Instituto Tecnológico de Tijuana

Subdirección Académica

Departamento de Sistemas y computación

Asignatura

Datos Masivos

Docente

JOSE CHRISTIAN ROMERO HERNANDEZ

Random Forest Classifier

Integrantes

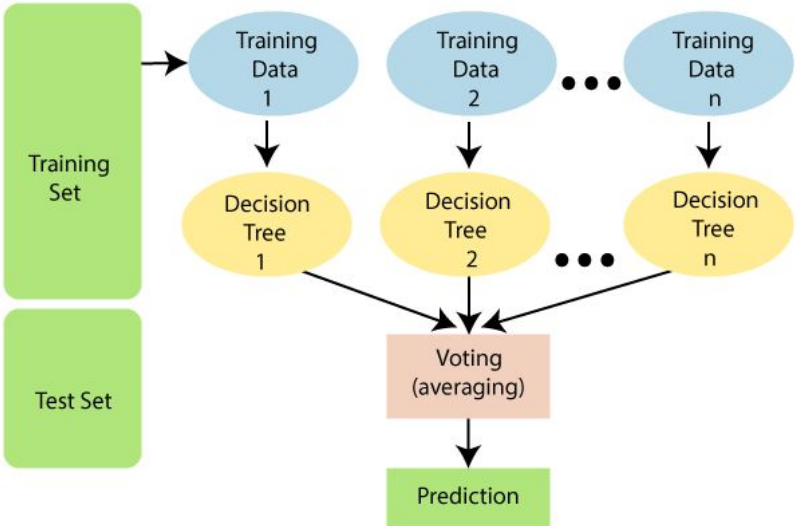
Garcia Torres Cristopher Arael	#17210892
Rivas Padilla Giselle	#17210902
Regalado Lopez Edgar Eduardo	#18212254
Guzman Morales Gregorio Manuel	#C17210760

Random Forest

Random Forest o Bosque aleatorio, es un algoritmo popular de aprendizaje automático que pertenece a la técnica de aprendizaje supervisado. Se puede utilizar tanto para problemas de clasificación como de regresión en ML. Se basa en el concepto de **aprendizaje en conjunto,** que es un proceso de *combinación de múltiples clasificadores para resolver un problema complejo y mejorar el rendimiento del modelo.*

Como su nombre indica, ***"Random Forest es un clasificador que contiene una serie de árboles de decisión en varios subconjuntos del conjunto de datos dado y toma el promedio para mejorar la precisión predictiva de ese conjunto de datos".*** En lugar de depender de un árbol de decisión, el bosque aleatorio toma la predicción de cada árbol y se basa en los votos mayoritarios de las predicciones, y predice el resultado final.

El mayor número de árboles en el bosque conduce a una mayor precisión y evita el problema del sobreajuste.

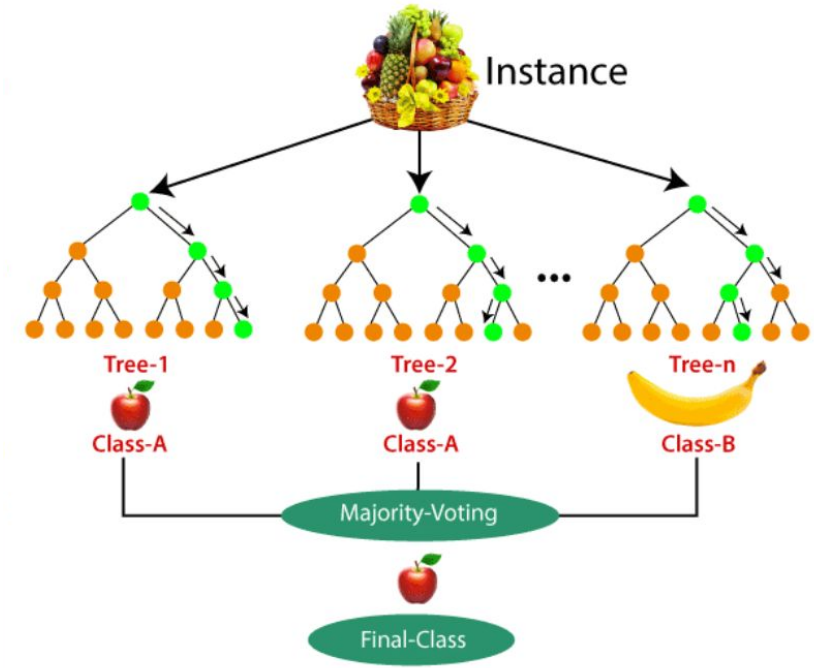


PREDICCIÓN

Para hacer una predicción en una nueva instancia, un bosque aleatorio debe agregar las predicciones de su conjunto de árboles de decisión. Esta agregación se realiza de manera diferente para la clasificación y la regresión.

Clasificación: Voto mayoritario. La predicción de cada árbol se cuenta como un voto para una clase. Se predice que la etiqueta será la clase que reciba más votos.

Regresión: Promediar. Cada árbol predice un valor real. Se predice que la etiqueta es el promedio de las predicciones del árbol.



¿Por qué usar random forest?

- Lleva menos tiempo de entrenamiento en comparación con otros algoritmos.
- Predice la salida con alta precisión, incluso para el gran conjunto de datos que ejecuta de manera eficiente.
- También puede mantener la precisión cuando falta una gran proporción de datos.

Aplicaciones de random forest

1. **Banca:** El sector bancario utiliza principalmente este algoritmo para la identificación del riesgo de préstamo.
2. **Medicina:** Con la ayuda de este algoritmo, se pueden identificar las tendencias de la enfermedad y los riesgos de la enfermedad.
3. **Uso del suelo:** Podemos identificar las áreas de uso similar de la tierra mediante este algoritmo.
4. **Marketing:** Las tendencias de marketing se pueden identificar utilizando este algoritmo.

VENTAJAS

- Random Forest es capaz de realizar tareas de clasificación y regresión.
- Es capaz de manejar grandes conjuntos de datos con alta dimensionalidad.
- Mejora la precisión del modelo y evita el problema de sobreajuste.

DESVENTAJAS

- Aunque el bosque aleatorio se puede usar tanto para tareas de clasificación como de regresión, no es más adecuado para tareas de regresión.

Video complementario

https://www.youtube.com/watch?v=jlJ4uKS9D5A

Referencias

Machine Learning Random Forest Algorithm - Javatpoint. (z.d.). Www.Javatpoint.Com. Geraadpleegd op 30 april 2022, van https://www.javatpoint.com/machine-learning-random-forest-algorithm

Ensembles - RDD-based API - Spark 2.4.8 Documentation. (z.d.). spark.apache. Geraadpleegd op 30 april 2022, van https://spark.apache.org/docs/2.4.8/mllib-ensembles.html#random-forests