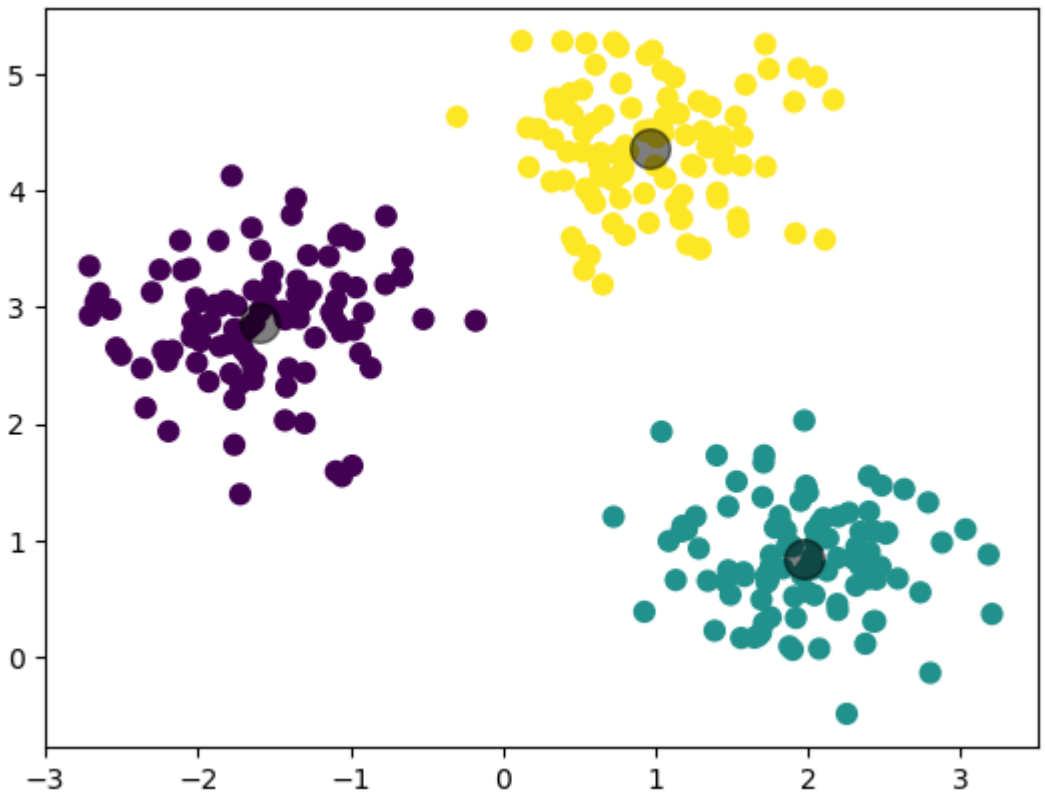
El aprendizaje supervisado implica la utilización de un conjunto de datos etiquetados para enseñarle al algoritmo a clasificar datos nuevos. Por ejemplo, gatos y perros, le decimos al algoritmo las características de ambos y más adelante cuando incluyamos nuevos datos, en base a los previos análisis, sabrá si es perro o gato.



(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

Miguel Ángel López Rojas
MCD Segundo semestre

Cuando es no supervisado entendemos que no se usan etiquetas, es decir, en el ejemplo de perros y gatos el algoritmo encuentra patrones y características en común y los agrupa según sus patrones, quedando al final dos grupos en base a sus semejanzas.



(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

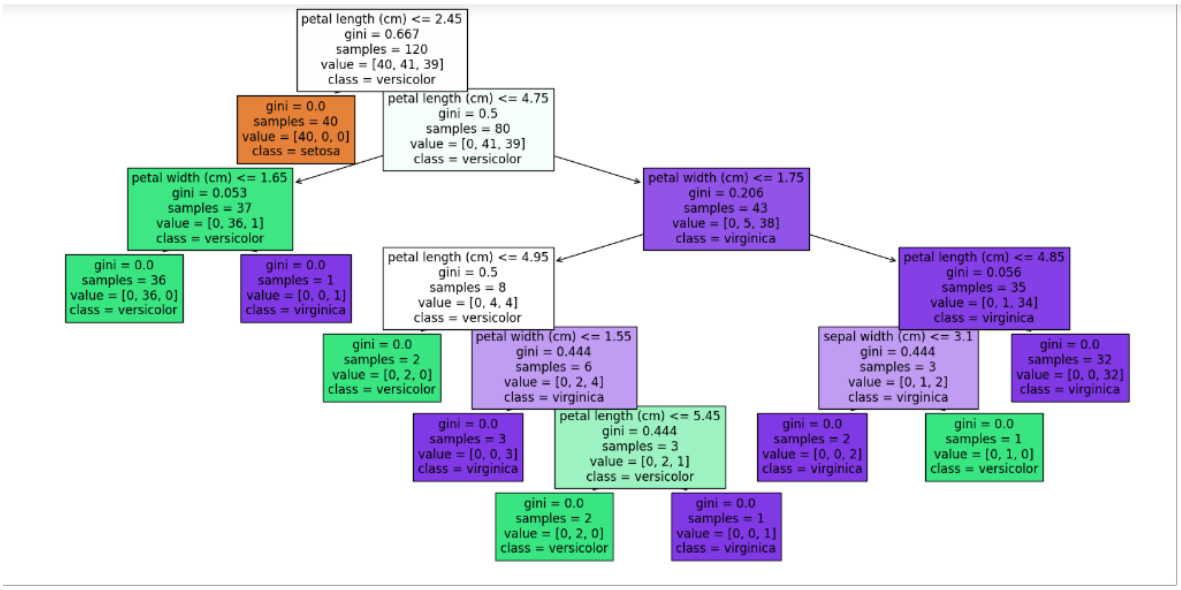
2. ¿Cuál es la diferencia hay entre agrupar y clasificar?; proporcione un ejemplo

