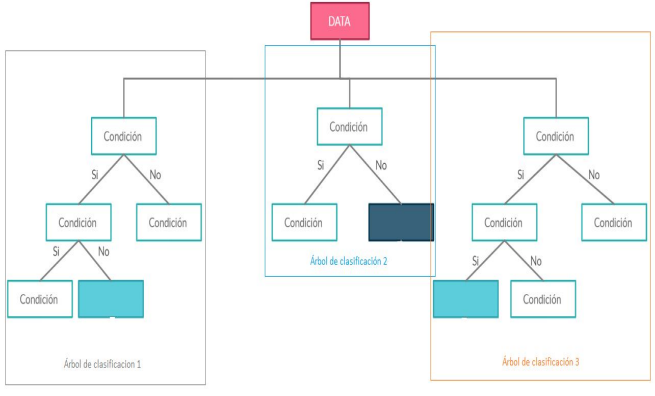
Distinción y Aplicación de Métodos Básicos de Clasificación

Rafael Martín R.

Métodos de Clasificación en Machine Learning

Explorando Árboles de Decisión, Random Forests y SVM

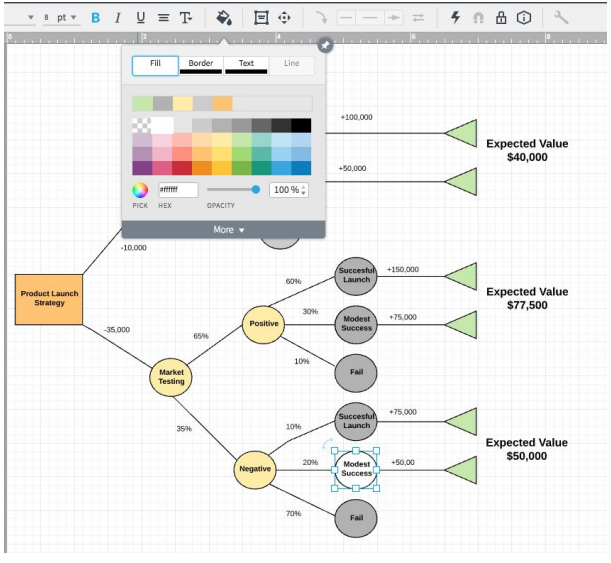
La clasificación en machine learning es fundamental para resolver problemas complejos de toma de decisiones. En esta presentación, exploraremos tres métodos poderosos: los Árboles de Decisión, que imitan el proceso de toma de decisiones humano; Random Forests, que aprovechan la sabiduría colectiva de múltiples árboles; y las Máquinas de Vectores de Soporte (SVM), que encuentran fronteras óptimas de separación entre clases. Cada método tiene sus fortalezas únicas y aplicaciones específicas en el mundo real.



Rafael Martín R.

Árboles de Decisión: Fundamentos

Los árboles de decisión son modelos de aprendizaje automático que imitan el proceso de toma de decisiones humano. Su estructura jerárquica consta de nodos de decisión que representan preguntas sobre las características de los datos, ramas que indican las posibles respuestas, y hojas que contienen las decisiones finales o predicciones. Esta estructura permite dividir sistemáticamente un problema complejo en subproblemas más simples, siguiendo un proceso lógico y transparente similar a cómo los humanos tomamos decisiones cotidianas.



Rafael Martín R.

Funcionamiento de Árboles de Decisión

División de Datos

El árbol evalúa cada característica y selecciona la mejor división posible basándose en la pureza de los subconjuntos resultantes, maximizando la ganancia de información.

Criterios de División

Se utilizan métricas como el Índice Gini o la Entropía para medir la impureza de los datos y determinar las divisiones óptimas en cada nodo del árbol.

Proceso de Clasificación

Los datos nuevos recorren el árbol desde la raíz, siguiendo las reglas de decisión en cada nodo hasta llegar a una hoja que determina la clasificación final.

Rafael Martín R.

