

# Algoritmos SVM y RF

---

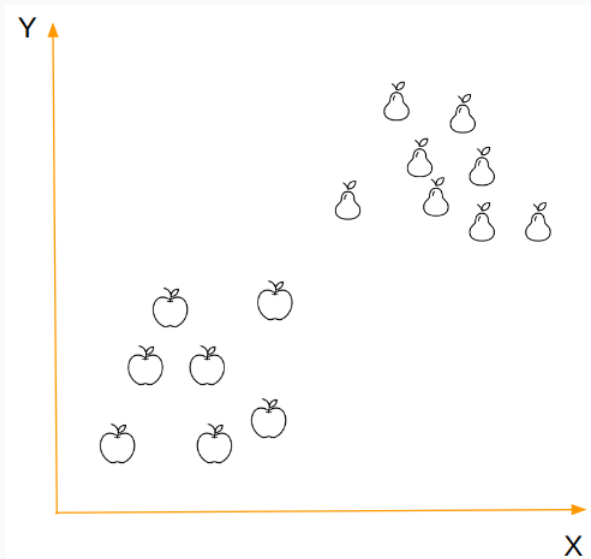
Blanca Vázquez

23 de octubre de 2023

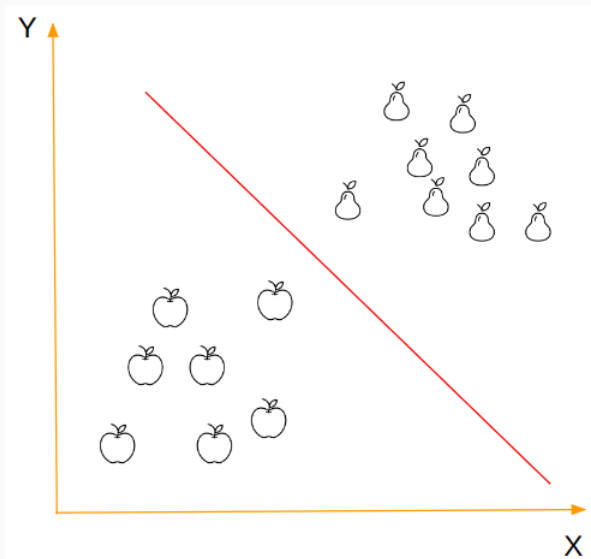
Las máquinas de soporte vectorial (SVM por las siglas en inglés de *Support Vector Machines*) son un conjunto de algoritmos de aprendizaje supervisado desarrolladas por Vladimir Vapnik en los años 90.

- Se utilizan en problemas de regresión y de clasificación.
- Es un clasificador discriminatorio definido por un hiperplano de separación.