ilustrativo que muestre las diferencias.

La principal diferencia se refiere al etiquetado. En la agrupación, nos basamos en encontrar patrones y relaciones entre ellas, como lo mencioné, sin tener información previa sobre las etiquetas o categorías a las que pertenece mientras que en la clasificación se le asigna una etiqueta en base a las características que los componen.

Miguel Ángel López Rojas
MCD Segundo semestre

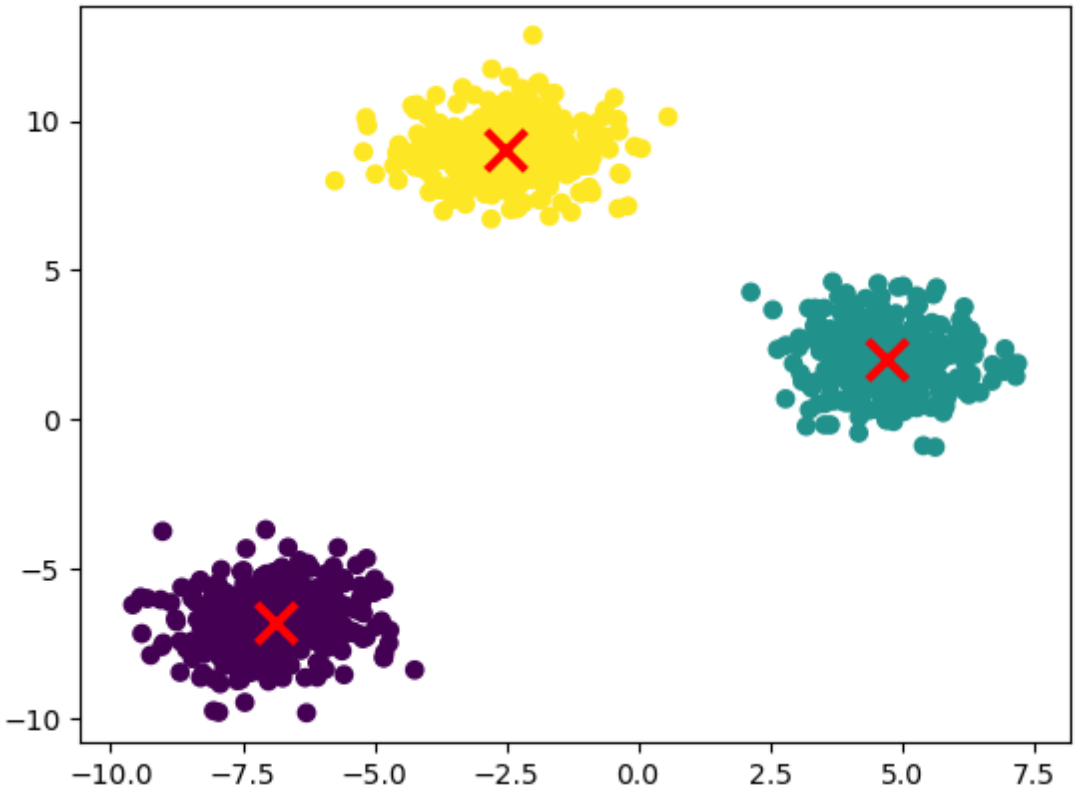
Clasificación:



(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

Agrupamiento

Miguel Ángel López Rojas
MCD Segundo semestre



(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

3.- ¿Que ventajas y desventajas hay entre los siguientes algoritmos de agrupamiento: K-media y jerárquico?

Nombre	Ventajas	Desventajas
K-Means	• Es eficiente en computo • Es escalable (Con muchos datos) • Aplicable para varias dimensiones • Facil de entender	• Depende de la elección de cuantos centroides le diste al inicio. • No funciona con otras formas (De tamaños diferentes o formas irregulares
Agrupamiento Jerárquico	• No hay que especificarle el número de grupos que hacer.	• Es muy pesado e ineficiente en computo.