Ventajas y Limitaciones de Árboles de Decisión

Ventajas Principales

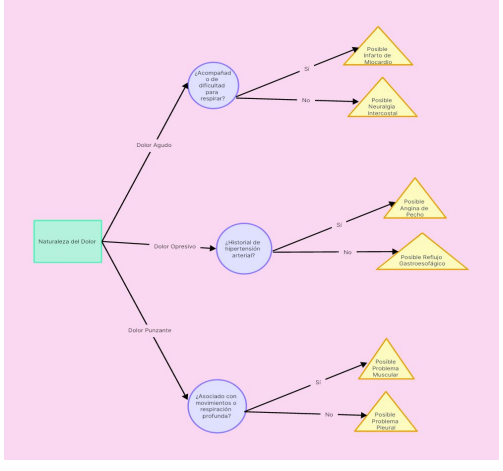
- Interpretación clara y transparente del proceso de decisión
- Visualización intuitiva mediante estructura de árbol jerárquico
- Manejo efectivo de datos numéricos y categóricos
- No requiere preprocesamiento extensivo de datos
- Capacidad para capturar relaciones no lineales naturalmente

Limitaciones Importantes

- Alta tendencia al sobreajuste, especialmente en árboles profundos
- Sensibilidad significativa a pequeñas variaciones en datos
- Puede generar estructuras complejas con datos ruidosos
- Dificultad para capturar relaciones lineales simples
- Inestabilidad en las predicciones con cambios en datos

Rafael Martín R.

Aplicaciones Prácticas de Árboles de Decisión



Diagnóstico Médico

Evaluación de Riesgo Crediticio

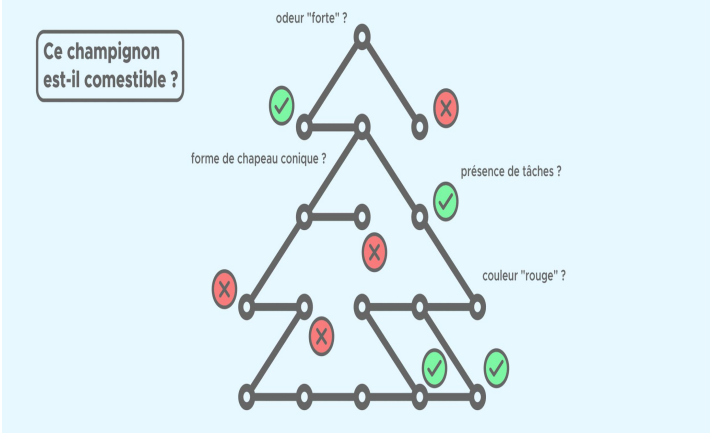
Sistemas de Recomendación

Rafael Martín R.

Random Forests: Mejorando los Árboles

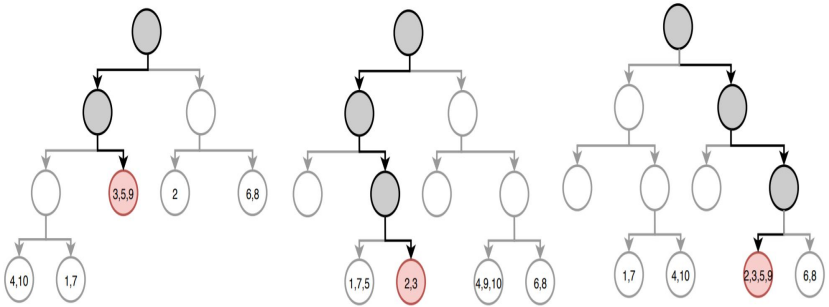
Combinando múltiples árboles para decisiones más precisas y robustas

Random Forests representa una evolución significativa de los árboles de decisión individuales, aplicando el principio de "sabiduría colectiva". Este método combina cientos o miles de árboles de decisión, cada uno entrenado con diferentes subconjuntos de datos y características. Al igual que un grupo diverso de expertos puede tomar mejores decisiones que un solo experto, Random Forests aprovecha la diversidad de múltiples árboles para generar predicciones más precisas y confiables.

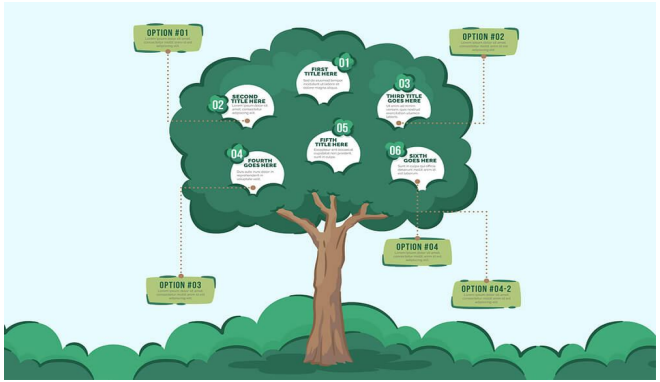


Rafael Martín R.

