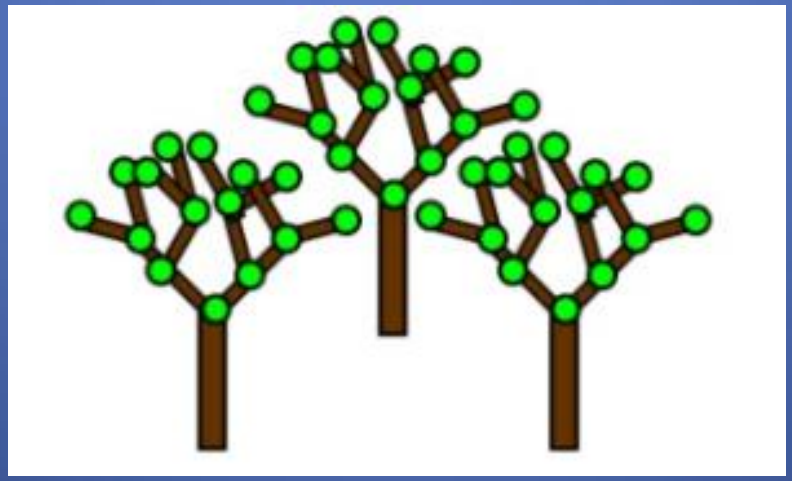
DIPLOMATURA EN PYTHON APLICADO A LA CIENCIA DE DATOS

Random Forest



Agenda

- Contenido Teórico
- Ventajas – Desventajas de Random Forest
- Uso del Paquete en Python

Random Forest – Contenido Teórico

- **Definición y Objetivo**:

  Random Forest consta de una gran cantidad de árboles de decisión individuales que operan como un conjunto.

  Cada árbol individual en el bosque aleatorio arroja una predicción de clase y la clase con más votos se convierte en la predicción del modelo.

  El objetivo es predecir con alta precisión y reducir la varianza utilizando bagging.-

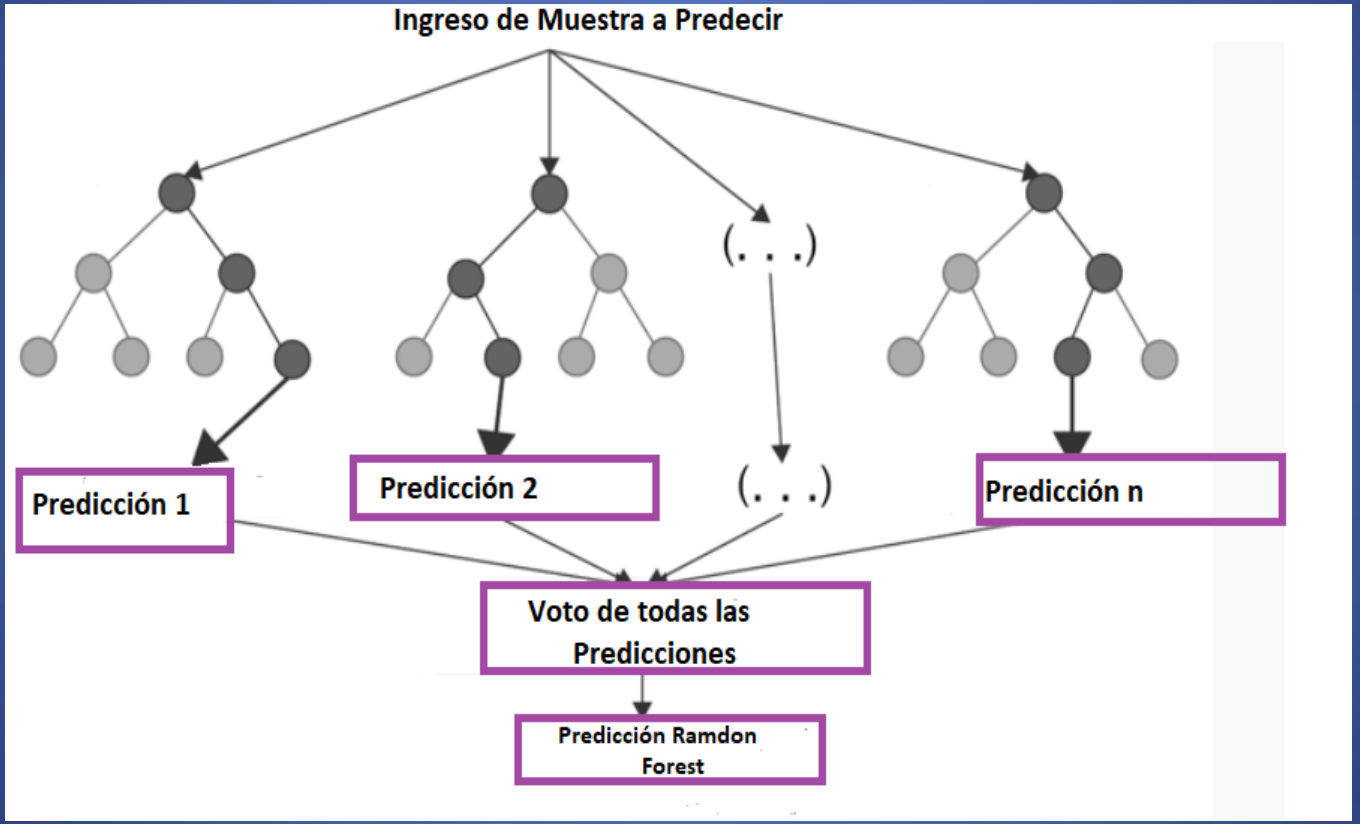
Tipo de Aprendizaje: Supervisado (Datos Etiquetados) .