	• Puede manejar muchas formas y tamaños, contrario a K-means	• No es adecuado para varias dimensiones. • Difícil de interpretar

4. Explique cuales son las diferencias entre los algoritmos de agrupamiento espectral,

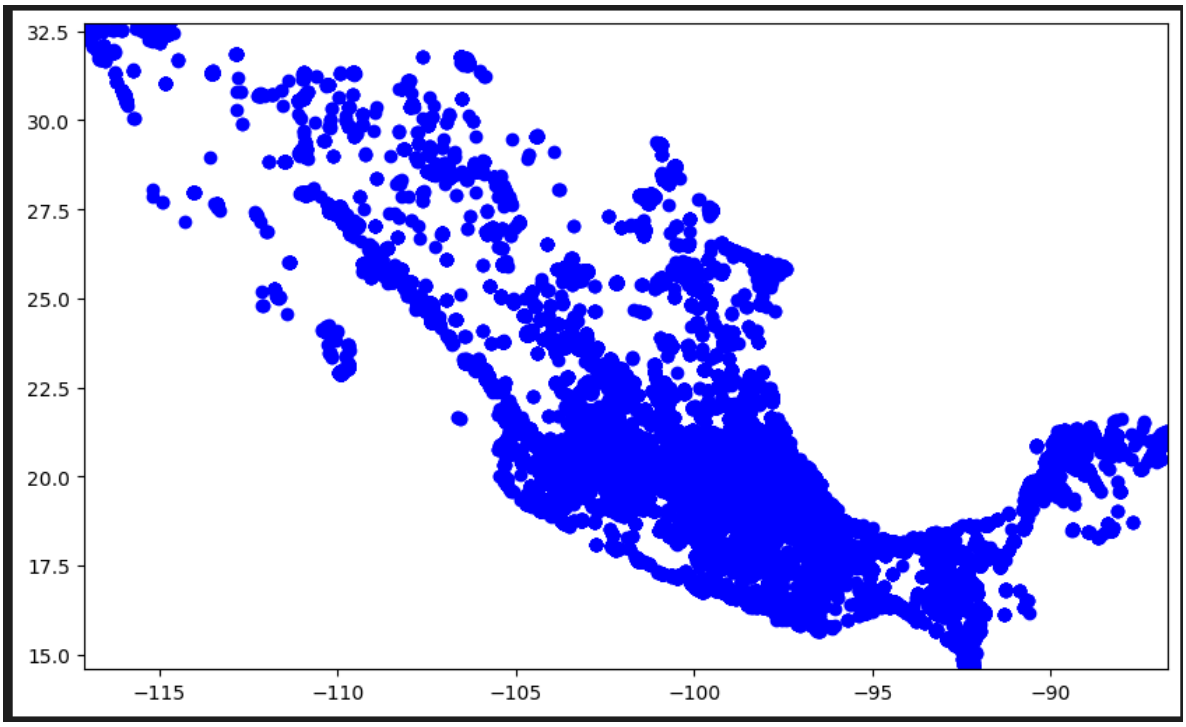
combinatorios (k-means) y de densidad (DBSCAN).

De agrupamiento espectral son aquellos que en base a su información, agrupa a aquellos que tienen características en común en base a sus datos, un ejemplo sería el tener un mapa de México, con las empresas que se dedican a los abarrotes, tengo un plot de donde se encuentran y quiero saber si son mayoristas o minoristas, entonces la distribución geográfica no es tan relevante para esta, sino datos de artículos vendidos y ventas, consolidarían este tipo de datos para saber si son mayoristas o minoristas y los agruparía en esos dos grupos.

Combinatorios son aquellos que tienen características en común, los agrupa en base a cuantos grupos quieres tener, toma todas las características y en base a la distancia que hay de ellos a sus centroides o media (Mean) calcula a que grupo pertenecen, es decir, ¿Qué grupo está más cerca de ese punto? Y así clasifica a cada uno de ellos. Bajo el mismo ejemplo de tener cuantas empresas de abarrotes hay en México, le podríamos decir que seccione la información en 32 (El número de Estados que hay en México) para saber a que estado pertenecen cada uno, aunque bien en este ejemplo, la información no sería tan precisa debido a que todas las formas de los estados tienen forma irregular.

De Densidad se refiere al agrupamiento de los datos en base a su concentración, es decir, agrupa en base a que tantos datos existen dentro de un grupo y crea un grupo más grande de estos objetos, lo que se encuentra fuera de estos cúmulos de información, los considera ruido. En el mismo ejemplo de México y sus abarrotes, pudiera servirnos para ver los espacios que no están tan explotados para poner nuestra abarrotera, medir esos espacios en blanco y considerarlos como un área de oportunidad para los mismos.

