

# Clasificador de semillas de trigo



## ¿De que va esto?

Nuestro problema se basa en los datos de 3 variedades de semillas de trigo, nuestro objetivo es que a partir de sus características podamos clasificar que semilla de trigo es, con el fin de crear una herramienta que nos permite identificar que semilla es.

## Análisis de los datos

El dataset es adecuado para un problema de clasificación, esta formado por las tres clases de semillas de trigo (Kama, Rosa y Canadian) y las siguientes características:

Área, perímetro, compacidad, longitud, ancho, asimetría y longitud surco



## Validación

División de los datos en train(80%), validation(10%) y test(10%).

Para su validación nos apoyamos en la matriz de confusión:

Recall, precisión, F1 score y accuracy.

Usamos GridSearchCV para agrupar y validar.

## Modelos y parámetros

Para la resolución de este clasificador, usamos los siguientes modelos:

KNN: K vecinos más cercanos, los vecinos votan con su clase.

- n\_neighbors, weight, p.

Regresión logística multiclasa: Clasifica usando una función logística:

- C, multi\_class, solver.

Árboles de decisión: Se recorren sus ramas hasta llegar al resultado:

- criterion, max\_depth, min\_sample\_split, min\_samples\_leaf, ccp\_alpha



## Resultados y distintas fases

Integración: Comprobar el dataset y cargar datos.

Preprocesamiento: Valores nulos, duplicados, outliers, selección variables..

Modelo: KNN, Regresión logística y Árboles de decisión.

Evaluación: Matriz de confusión, grid y error cuadrático medio.

Interpretación y difusión: Aplicación del modelo en el mundo real.

## Conclusiones

Tras realizar los pasos anteriores, obtenemos un programa que a partir de los parámetros nos devuelve la clasificación de la semilla. Su desarrollo nos ayuda a entender la metodología KDD

