1.

En el **aprendizaje supervisado**, el modelo aprende usando datos etiquetados, es decir, con ejemplos ya clasificados. Se utiliza para problemas de clasificación. Los algoritmos comunes son: **regresión logística**, **KNN**, **árboles de decisión** y **SVM**.

En el **aprendizaje no supervisado**, el modelo trabaja con datos sin etiquetas, buscando patrones ocultos. Es útil para agrupaciones. Los algoritmos comunes son: **K-Means**, **DBSCAN** y **clustering jerárquico**.

2.

La **segmentación** es importante para agrupar datos en subgrupos con características similares, lo que permite entender mejor los patrones y tomar decisiones informadas. Esto es útil en marketing, personalización de productos y análisis de clientes.

Para que un clúster sea confiable se debe cumplir que sea:

Homogéneo: los datos dentro del grupo son similares.

Heterogéneo: los grupos son diferentes entre sí.

Estable: los clústeres no cambian mucho con nuevos datos.

Interpretativo: los resultados tienen sentido práctico.

Los algoritmos que se usan para segmentación son **K-Means**, **DBSCAN**, **Clustering jerárquico** y **Mezcla Gaussiana (GMM)**, cada uno adaptado para distintos tipos de datos y patrones.

3.

En **clasificación** los datos clave del negocio incluyen información demográfica, historial de compras y comportamiento del cliente.

Los algoritmos que suelen usarse son **regresión logística**, **árboles de decisión**, **Random Forest**, **KNN** y **SVM**, que ayudan a clasificar o predecir resultados basados en estos datos.

4.

Un modelo de **regresión** se usa en situaciones como predecir ventas futuras, calcular el precio de un producto o estimar el crecimiento de una empresa. Los algoritmos que ayudan a hacer estas predicciones son **regresión lineal**, **regresión polinómica**, **Ridge** y **SVR**, estos hacen relaciones entre las variables para hacer estimaciones numéricas.

5.

Los principales objetivos de una serie de tiempo son identificar patrones y tendencias, predecir valores futuros y entender cómo se relacionan las variables. Para construirla, necesitas dos tipos de variables: el tiempo esta es la base de medición y la variable que deseas medir como por ejemplo, ventas o temperaturas.

Para una serie de Tiempo confiable debe cumplir con lo siguiente:

1. **Reúne datos:** Consigue datos precisos en intervalos regulares, como diarios o mensuales.
2. **Limpia los datos:** Corrige errores y maneja cualquier dato faltante.
3. **Grafica los datos:** Haz gráficos para visualizar patrones y tendencias.
4. **Descompón los datos:** Separa la serie en sus componentes: tendencia, estacionalidad y ruido.
5. **Modela los datos:** Utiliza métodos estadísticos para hacer predicciones.
6. **Evalúa el modelo:** Revisa y ajusta el modelo según sea necesario.

6.

Componentes:

1. **Nivel (L):** El promedio actual de la serie de tiempo.
2. **Tendencia (T):** Nos dice si la serie está aumentando o disminuyendo.
3. **Estacionalidad (S):** Refleja los patrones que se repiten en intervalos regulares, como las ventas que suben en ciertas épocas del año.

Parámetros:

- **α (alfa):** Controla cuánto peso tienen los datos recientes en el nivel.
- **β (beta):** Determina la influencia de los cambios recientes en la tendencia.
- **γ (gamma):** Ajusta la importancia de los datos estacionales recientes.

7.

Es un modelo de series de tiempo que combina tres componentes:

auto-regresivo (AR): que utiliza valores pasados para hacer predicciones

Integrado (I): que se refiere a la diferencia de los datos para eliminar tendencias y hacer la serie más estable

Media Móvil (MA):que ajusta los pronósticos considerando errores pasados.

Es recomendable usar ARIMA cuando la serie de tiempo no presenta estacionalidad o esta ha sido eliminada, cuando hay tendencias presentes que necesitan estabilizarse y cuando se dispone de suficientes datos histórico.

8.

*****Respuesta en la hoja

9.

El modelo de asociación se utiliza para descubrir relaciones entre variables en grandes conjuntos de datos, como en análisis de mercado o recomendaciones de productos.

Algoritmos de Asociación:

Apriori: Busca ítems frecuentes mediante un enfoque de "cortar y pegar". Es eficiente en datos pequeños a medianos, pero puede ser lento en grandes conjuntos.

FP-Growth: Crea un árbol compacto (FP-tree) para representar los datos y encuentra patrones sin generar combinaciones, lo que lo hace más rápido y eficiente en memoria.

Eclat: Usa un enfoque de búsqueda vertical, donde los ítems se representan en listas de transacciones, realizando intersecciones para encontrar ítems frecuentes. Es más efectivo con muchos ítems.

10.

El algoritmo **DBSCAN** identifica clústeres agrupando puntos cercanos según su densidad. Busca **puntos centrales** que tengan suficientes vecinos dentro de un radio específico (lamado **epsilon** y agrupa esos vecinos juntos. Los puntos que no se agrupan se consideran ruido.

Parámetros:

Epsilon (ε): Define el radio de búsqueda. Un valor pequeño puede resultar en muchos clústeres pequeños, mientras que uno grande puede fusionarlos.

MinPts: Número mínimo de puntos necesarios para que un punto sea considerado central. Un valor más alto genera menos clústeres.

11.

El algoritmo **SVM** se usa para clasificación y regresión, buscando la mejor línea o diferencial que separa diferentes clases.

Características:

Separación óptima: Encuentra el margen máximo entre clases, mejorando la precisión.

Kernel trick: Maneja datos no lineales al transformar el espacio de características con diferentes funciones kernel.