Miguel Ángel López Rojas
MCD Segundo semestre



5. Explique las diferencias entre los índices internos y externos, empleado para la

validación de agrupamiento de datos. Proporcione un ejemplo para ilustrar los

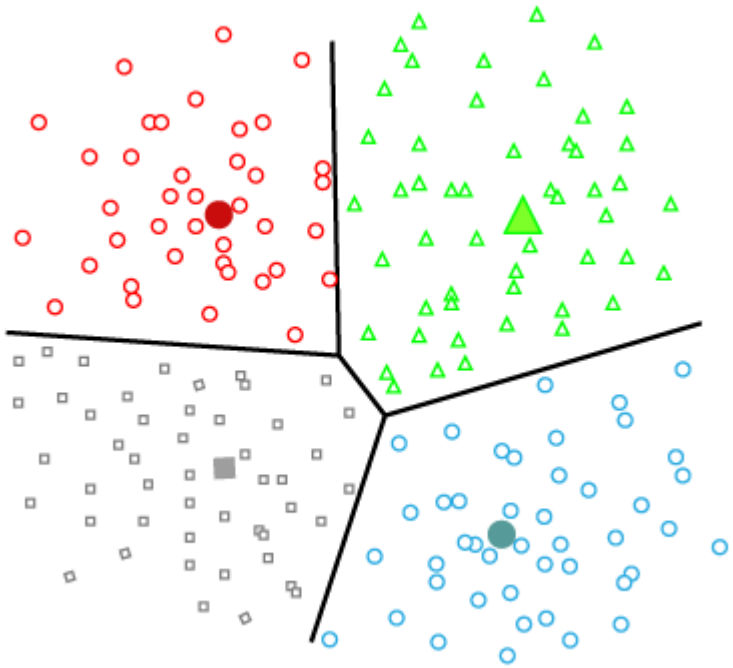
conceptos.

La validación en el clustering es evaluar que tan ben está etiquetando nuestro algoritmo los datos,

Los índices internos miden que tan compactas o separadas están las agrupaciones entre sí, mientras que los índices externos miden que tanto corresponden las etiquetas reales con los grupos obtenidos.

Para los índices internos existen métodos como el de Silhouette y el coeficiente de Davies-Bouldin, mientras que para el externo un ejemplo son los de son el índice Rand y el índice de Jaccard.

Miguel Ángel López Rojas
MCD Segundo semestre



https://static.javatpoint.com/tutorial/machine-learning/images/clustering-in-machine-learning2.png

Para este ejemplo, pensemos que estamos identificando en un campo dividido por un río, un camino de piedra, árboles y por último donde se han hecho fogatas en ese campo y ponerles un pin de donde se encuentran. Después del proceso agrupo y para validar que estén bien etiquetados, primero mediríamos con índices internos que estén bien distribuidos los cuatro grupos de manera adecuada por medio de alguna de las técnicas antes mencionadas como Silhouette, mediría que tanta diferencia de distancias hay entre los grupos y con los índices externos podríamos ver que los arboles correspondan realmente a los arboles y que no haya agua entre ellos, por ejemplo.

Miguel Ángel López Rojas
MCD Segundo semestre

6. Que ventajas y desventajas hay entre los siguientes algoritmos de clasificación: SVM,

K-vecinos cercanos y arboles de decisión.

Nombre	Ventajas	Desventajas
SVM	• Eficiente en conjuntos de datos con alta dimensionalidad • Funciona bien con un margen claro entre clases	• No es adecuado para grandes conjuntos de datos • Puede ser sensible a datos ruidosos
K- Vecino más cercano	• Fácil de entender y de implementar • Funciona bien con conjuntos de datos chico • Puede ser usado para problemas de clasificación y de regresión	• Puede ser sensible a la elección de los parámetros, como el número de vecinos • Uso computacional • No funciona bien con datos de alta dimensionalidad
Árboles de decisión	• Fácil de entender e interpretar • Puede manejar datos numéricos y categóricos • No requiere la normalización de datos • Útil para identificar variables importantes en la clasificación	• Puede ser propenso a overfitting si no se controla • Puede ser sensible a datos ruidosos y atípicos • Pequeños cambios en los datos pueden resultar en árboles de decisión completamente diferentes

Miguel Ángel López Rojas
MCD Segundo semestre

7. Que diferencia existe entre las siguientes métricas para evaluar el desempeño de los

algoritmos de clasificación:

Matriz de confusión, F medición y análisis ROC. Proporcione un ejemplo.

La matriz de confusión es una tabla que muestra la cantidad de veces que un modelo predijo correctamente o errónea, cada clase. Por ejemplo, si estamos clasificando imágenes de gatos y perros, la matriz de confusión pondría cuántas veces el modelo predijo de manera acertada que era un gato y cuantas no.

La F-medida es una medida que combina la precisión y la "exhaustividad". La precisión es la cantidad de veces que un modelo predice correctamente una clase en comparación con el número total de veces que predijo esa clase.

La "exhaustividad" se refiere a la cantidad de veces que el modelo predice correctamente una clase en comparación con el número total de instancias de esa clase en el conjunto de datos, entonces, la F-medida, combina las dos medidas para proporcionar una sola y única medida de rendimiento. Un valor de F-medida alto indica que el modelo tiene tanto una alta precisión como una alta "exhaustividad".

Por último, el análisis ROC (Receiver Operating Characteristic) también es una herramienta utilizada para evaluar la capacidad de un modelo para distinguir entre dos clases.

Representa la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) para diferentes umbrales de clasificación. Un modelo ideal tendría una curva ROC que se acercara a la esquina superior izquierda del gráfico, lo que indica una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.

