

Clasificación de Frutas y Verduras

Objetivo

Entrenar un modelo de aprendizaje automático para clasificar alimentos como frutas o verduras, basándose en sus propiedades nutricionales. La clasificación se realiza a partir de un conjunto de entrenamiento de 35 ejemplos, utilizando variables numéricas como vitaminas, minerales, calorías, etc.

1. Investigación de variables y dominio

Cada fila del dataset representa un alimento identificado por su **name**. Las columnas contienen variables nutricionales como:

- Vitaminas A y C (% del valor diario)
- Minerales: calcio, hierro, magnesio, potasio
- Macronutrientes: proteína y fibra
- Calorías por cada 100g

La variable objetivo es **classification**, con valores **fruit** o **vegetable**.

2. Análisis descriptivo y detección de outliers

Se eliminaron las dos primeras filas del archivo original (contenían metadatos del software Orange). Luego se convirtieron las columnas a valores numéricos para análisis.

Estadísticas:

- **vitamin A %** tiene valores extremos (máx. 334%), lo cual puede afectar al modelo.
- **calories (per 100g)** varía entre 16 y 160.
- **fiber (g)** y **protein (g)** muestran buena dispersión.

- La variable con mayor correlación con la clase es calorías (0.43), seguida de fibra.

Se detectaron valores atípicos en algunas columnas, pero se decidió conservarlos dado que representan diferencias naturales entre alimentos.

3. Verificación y ajustes de tipos de variables

Se realizó:

- Conversión de todos los atributos numéricos desde `object` a `float`.
 - Eliminación de registros con valores faltantes (`NaN`).
 - Validación de que no existan duplicados ni inconsistencias.
-

4. Manejo de datos ausentes o atípicos

- Se eliminaron registros no numéricos al limpiar el dataset.
 - No hubo datos ausentes tras la limpieza.
 - Los outliers detectados en variables como `vitamin A` o `calories` se conservaron para no perder información, dado el tamaño reducido del dataset.
-

5. Análisis de correlaciones

Se codificó la variable objetivo `classification` como binaria (`fruit=1`, `vegetable=0`). Se obtuvieron las siguientes correlaciones:

Variable	Correlación con ser fruta
Calories	0.43
Fiber	0.24

Vitamin C -0.09

Iron -0.31

Protein -0.41


Variables como **protein**, **calcium**, **magnesium** y **vitamin A** se asocian más con verduras, mientras que calorías y fibra están más asociadas con frutas.

Preprocesamiento aplicado

- Limpieza de filas no válidas.
 - Conversión de tipos a **float**.
 - Estandarización de variables numéricas.
 - Codificación de la clase (**fruit**, **vegetable**) como etiqueta binaria para modelado.
-

Modelado y evaluación

Se entrenaron varios modelos utilizando validación cruzada:

Modelo	Accuracy
Regresión Logística	88.6%
KNN (k=3)	91.4%
Árbol de decisión	88.6%
Random Forest	94.3% 

El modelo Random Forest obtuvo el mejor rendimiento. Se seleccionó como modelo final.