Robustez: Funciona bien en espacios de alta dimensión y con pocos datos de entrenamiento.

Ejemplo:

Clasificación de Perros y Gatos

Características: Las características de las imágenes pueden incluir el color predominante, la textura y la forma de las orejas.

Entrenamiento del modelo: Usas un conjunto de imágenes ya etiquetadas como perros o gatos para entrenar el modelo. SVM analizará estas imágenes y buscará patrones que ayuden a distinguir entre ambas categorías.

Clasificación: Cuando subes una nueva imagen, SVM utiliza lo aprendido para determinar si la imagen es de un perro o un gato, trazando una línea que separe ambas clases.

12.

Los principales componentes de una red neuronal son:

las neuronas: que procesan la información mediante entradas ponderadas y funciones de activación

las capas: que incluyen la de entrada, ocultas y de salida

los pesos y sesgos: que ajustan la importancia de las entradas.

Estos elementos trabajan en conjunto para permitir que la red aprenda patrones.

Los tipos de Redes Neuronales vistas son:

Redes Neuronales Artificiales (ANN):utilizadas para clasificación de imágenes

Redes Neuronales Convolucionales (CNN): especializadas en procesar imágenes y videos

Redes Neuronales Recurrentes (RNN): manejan datos secuenciales, ideales para traducción automática y análisis de series de tiempo.

13.

Las actividades para implementar modelos predictivos en la Minería de Datos son:

Definición del problema: Establecer claramente el objetivo del análisis, como predecir el comportamiento del cliente o detectar fraudes.

Recolección de datos: Recopilar datos de diversas fuentes, como bases de datos internas, encuestas o registros históricos.

Preparación de datos: Limpiar los datos, manejando valores faltantes y eliminando duplicados. Realizar transformaciones necesarias para asegurar la calidad.

Exploración de datos: Utilizar herramientas de visualización y análisis descriptivo para entender mejor los patrones y tendencias en los datos.

Selección de características: Identificar y seleccionar las variables más relevantes que impactan el modelo, utilizando técnicas como análisis de correlación.

Modelado: Aplicar algoritmos adecuados, como regresión, árboles de decisión o redes neuronales, para construir el modelo predictivo.

Evaluación del modelo: Medir el rendimiento con métricas como precisión y recall, ajustando el modelo según sea necesario para mejorar su efectividad.

Implementación y monitoreo: Integrar el modelo en el sistema existente y realizar un seguimiento continuo de su rendimiento para asegurar su efectividad en el tiempo.

14.

El propósito de evaluar el rendimiento de un modelo predictivo es asegurar que haga buenas predicciones y pueda generalizar a nuevos datos. Esto ayuda a elegir el mejor modelo y a mejorar su efectividad.

Medidas de rendimiento por tipo de algoritmo:

Regresión:

Error Cuadrático Medio (MSE): Mide el promedio de los errores al cuadrado entre las predicciones y los valores reales.

R^2: Indica cuánta variabilidad de los datos se explica con el modelo.

Clasificación:

Precisión: Porcentaje de predicciones correctas sobre el total de predicciones.

Recall: Proporción de verdaderos positivos sobre todos los casos positivos reales.

F1-Score: Combina precisión y recall en una sola medida.

Matriz de Confusión: Muestra los verdaderos y falsos positivos y negativos.

Series de tiempo:

Error Absoluto Medio (MAE): Promedio de los errores absolutos entre las predicciones y los valores reales.

RMSE: Raíz cuadrada del MSE, que da una medida en las mismas unidades que los datos.

15.

Para poder seleccionar el mejor algoritmo es indispensable conocer:

1. **Tipo de problema:** Determina si necesitas clasificar, predecir valores, agrupar datos o detectar anomalías.
2. **Características de los datos:** Evalúa el tamaño, calidad y tipo de datos (numéricos o categóricos) y si hay valores faltantes.
3. **Interpretabilidad:** Decide si necesitas entender cómo funciona el modelo. Algunos algoritmos, como los árboles de decisión, son más fáciles de interpretar.
4. **Tiempo de entrenamiento:** Considera cuánto tiempo y recursos requiere el algoritmo, especialmente con grandes volúmenes de datos.
5. **Precisión:** Comprueba qué tan bien el algoritmo puede hacer predicciones en datos nuevos.
6. **Escalabilidad:** Asegúrate de que el algoritmo pueda manejar más datos si es necesario.

16.

a) **Análisis de reclamaciones:**

Modelo: Clasificación (como árboles de decisión o regresión logística).

Resultados: Identificación de patrones en reclamaciones fraudulentas y mejora en la detección de fraudes.

b) **Optimización de precios:**

Modelo: Regresión (para predecir la demanda en función del precio) o algoritmos de optimización.

Resultados: Precios ajustados que maximizan ingresos y mejora en la competitividad del producto.

c) **Simulación de presupuestos:**

Modelo: Series de tiempo o análisis predictivo.

Resultados: Proyecciones de ingresos y gastos, permitiendo una mejor planificación financiera.

d) **Optimización de campañas:**

Modelo: Análisis de clusters y modelos de clasificación.

Resultados: Segmentación de clientes para dirigir campañas más efectivas y mejorar el retorno de inversión.

e) **Demanda de inventarios:**

Modelo: Series de tiempo o regresión.

Resultados: Predicciones precisas de la demanda, lo que ayuda a gestionar el inventario de manera más eficiente y reducir costos.

f) **Diagnósticos médicos:**

Modelo: Clasificación (como redes neuronales o máquinas de soporte vectorial).

Resultados: Diagnósticos más precisos y rápidos, mejorando la atención al paciente y la detección temprana de enfermedades.