Un ejemplo sería que se tiene un modelo de clasificación para identificar correos electrónicos spam. Si el modelo clasifica 100 correos electrónicos como spam y 90 de ellos son realmente spam, pero 10 son correos electrónicos legítimos que se clasifican mal, se clasifican como spam, entonces el modelo tiene 90 verdaderos positivos y 10 falsos positivos.

La precisión del modelo sería del 90%, pero esto no refleja que algunos correos electrónicos legítimos se hayan identificado mal, como spam. El resultado del modelo sería del 90% pero esto no refleja que algunos correos electrónicos regulares se hayan identificado mal, es decir, como spam.

La curva ROC del modelo mostraría cómo varía la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos a medida que se ajusta el umbral de clasificación.

Miguel Ángel López Rojas
MCD Segundo semestre

8. Lo métodos utilizados convencionalmente para clasificar (SVM, K-vecinos cercanos

y arboles de decisión) pueden ser utilizados para otros propósitos. Proporcione al

menos 2 aplicaciones potenciales, que no sean la de clasificar.

SVM: Tienes muchos datos y quieres detectar si hay patrones en los datos, por ejemplo, si tienes una base de datos de personas y quieres identificar si existe una relación entre la edad y los hábitos alimenticios. El SVM puede ayudarte a encontrar ese patrón.

K-vecinos cercanos: Tienes una imagen y quieres saber si hay objetos en ella. Puedes usar K-vecinos cercanos para buscar en la imagen patrones que se parezcan a objetos conocidos, como autos o personas.

Árboles de decisión: Estas comprando online y quieres hacer recomendaciones personalizadas a los clientes. Puedes usar árboles de decisión para analizar los datos de compra de cada cliente y ofrecerles productos que se adapten a sus preferencias.

9. Asumiendo, que la etapa de limpieza de datos no fue eficiente, que efecto tendría

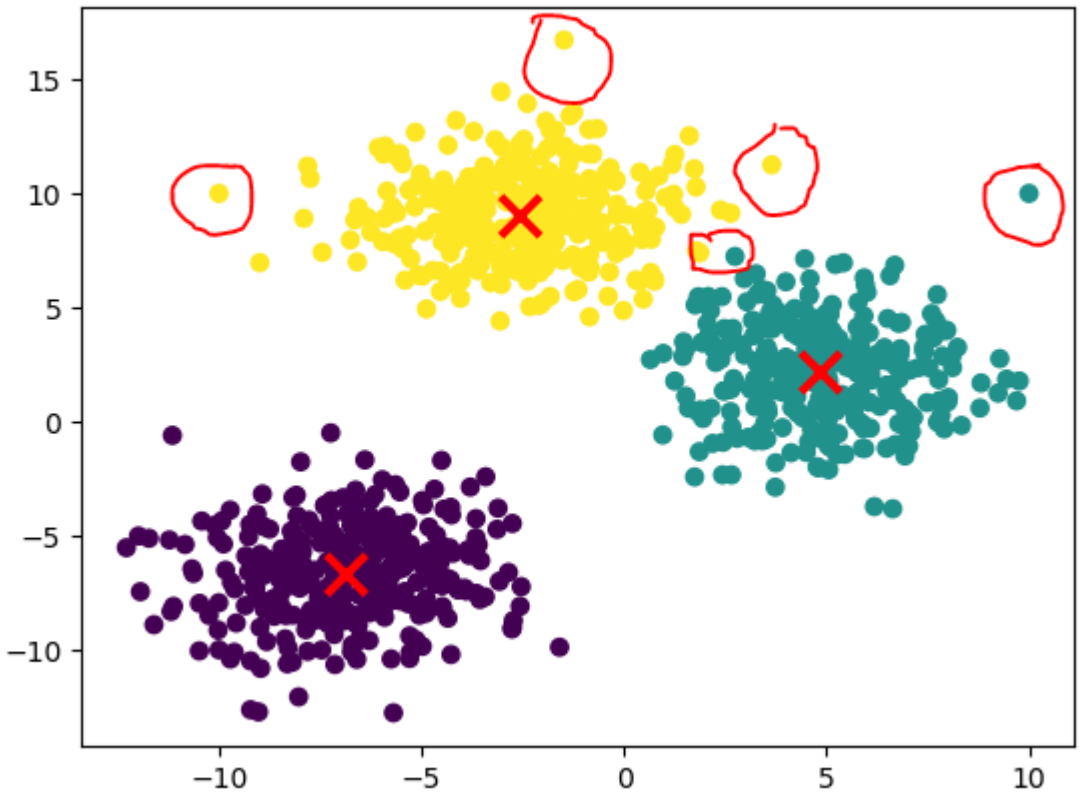
el ruido, datos atípicos y datos faltantes, en los algoritmos de agrupamiento y

clasificación. Proporcione ejemplos ilustrativos de los efectos de cada uno de ellos.