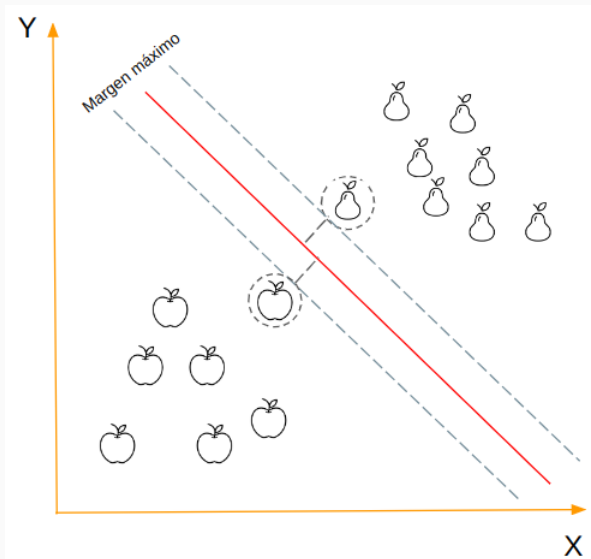
# Clasificación de datos



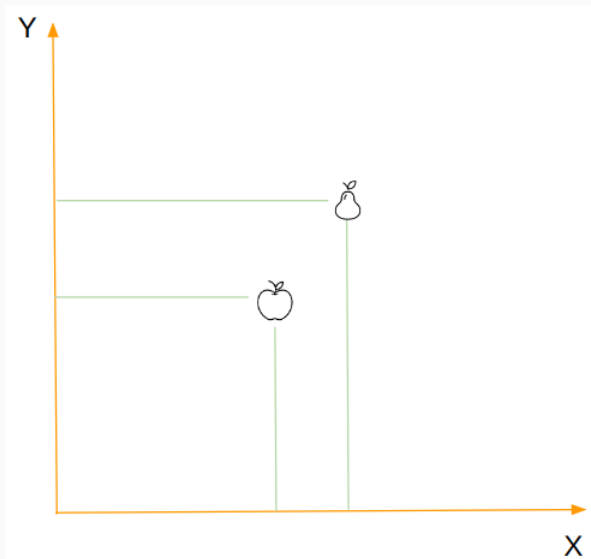
## Establecemos un hiperplano óptimo



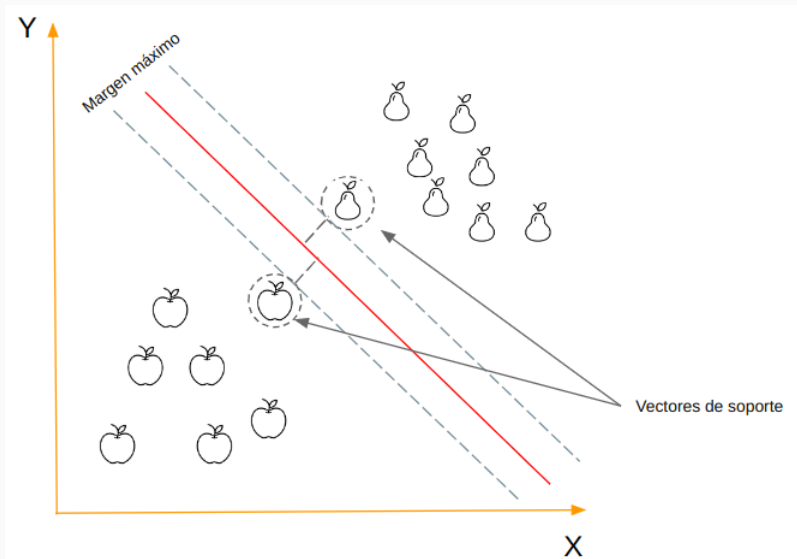
# Vectores de soporte



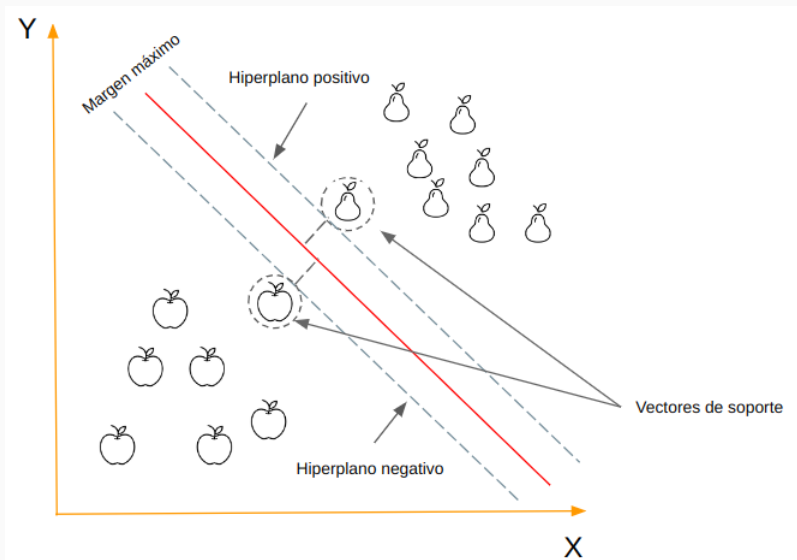
## Vectores de soporte



# Hiperplano positivo y negativo

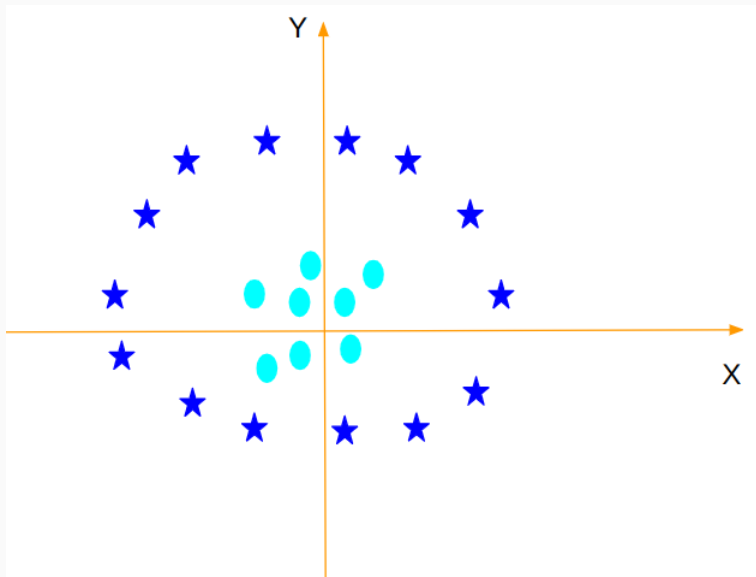


# Elementos en nuestro modelo SVM





¿Y si el problema no es lineal?

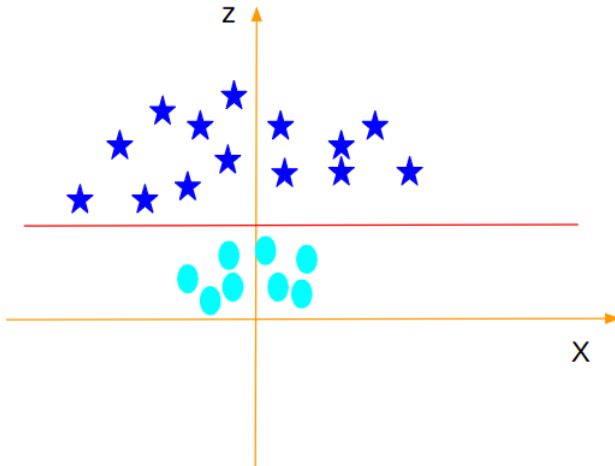


## ¿Y si el problema no es lineal?

- El algoritmo permite usar funciones Kernel (no lineales)
- Estas funciones trasladan los datos a un espacio donde el hiperplano tiene una solución lineal.
- Existen diferentes funciones kernel: lineal, de polinomios, de función base radial, sigmoide, etc.