- ¿Dónde Aplicar Árboles de Decisión?
  - Financiero – Detectar clientes con más posibilidades de pagar deuda
  - Medicina – Identificar enfermedades
  - Comercio electrónico – si a un cliente le va a gustar un determinado producto

Random Forest – ¿Cómo Funciona?

1. Dado un dataset inicial de entrada
2. Random Forest va a generar una gran cantidad de árboles de decisión individuales que operan como un conjunto.
3. Cada árbol individual en el bosque aleatorio arroja una predicción de clase y la clase con más votos se convierte en la predicción del modelo, como lo vemos en la figura a continuación.
4. Cada árbol va a crecer hasta su máxima extensión posible y **NO hay proceso de poda**.
5. Nuevas instancias se predicen a partir de la agregación de las predicciones de los x árboles (i.e., mayoría de votos para clasificación, promedio para regresión)

Random Forest – Contenido Teórico



Random Forest – Ventajas - Desventajas

- Ventajas
  - ✓ Es uno de los algoritmos de aprendizaje más precisos disponibles. Para muchos conjuntos de datos, produce un clasificador de alta precisión.
  - ✓ Funciona de manera eficiente en grandes bases de datos.
  - ✓ Puede manejar miles de variables de entrada sin eliminarlas.
  - ✓ Proporciona estimaciones de las variables que son importantes en la clasificación.
  - ✓ Genera una estimación interna no sesgada del error de generalización a medida que avanza la construcción del bosque.
  - ✓ Tiene un método eficaz para estimar los datos faltantes y mantiene la precisión cuando falta una gran proporción de los datos.
  - ✓ Puede resolver problemas de clasificación y regresión

Random Forest – Ventajas - Desventajas

- Desventajas
  - ✓ Se ha observado que los bosques aleatorios sobre ajustan para algunos conjuntos de datos con tareas de clasificación / regresión ruidosas.
  - ✓ Para los datos que incluyen variables categóricas con diferente número de niveles, los bosques aleatorios están sesgados a favor de aquellos atributos con más niveles. Por lo tanto, los puntajes de importancia variable del bosque aleatorio no son confiables para este tipo de datos.