Los efectos de una mala limpieza pueden ser catastróficos pues pueden afecta por completo el outcome de un resultado, como dice el dicho en la industria “Mete basura a tu operación y obtendrás basura”, por ejemplo, trabamos con datos que consideran un todo, como K-means, y tengo 1 dato que quiero clasificar, y el valor nuevo es 10, para evaluar a que grupo pertenecen, el primer grupo su valor valor promedio es 150 (sus tres datos son: 140,160 y 150), el segundo grupo tiene una media de 15 (sus datos son 14, 15 y 16) y pero el tercero vale 110 su media(Sus datos son 10,10 y un outlier de 310), entonces el que más se acerca sería el grupo de 15 y lo clasificaría en ese segundo grupo, cuando el tercero, la mayoría de los números corresponden más pero existe un valor atípico que modifica todo el ejercicio.

A continuación, una ilustración de outliers o basura que podrían estar afectando el ejercicio:

Miguel Ángel López Rojas
MCD Segundo semestre



(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

10. ¿Cuándo es necesario normalizar los datos?, proporcione un ejemplo.

Cuando los datos no se encuentran dentro de la misma escala o las distancias en la escala son bastante grandes o bastantes pequeñas. Pueden ser, por ejemplo, una un conjunto de datos que tiene información sobre la edad y el sueldo de una persona, la edad puede estar en el rango de 0 a 100 años y el sueldo de 1 a 100,000 dólares.

11. ¿Hay diferentes formas de normalizar los datos? (zcores, min-max, unit vector,

estándardization). ¿Qué efecto tiene en los algoritmos de aprendizaje máquina?

Normalizar los datos puede ayudar a mejorar rendimientos, la demás de ser unos más fáciles de entender y apoya a algunos algoritmos de aprendizaje automático, especialmente cuando las características tienen diferentes escalas.

Miguel Ángel López Rojas
MCD Segundo semestre

La normalización puede mejorar el proceso del algoritmo y hacer que los resultados sean más interpretables.

Las diferentes técnicas de normalización, como los z-scores, el min-max, la normalización de vectores unitarios y la estandarización, tienen efectos diferentes en los datos Es importante experimentar con diferentes técnicas y evaluar su impacto en el rendimiento del modelo pero hay modelos que se suponen para ellos:

- Z-score: normaliza los datos para que tengan una media de cero y una desviación estándar de uno, lo que permite comparar valores relativos a la media y la dispersión de los datos. Normalmente se usa cuando tienen una distribución “Normal”, o sea cargada al centro.
- Min-max: transforma los datos para que se encuentren en un rango específico, generalmente entre 0 y 1. Los valores más bajos son transformados en 0 y los más altos en 1.
- Unit vector: escala los datos para que la norma del vector de características sea igual a 1, lo que los convierte en un vector unitario. Este enfoque es útil para algoritmos que utilizan medidas de distancia entre vectores.
- Estándar: normaliza los datos para tener una media de cero y una desviación estándar de uno, pero a diferencia del z-score, utiliza una estimación diferente para la desviación estándar si el tamaño de la muestra es pequeño (Para una distribución NO normal)

12. Que diferencias hay entre escalar, normalizar y estandarizar de los datos.

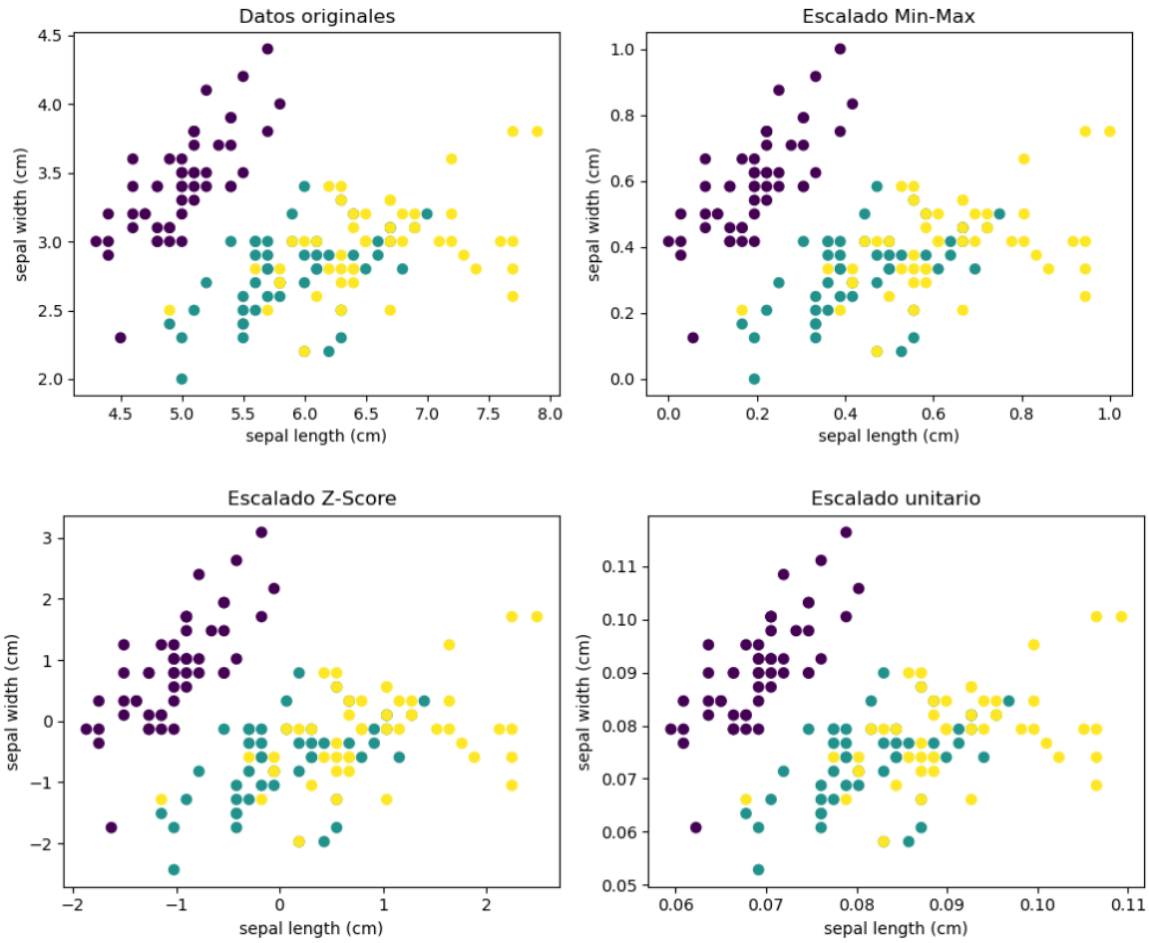
Proporcione un ejemplo que ilustre estos efectos.

