Proyecto Final: Módulo 3

Análisis y modelado de calidad de manzanas

Profesor: M.I. Canek Mota Delfin

Fecha de entrega: 23/05/2025

Objetivo

Desarrollar pensamiento crítico en torno a un problema real: clasificar la calidad de manzanas para reducir desperdicio y optimizar la cadena de suministro. Se evaluarán *más* los argumentos y hallazgos analíticos que la complejidad algorítmica.

1. Contexto y motivación

La calidad de la fruta impacta en precios, satisfacción del consumidor y pérdidas económicas. Trabajarás con medidas físicas y químicas para responder:

¿Qué hace "buena" a una manzana y qué tan bien podemos predecirlo?

2. Entregables

1. **Notebook (*.ipynb)** reproducible y comentado.
2. **Reporte** (PDF.) resumiendo hallazgos clave y recomendaciones.

Descripción del Dataset

El conjunto se obtuvo de Kaggle (Se realizaron modificaciones al original):

- **Enlace:** `https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality`

Variables disponibles:

- **Tamaño:** Milímetros (mm) de diámetro
- **Peso:** Gramos (g)
- **Dulzor:** Grados Brix (°Brix)

- **Crujiente:** Kilogramos-fuerza (kgf)
- **Jugosidad:** Porcentaje (%) de agua
- **Maduración:** Índice de almidón
- **Acidez:** pH
- **Calidad:** Etiqueta categórica (*"good"* o *"bad"*)

Instrucciones del proyecto

1. Carga y exploración inicial

- Carga el archivo CSV en un DataFrame.
- Usa `head()`, `info()` y `describe()` para conocer el contenido.

2. Limpieza de datos

- Verifica la existencia de valores nulos o duplicados.
- Asegúrate de que los tipos de datos sean correctos.
- Trata los outliers si se identifican.

3. Análisis exploratorio de datos (EDA)

- Analiza las distribuciones de cada variable (histogramas, etc).
- Usa gráficas de dispersión y boxplots para comparar variables con la calidad.
- Muestra la matriz de correlación y un mapa de calor.

4. Preparación para el modelado

- Codifica la variable `calidad` como binaria (en caso de ser necesario).
- Aplica normalización o estandarización si es necesario.
- Divide los datos en conjuntos de entrenamiento y prueba (80/20).

5. Modelado predictivo

Entrena al menos dos modelos de clasificación supervisada entre los siguientes:

- Regresión
- Árboles de Decisión
- Random Forest
- SVM

Evalúa cada modelo utilizando:

- Accuracy
- Matriz de confusión
- Precisión, Recall y F1-Score

6. Conclusiones

Redacta una sección con respuestas a:

- ¿Qué modelo funcionó mejor y por qué?
- ¿Qué variables fueron más relevantes?
- ¿Qué limitaciones tuvo el análisis?
- ¿Cómo se podría mejorar el modelo con más datos?

Recomendaciones técnicas

- Usa librerías como `pandas`, `numpy`, `matplotlib`, `seaborn`, `scikit-learn`
- Documenta claramente tu notebook: explica cada paso como si otro estudiante necesitara entenderlo desde cero.
- El notebook debe ser funcional, reproducible y estar bien organizado.

3. Rúbrica de evaluación

Criterio	**Peso**
EDA profundo e interpretación	35 %
Modelos	25 %
Interpretabilidad y narrativa	20 %
Reproducibilidad	10 %
Presentación y claridad	10 %