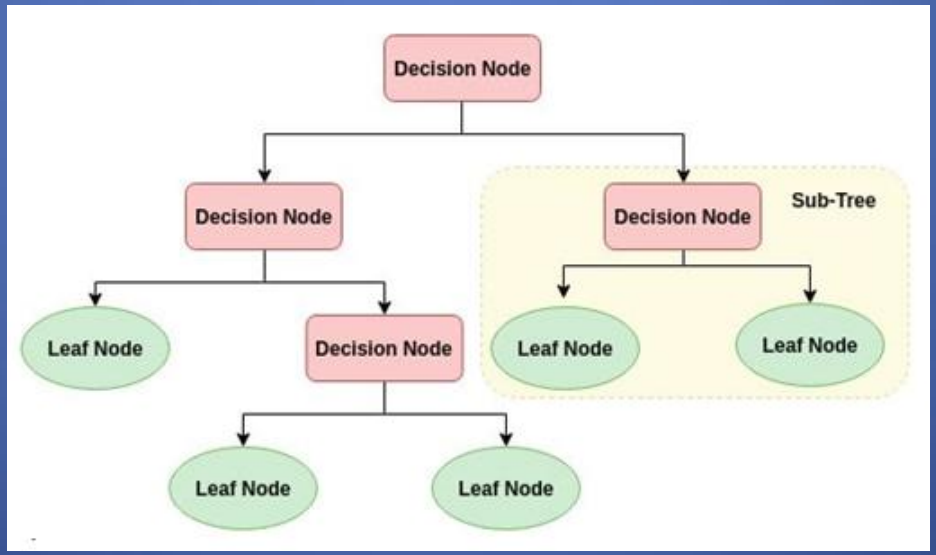
Random Forest - Terminología

Con respecto a la terminología de cada subárbol se aplican los mismos conceptos que los árboles de decisión

Nodo Raíz

División

Nodo de hoja



Rama / Subárbol

Poda

Nodo padre e hijo

Random Forest

Random Forest - Bootstraping

Bootstrap

¿Qué es?

Técnica

Se centra en el re muestreo de datos dentro de una muestra aleatoria o al azar

Objetivo

Aproximar la varianza gracias a la realización de re muestreos aleatorios de la muestra inicial y no de la población

Característica

Supone un re muestreo posterior para poder obtener expresiones cerradas y solucionar la complejidad matemática de estas operaciones.-

Random Forest

Parámetros de la Función - Clasificación

class sklearn.ensemble.**RandomForestClassifier**(***n_estimators=100***, **, criterion='gini', max_depth=None, min_samples_split=2*, ***min_samples_leaf=1***, *min_weight_fraction_leaf=0.0*, ***max_features='auto'***, *max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None*, ***bootstrap=True***, ***oob_score=False***, ***n_jobs=None***, *random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None*)

Parámetros de la Función - Regresión

class sklearn.ensemble.**RandomForestRegressor**(***n_estimators=100***, **, criterion='mse', max_depth=None, min_samples_split=2,* ***min_samples_leaf=1***, *min_weight_fraction_leaf=0.0*, ***max_features='auto'***, *max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None*, ***bootstrap=True***, ***oob_score=False***, ***n_jobs=None***, *random_state=None, verbose=0, warm_start=False, ccp_alpha=0.0, max_samples=None)*

Random Forest - Ejemplo

- Se tiene una lista de candidatos a ingresar a un puesto vacante. Para ello se solicita examen de ingreso, psicofísico, experiencia laboral y edad.
- Los datos correspondientes se ingresarán al dataset y la etiqueta será – Admitido = 2, Lista de espera = 1 y no admitido = 0

DIPLOMATURA EN PYTHON APLICADO A LA CIENCIA DE DATOS

Unidad – Random Forest

Fin

Muchas Gracias!