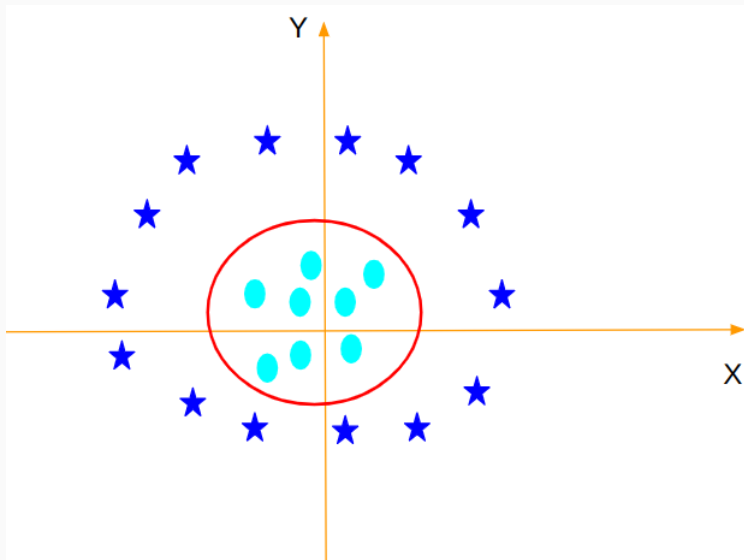
# Clasificación de datos

Agregamos un eje  $z$  en donde los datos se convirtieron en un problema lineal.



# Clasificación de datos

Una vez conseguida la solución se transforma al espacio original



## Ventajas del algoritmo de SVM

1. Funcionan muy bien cuando los datos son linealmente separables.
2. Para los datos que no son linealmente separables, podemos proyectar los datos a un espacio donde sean perfectamente o casi linealmente separables.

- El algoritmo de árboles aleatorios (RF, por las siglas en inglés de *Random Forest*) es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento.
- Los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado.

## ¿Cómo se construyen los árboles

1. Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables  $m$  (predictores) de forma que  $m < M$  (donde  $M =$  total de predictores).
2. Cada árbol crece hasta su máxima extensión.
3. Cada árbol contiene un subconjunto de datos aleatorios (seleccionados mediante *bootstrap*).
4. Las observaciones no consideradas en los árboles, se utilizan para validar el modelo.
5. Las salidas de todos los árboles se combinan en una salida final (ensamblado) que se obtiene mediante una regla.
6. La regla puede ser promedio: cuando las salidas son numéricas.
7. La regla puede ser conteo de votos: cuando las salidas son categóricas.

# Funcionamiento del algoritmo de árboles aleatorios

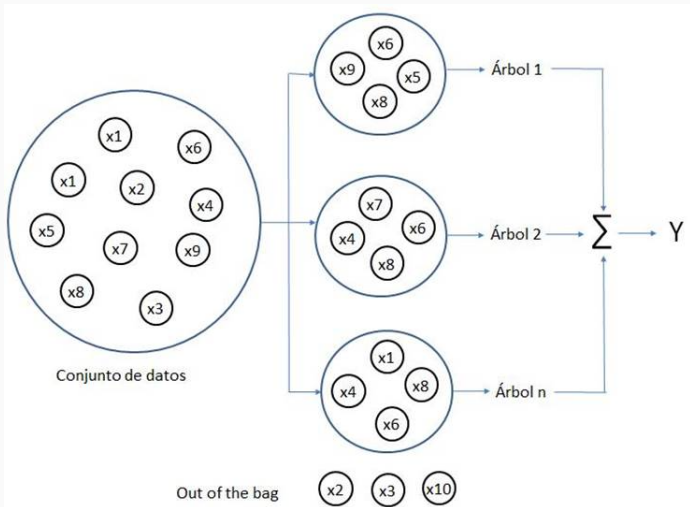


Imagen tomada de Espinosa-Zúñiga, 2020.



# Principales ventajas

- Pueden usarse para clasificación o regresión.
- Para clasificación: cada árbol vota por una clase y el resultado del modelo es la clase con mayor número de votos en todos los árboles.
- En la clasificación, cada nueva observación se presenta a cada uno de los árboles y se asigna a la clase más votada.
- Para regresión: el resultado del modelo es el promedio de las salidas de todos los árboles.
- Es un modelo simple de entrenar.

## Principales desventajas

- Pueden llegar a ser costosos.
- En grandes volúmenes de datos pueden requerir recursos de hardware importantes.
- Pueden llegar a sobreajustarse