Un escalar es como su nombre lo dice, convertir a una escala, como lo mencioné en una pregunta anterior, se usa cuando no tenemos la misma escala, cuando los datos abarcan mucho espacio o un espacio muy pequeño, por ejemplo, convertir todos los números en un rango del 0 al 100 o entre 0 y 1.

Mientras que normalizar, se refiere a la distribución de los datos, se transforman para cumplir una distribución. Por ejemplo, quiero hacer una campana de Gauss y tengo cientos de datos y le quiero dar la forma de la campana, pues sus datos están cargados en la media, por tanto, los normalizo.

Y por último la estandarización se refiere a centrar los datos en cero y una desviación estándar de 1, es decir, a todos les damos la misma distancia.

Miguel Ángel López Rojas
MCD Segundo semestre



(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

En todos los métodos nos podemos observar que la distribución de los datos es la misma, sin embargo, en los ejes X y Y que son diferentes escalas.

13. ¿Cómo se reentrena el modelo de aprendizaje maquina con nuevos datos?

1. Se recopilan los nuevos datos.
2. Se preprocesan, aquí hay que tomar en cuenta si se tenían escalados, si hay que limpiarlos para que no existan outliers o ruido y someterlos a los mismos pasos que los datos anteriores.

3. Combinar los datos nuevos con los que ya se tenía, esto es muy importante pues el chiste del entrenamiento es que se aproveche del entrenamiento anterior y mejore el modelo.
4. Reentrenar el modelo con los datos combinados (Los nuevos y los viejos)
5. Evaluar el rendimiento del modelo, aquí se ve la precisión y relevancia de este nuevos set de datos.

14. Que efecto tiene el desbalance de los datos en los modelos de aprendizaje máquina.

Ilustre la respuesta con un ejemplo.

El desbalance se da cuando una clase o categoría tiene muchos más datos de entrenamiento, que el resto, lo que afecta al aprendizaje de máquina y generalmente produce un sesgo y afecta a su predicción.

In [51]:
```
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Generar datos desbalanceados
X, y = make_classification(n_samples=10000, weights=[0.9, 0.1], random_state=42)

# Separar datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Entrenar modelo de regresión logística
model = LogisticRegression(random_state=42)
model.fit(X_train, y_train)

# Predecir en datos de prueba
y_pred = model.predict(X_test)

# Calcular la precisión del modelo
accuracy = accuracy_score(y_test, y_pred)
print("Precisión del modelo:", accuracy)
```

```
Precisión del modelo: 0.936
```

(Ejemplo y código en documento anexo: Examen_2_parte_1.ipynb)

En ese ejemplo, el 90% de los datos perteneces a la clase 0 y el 10% a la 1 (Y), a este le será fácil predecir la clase 0, más no la 2, debido a la falta de datos.

15. Describa cinco métodos para manejar el desbalance de los datos.

a) Oversampling: Sirve para aumentar la frecuencia de los datos minoristas (Los que hay menos) al agregar copias sintéticas de los ejemplos que ya existen.

b) Undersampling: Reducimos los datos por medio de eliminación de datos al azar.

c) Generación de datos sintéticos: Se crean datos para la clase minorista por medio de los datos existentes. Suele hacerse por métodos de interpolación o extrapolación.