Funcionamiento de Random Forests



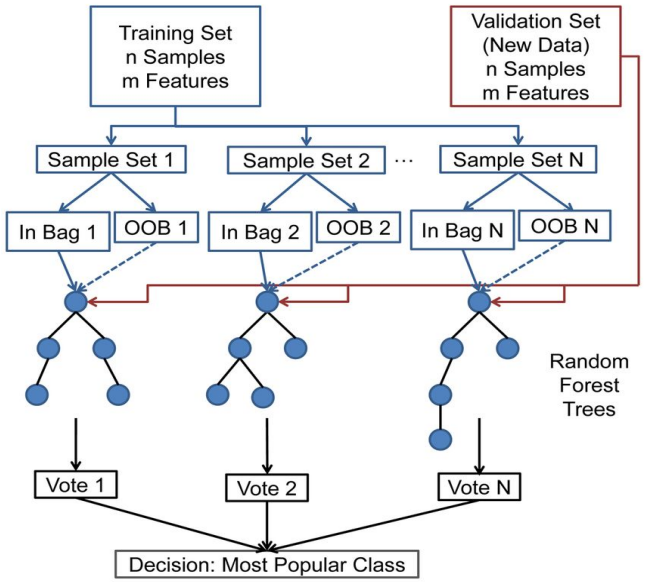
Proceso de Bagging



Combinación de Predicciones

Rafael Martín R.

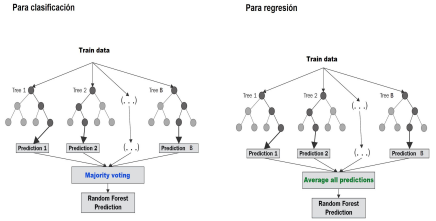
Ventajas de Random Forests



Superando limitaciones con el poder del conjunto

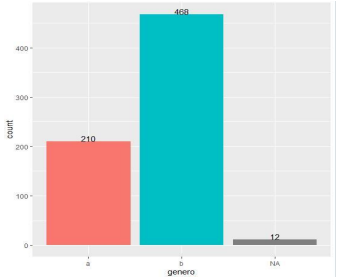
- Mayor precisión en predicciones gracias al promedio de múltiples árboles independientes
- Reducción significativa del sobreajuste mediante la diversificación de modelos
- Manejo robusto de valores atípicos y datos faltantes sin preprocesamiento extensivo
- Capacidad para manejar grandes conjuntos de datos con múltiples variables
- Estimación incorporada de la importancia de las variables para selección de características
- Menor necesidad de ajuste de hiperparámetros comparado con otros algoritmos avanzados

Aplicaciones de Random Forests



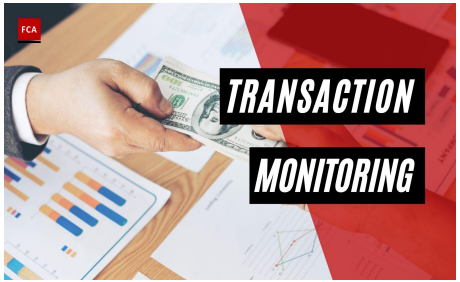
Predicción en Medicina

- Diagnóstico temprano de enfermedades cardíacas mediante análisis de múltiples factores
- Predicción de riesgo de diabetes basado en historial clínico y biomarcadores
- Identificación de patrones en imágenes médicas para detección de tumores



Análisis de Mercado y Finanzas

- Predicción de tendencias de mercado analizando múltiples variables económicas
- Segmentación avanzada de clientes basada en patrones de comportamiento
- Evaluación de riesgo crediticio considerando diversos factores financieros



