# Tarea de acción 4

## Identificación del estudiante

| | |
|---|---|
| **Nombre** | |
| **Profesión** | |
| **Institución** | |
| **Ciudad - País** | |
| **Correo electrónico** | |
| **Módulo 4** | Modelo lineal |

Para el desarrollo de esta tarea es necesario estudiar el material correspondiente a este módulo, junto con las clases sincrónicas.

En esta actividad, nos sumergiremos en el emocionante mundo del aprendizaje automático aplicado a la agricultura. ¿Alguna vez te has preguntado cómo las máquinas pueden ayudarnos a predecir calidad de las frutas? Bueno. Hoy exploraremos precisamente esto.

Tendremos a disposición una base de datos que contiene una variedad de información sobre manzanas: su madurez, peso, tamaño, dulzor y otras. Además, contamos con un dato adicional para cada manzana, una etiqueta que nos indica la calidad general de esta (Good or Bad).

El objetivo de esta actividad es entrenar una máquina de soporte vectorial para predicción de calidad de manzanas.

Además, se deberá entrenar un modelo de regresión lineal múltiple para estimar el grado de dulzor de la manzana.

La base de datos viene en formato csv (**apple_quality.csv**). Los campos disponibles son:

- **A_id**: Identificación único de cada fruta
- **Size**: Tamaño de la fruta
- **Weight**: Peso de la fruta
- **Sweetness**: Dulzor de la fruta
- **Crunchiness**: Textura que indica el carácter crujiente de la fruta
- **Juiciness**: Grado de jugosidad de la fruta
- **Ripeness**: Grado de madurez de la fruta

- **Acidity**: Grado de acidez de la fruta
- **Quality**: Calidad general de la fruta

## Actividades:

1. Cargar la base de datos **apple_quality.csv** y realizar proceso de limpieza, identificando valores outliers, ausentes y valores con error.

   Realizar un análisis exploratorio descriptivo de los datos analizando la relación entre las variables explicativas con la variable dependiente, use representaciones visuales adecuadas, y comente cada gráfico y concluya. **(5 pts.)**

2. Divida los datos en 70% para entrenamiento y 30% para test, luego entrene tres modelos SVM cada uno con diferente función de kernel: linear, rbf y poly, variando para cada uno de ellos el valor de C para siete valores diferentes que deben ir desde 0.001 hasta 0.02. Determine cual de todos los modelos genera el mejor accuracy. **(5 pts.)**

3. Entrene un modelo de regresión lineal para predecir el nivel de dulzor de la manzana (Sweetness) usando como variables regresoras las otras características sin considerar la variable de Quality.

   Analice los residuos del modelo entrenado usando los datos de entrenamiento y verifique que distribuyan de acuerdo a la distribución normal con media cero y que su variabilidad sea homocedastica. Para esto se sugiere graficar los valores ajustados con sus residuos. **(8 pts.)**

4. Realice una evaluación del modelo de regresión lineal usando una métrica adecuada he interprete el efecto de las variables regresoras en la variable dependiente. Describa si el modelo es aceptable o no y cómo podemos mejorarlo. **(7 pts.)**