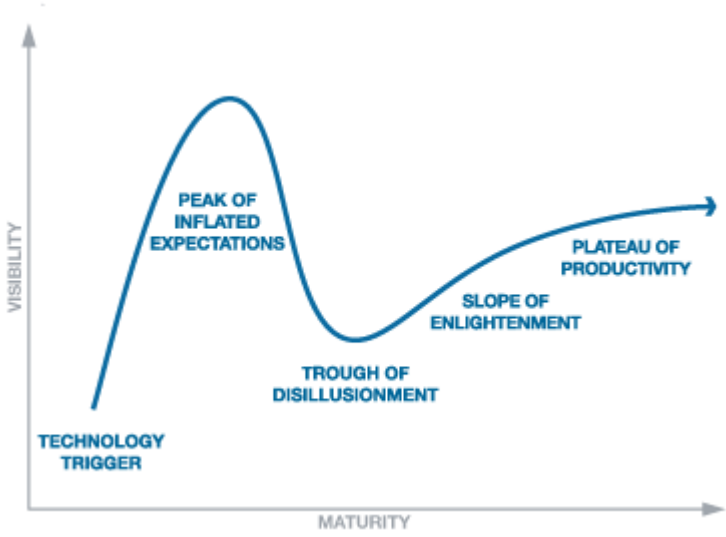
d) Cost-sensitive learning: Modifica el algoritmo de aprendizaje para tener encuentra la asimetría y penaliza los errores en la clase minorista

e) Ensemble learning: Que utiliza varios modelos para mejorar la precisión de las predicciones. Generalmente usa una combinación de sobremuestreo con undersampling.

16. ¿Qué es el Hype Cicle tecnológico de Garner?

Se utiliza para evaluar el ciclo de vida de un producto, es decir, madurez, adopción y aplicación, se llama de Garner porque lo inventó la consultora Garner Inc. Como modelo gráfico para representar este mismo ciclo.

a) Desencadenante: una tecnología se presenta por primera vez y comienza a generar interés.
b) Pico de expectativas infladas: la tecnología gana popularidad
c) Valle de la desilusión: la tecnología no cumple con las expectativas
d) Cuesta del conocimiento: la tecnología comienza a ser mejor comprendida
e) Meseta de productividad: la tecnología se integra en el mercado y se utiliza ampliamente en aplicaciones comerciales.



https://www.bigdata-social.com/wp-content/uploads/2016/03/Gartner-Hype-Cycle.png

17. Analice la curva de Gartner sobre inteligencia artificial 2023.

a) Desencadenante: La introducción de Chat gpt 3.
b) Pico de expectativas infladas: Se escuchó por todos lados a principio del año, sobre como reemplazaría nuestros trabajos
c) Valle de la desilusión: El mercado se da cuenta que no tiene acceso a internet, sobre todos los errores que da y que olvida lo que se le preguntó después de un tiempo.

d) Cuesta del conocimiento: Ya no es un miedo a ser reemplazados, nos damos cuenta que con una buena descripción de lo que buscamos, podemos obtener resultados mejores
e) Meseta de productividad: complementamos nuestros trabajos con el uso de la herramienta y toda la revolución que trajo esta inteligencia artificial.

18. Proporcione un ejemplo de cómo usaría alguna de las herramientas aprendidas en

el curso (reducción de dimensión, agrupamiento, modelado con datos, clasificación,

..) con métodos para la visualización de los resultados, por ejemplo Folium o Qgis.

En mi tesis estoy implementando varios recursos vistos en esta clase. Lo que busca mi tesis es encontrar el mercado más interesante para abarrotes mayorista, por lo que por medio de Folium mapeamos todas las empresas abarroteras, después viene el calculo de DBSCAN para medir los espacios entre los grupos, así comenzaremos a enfocarnos en el Estado que tenga el menor número de empresas mayoristas, para investigarlo de manera particular.

19. Analice las hojas de características (datasheet) de los algoritmos de aprendizaje máquina de Azure y scikit-learn. Mencione algunas diferencias.

Las hojas de comportamiento o características, son documentos que describen los comportamientos de un modelo de machine learning, desde sus entradas, salidas, limitaciones y hasta restricciones éticas.

Las de Azure y Scikitlearn son diferentes las de Azure consisten en un formato más tabular, mientras que las de Sklearn son un poco más descriptivas o narrativas.

las hojas de características de Azure son más detalladas y completas en la descripción de los parámetros del modelo, las técnicas de optimización y los requisitos de hardware, las hojas de características de scikit-learn son más concisas y se centran principalmente en la descripción de las entradas y salidas del modelo.

Y por último las hojas de Azure son más especificas para este modelo, mientras que las de Scikit-learn son más generales y se pueden usar para otros modelos.

20. En un párrafo proporcione una descripción general del proyecto final del curso que

presentara.

El proyecto final será el menciona en la pregunta 18, el fin es el encontrar el mercado más atractivo para poner una abarrotera mayorista. Al estar tan saturado el mercado, el encontrar el lugar menos saturado es clave, por tanto, el uso de paqueterías como Folium para ubicar los datos en la gráfica

vienen en primera instancia. Después es fundamental el agrupar y calcular la distancia entre los puntos por medio de DBSCAN, así aterrizaremos cuál es el mercado sobre el cuál enforcarme y comenzar a analizar a los mismos.

SEGUNDA PARTE DEL EXAMEN RESPONDIDA Y EXPLICADA EN EL DOCUMENTO: Examen_2_parte_1.ipynb