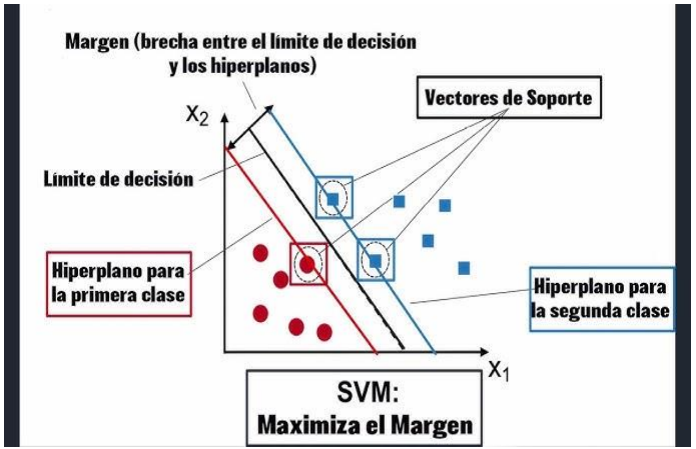
Detección de Fraudes

- Identificación de transacciones fraudulentas en tiempo real mediante patrones
- Detección de anomalías en comportamientos de usuarios y sistemas
- Prevención de fraudes en seguros mediante análisis de reclamaciones

Rafael Martín R.

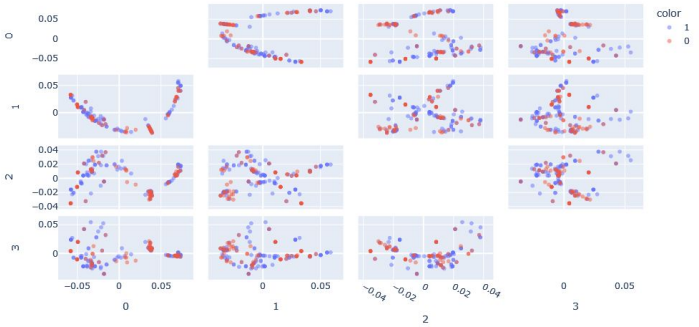
SVM: Conceptos Fundamentales

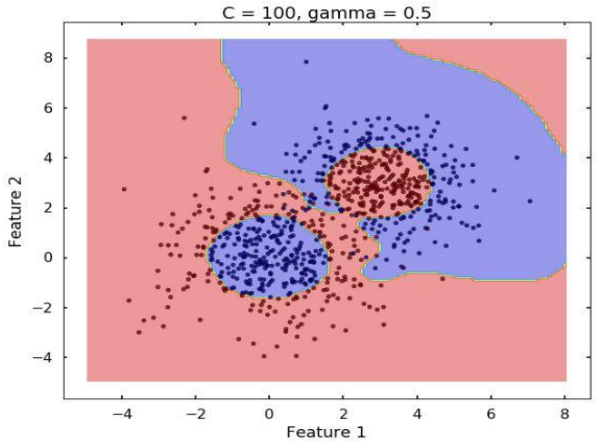
Las Máquinas de Vectores de Soporte (SVM) son algoritmos poderosos que buscan encontrar el hiperplano óptimo que mejor separa las clases en el espacio de características. Este hiperplano se determina maximizando el margen, que es la distancia entre el hiperplano y los puntos más cercanos de cada clase, llamados vectores de soporte. Estos vectores son cruciales ya que son los únicos puntos necesarios para definir el límite de decisión, lo que hace que SVM sea eficiente y robusto.



Rafael Martín R.

SVM: Más Allá de la Linearidad





Transformación del Espacio con Kernels

Problemas No Linealmente Separables

Rafael Martín R.

Ventajas y Desafíos de SVM

Ventajas Principales

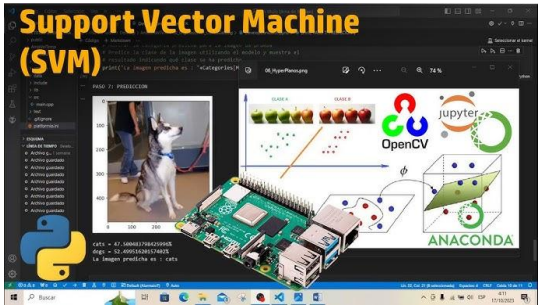
- Excelente rendimiento en espacios de alta dimensionalidad
- Sólido fundamento teórico y garantías matemáticas
- Eficaz en problemas de clasificación no lineales
- Alta precisión y capacidad de generalización
- Resistente al sobreajuste con parametrización adecuada

Desafíos Importantes

- Alto costo computacional para grandes conjuntos de datos
- Compleja selección de parámetros y función kernel
- Sensibilidad a la calidad de los datos
- Difícil interpretación del modelo resultante
- Requiere preprocesamiento cuidadoso de los datos

Rafael Martín R.

