# Selección del Clasificador para el DataSet

## Introducción

El problema en cuestión es de aprendizaje supervisado, ya que el DataSet incluye una columna objetivo etiquetada (NObeyesdad) que clasifica a los individuos en niveles de obesidad. Esto nos lleva a seleccionar un modelo que pueda manejar múltiples clases y datos mixtos (numéricos y categóricos). Además, según las condiciones, el modelo no debe depender estrictamente del preprocesamiento.

## Selección del Clasificador

El modelo seleccionado es el Bosque Aleatorio (Random Forest), que cumple con los requisitos del problema y las especificaciones del enunciado. Este modelo fue elegido por las siguientes razones: :  
  
**1.** Capacidad para datos mixtos: Maneja tanto variables numéricas como categóricas sin requerir normalización o escalado. .  
**2.** Problemas multiclase: Es inherentemente capaz de clasificar múltiples categorías sin necesidad de adaptaciones adicionales. .  
**3.** Resistencia al ruido: Tolera la presencia de datos redundantes o correlacionados, algo común en este tipo de conjuntos de datos. .  
**4.** Flexibilidad en el preprocesamiento: Aunque la codificación de variables categóricas puede mejorar su desempeño, no es estrictamente necesaria para su funcionamiento.  
  
En comparación, modelos como las Máquinas de Soporte Vectorial (SVM) y las Redes Neuronales fueron descartados por su dependencia del preprocesamiento (como escalado de datos) y su complejidad respecto al volumen de datos.

## Justificación del Clasificador

El Bosque Aleatorio es ampliamente reconocido por su robustez y flexibilidad en problemas supervisados. Según Breiman (2001), este modelo utiliza un conjunto de árboles de decisión entrenados con datos aleatorios, lo que lo hace menos propenso al sobreajuste en comparación con un único árbol de decisión. Además, su capacidad para manejar problemas multiclase lo hace ideal para este conjunto de datos.  
  
Desde un punto de vista práctico, el Bosque Aleatorio es adecuado porque:  
- Proporciona métricas interpretables como la importancia de las características, lo que ayuda a entender el impacto de cada variable en la predicción.  
- No depende estrictamente del preprocesamiento, como lo confirma Bishop (2006), quien destaca que este modelo no requiere normalización de los datos de entrada, a diferencia de métodos basados en distancia como KNN o SVM.

En caso de ser no supervisado se utiliza el clasificador **K-means** es un algoritmo de agrupamiento (clustering) que tiene como objetivo dividir un conjunto de datos en un número específico de **K clusters** (grupos) basándose en las características de los datos. Cada cluster contiene puntos de datos similares entre sí y diferentes de los puntos en otros clusters. El algoritmo busca minimizar la **varianza intra-cluster** (la diferencia dentro de los puntos de un mismo cluster) y maximizar la **varianza inter-cluster** (la diferencia entre clusters).

## Conclusión

El Bosque Aleatorio es la mejor elección para este problema de clasificación multiclase supervisada. Su robustez, flexibilidad frente al preprocesamiento y su capacidad para manejar datos mixtos cumplen con las condiciones del enunciado y garantizan un desempeño óptimo en este contexto.

Para datos no supervisados, un clasificador comúnmente utilizado es el **K-means clustering**. Este algoritmo agrupa los datos en K clusters basándose en las características de los datos, sin necesidad de etiquetas. El número de clusters, K, debe especificarse antes de ejecutar el algoritmo.

## Referencias

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. DOI: https://doi.org/10.1023/A:1010933404324  
2. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. ISBN: 978-0387310732.