

Análisis de Caso: Fundamentos del Aprendizaje de Máquina

Empresa: DataSolutions S.A.

1. Clasificación del problema

Problema	Descripción	Tipo de Tarea
A	Predecir el monto de ventas semanales de una cadena de supermercados	Regresión supervisada
B	Detectar si un cliente abandonará un servicio de streaming	Clasificación supervisada
C	Agrupar a los clientes bancarios por comportamiento de gasto sin etiquetas	Aprendizaje no supervisado (Clustering)

2. Selección de modelo y justificación

Problema	Modelo sugerido	Justificación	Ventajas	Limitaciones
A	Regresión lineal o polinómica	El objetivo es predecir un valor numérico continuo	Simplicidad, interpretabilidad	Puede subestimar relaciones no lineales
B	Árbol de decisión o Random Forest	Clasificar abandono, útil con variables categóricas y numéricas	Robusto a outliers, no necesita escalado	Puede sobreajustarse
C	K-means clustering	No hay etiquetas, se necesita agrupar por similitud	Simplicidad, segmentación útil	Sensible a escala y a elección de K

3. Riesgos y desafíos

Riesgo	Descripción
Overfitting	El modelo aprende demasiado el entrenamiento y generaliza mal.
Datos desbalanceados	Puede sesgar los modelos hacia la clase mayoritaria.
Escalamiento necesario	Modelos basados en distancia fallan si no se escalan los datos.
Datos faltantes	Reducen calidad del entrenamiento y la predicción.
Maldición de la dimensionalidad	Muchos atributos irrelevantes dificultan agrupaciones y predicciones.

4. Flujo de trabajo para el Problema B

1. Recolección de datos históricos de usuarios
2. Limpieza y tratamiento de datos nulos
3. Codificación de variables categóricas
4. División en conjuntos de entrenamiento/test
5. Escalamiento si aplica
6. Selección de modelo (árbol, random forest, etc.)
7. Entrenamiento del modelo
8. Evaluación con métricas (precisión, recall, F1)
9. Ajuste de hiperparámetros si es necesario
10. Implementación y monitoreo continuo

5. Reflexión personal

Aplicar estos conceptos permite identificar el tipo correcto de problema y evitar errores comunes, como usar clasificación en tareas de regresión o no escalar datos en modelos que lo requieren. Además, permite anticipar riesgos y elegir métricas adecuadas según el caso. Este enfoque sistemático profesionaliza el análisis de datos y mejora la calidad de las decisiones basadas en modelos predictivos.