Aplicaciones Prácticas de SVM



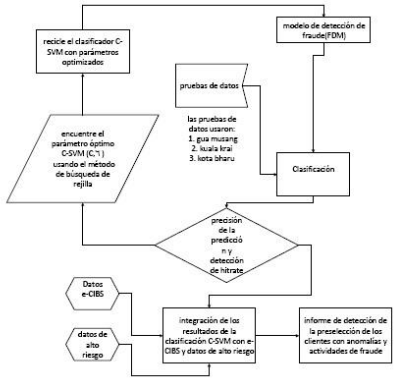
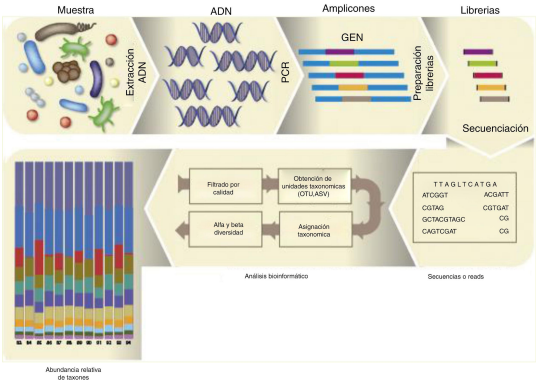


Figura 6. Diagrama de flujo del motor de validación FDM para la detección de presuntos clientes (clientes con anomalías y actividades de fraude)

Fuente: Nagi *et al.* (2010).



Reconocimiento de Imágenes

Clasificación de Textos

Aplicaciones en Bioinformática

Rafael Martín R.

Comparativa de Métodos

Características y Complejidad

- Árboles: Simple y rápido, interpretabilidad alta, propenso a sobreajuste
- Random Forests: Complejidad media, mejor rendimiento, requiere más memoria
- SVM: Alta complejidad computacional, excelente en datos de alta dimensión

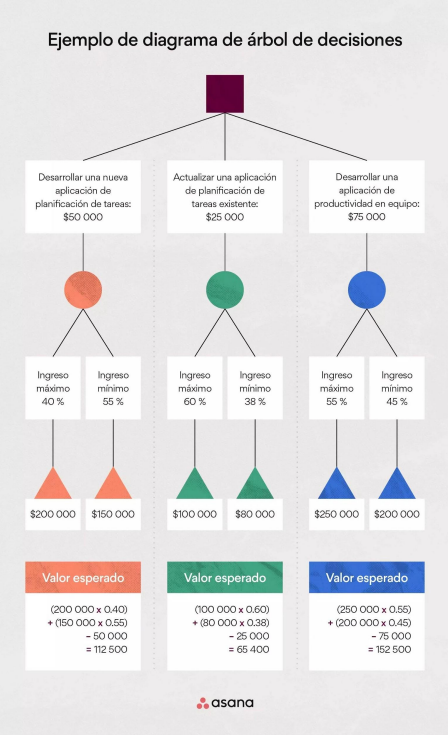
Escalabilidad y Recursos

- Árboles: Excelente escalabilidad, recursos computacionales mínimos requeridos
- Random Forests: Escalabilidad moderada, paralelizable para mejor rendimiento
- SVM: Escalabilidad limitada, alto consumo de memoria y procesamiento

Tipos de Datos y Aplicabilidad

- Árboles: Ideal para datos mixtos, categóricos y numéricos
- Random Forests: Versátil en diversos tipos de datos, maneja valores faltantes
- SVM: Óptimo para datos numéricos, requiere preprocesamiento específico

Guía de Selección de Método



- Tamaño del Dataset: Árboles para conjuntos pequeños, Random Forests y SVM para medianos y grandes
- Tipo de Datos: Árboles y Random Forests para datos mixtos, SVM excelente para datos numéricos de alta dimensión
- Interpretabilidad: Árboles de Decisión ofrecen la mayor claridad, Random Forests balance, SVM menos interpretables
- Recursos Computacionales: Árboles son ligeros, Random Forests moderados, SVM demandan más recursos
- Precisión Requerida: SVM y Random Forests para alta precisión, árboles para precisión moderada
- Datos Ruidosos: Random Forests manejan mejor el ruido, seguido por SVM, árboles son más sensibles
- Tiempo de Entrenamiento: Árboles son rápidos, Random Forests moderados, SVM pueden ser lentos

Rafael Martín R.