**Escuela de Estudios de Postgrado**

**Facultad de Ingeniería**

**Universidad San Carlos de Guatemala**

**Maestría en Ingeniería para la Industria con**

**Especialización en Ciencias de la Computación.**

**Minería de Datos**

**Ing. Ms. Kevin Lajpop**

**PROYECTO 2**

<https://github.com/obduliogonzalez/proyecto2_mineria_datos>

**Néstor Obdulio González López**

**Carné: 999013150**

**1. Entender la estructura del árbol**

Un árbol de decisión utiliza divisiones (o nodos) basadas en los valores de las variables predictoras para llegar a una decisión final. Estos nodos son puntos de decisión donde se selecciona el mejor predictor y su umbral de división en función de una métrica como Gini o Entropía.

**Elementos importantes:**

* **Raíz (Root)**: Es el nodo inicial que representa toda la muestra de datos. La primera división se hace usando la variable más influyente.
* **Nodos intermedios**: Representan divisiones adicionales basadas en otras variables predictoras.
* **Hojas (Leaves)**: Son los nodos terminales del árbol que representan una predicción final.

**2. Interpretación de los modelos específicos**

**Caso 1: Predicción de Estado de Ebriedad**

**Variables**: SEXO, EDAD, EST\_CONYUGAL, MES, AREA\_GEOGRAFICA

* **Significado**: Este modelo busca determinar si una persona se encuentra en estado de ebriedad basándose en las características demográficas y contexto del evento.
* **Interpretación del árbol**:
  + Por ejemplo, si EST\_CONYUGAL tiene una división en <7 y >=7, indica que el estado conyugal es una variable relevante para predecir el estado de ebriedad.
  + Si en un nodo, la proporción de una clase (ebriedad sí/no) es muy alta, podemos inferir que las características del nodo son fuertes indicadores de esa clase.

**Caso 2: Predicción de Área Geográfica**

**Variables**: EDAD, ETNIA, ESCOLARIDAD, EST\_CONYUGAL

* **Significado**: Determina el área geográfica probable de una persona basada en su perfil demográfico y educativo.
* **Interpretación del árbol**:
  + Las divisiones iniciales, como EDAD<30, pueden sugerir que la edad es una variable clave para segmentar áreas geográficas.
  + Si ETNIA aparece en nodos superiores, implica que el origen étnico tiene un fuerte impacto en la clasificación geográfica.

**Caso 3: Predicción de Condición de Alfabetismo**

**Variables**: SEXO, EDAD, EST\_CONYUGAL, AREA\_GEOGRAFICA

* **Significado**: Predice si una persona sabe leer y escribir, con base en su contexto personal y demográfico.
* **Interpretación del árbol**:
  + Si EDAD<18 aparece como una división inicial, esto indicaría que los jóvenes tienen una alta probabilidad de alfabetismo.
  + La combinación de variables como AREA\_GEOGRAFICA y EST\_CONYUGAL puede destacar patrones específicos para ciertas áreas o grupos demográficos.

**Caso 4: Predicción del Nivel de Escolaridad**

**Variables**: EDAD, EST\_CONYUGAL, AREA\_GEOGRAFICA, ETNIA

* **Significado**: Estima el nivel de escolaridad de una persona basándose en su contexto social y cultural.
* **Interpretación del árbol**:
  + Si ETNIA tiene un impacto fuerte, sugiere que ciertos grupos étnicos tienen un acceso diferencial a la educación.
  + Una división como EDAD>=25 podría indicar que la educación terciaria se alcanza generalmente después de los 25 años.

**3. Evaluación de la Relevancia de las Predicciones**

Para interpretar las predicciones realizadas por el modelo en casos manuales:

* **Probabilidades asignadas**: Los modelos de árboles pueden proporcionar probabilidades para cada clase en los nodos terminales. Estas probabilidades indican la confianza del modelo en la predicción.
* **Errores de clasificación**: Al comparar las predicciones del modelo con los datos reales, se puede determinar qué tan bien el modelo está capturando los patrones.

Ejemplo de interpretación:

* Si el modelo predice correctamente que una persona con EDAD=30, ETNIA=2, ESCOLARIDAD=3 y EST\_CONYUGAL=1 tiene AREA\_GEOGRAFICA=2, esto sugiere que las características del caso se alinean bien con los patrones de datos históricos.

**4. Evaluar la calidad y utilidad del modelo**

* **Precisión**: Proporción de predicciones correctas.
* **Relevancia de las variables**: Identificar cuáles fueron más importantes para el modelo (esto se puede obtener con funciones específicas como varImp).
* **Aplicabilidad práctica**: Considerar si las predicciones coinciden con el contexto o el propósito para el cual se diseñó el modelo.

**Conclusión**

Cada árbol proporciona una estructura visual y decisiones basadas en datos que deben contextualizarse para entender su significado. Analizar los nodos y la importancia de las variables permite interpretar si el modelo está alineado con los objetivos del análisis.

**Paso 3: Interpretación y análisis**

Para cada modelo, puedes interpretar los resultados de la siguiente manera:

1. **Resumen del Modelo:**
   * El print(modelo\_rf\_ebriedad) o print(modelo\_rf\_area\_geo) te dará un resumen general del modelo entrenado, que incluye información sobre los errores de clasificación, la precisión, entre otros.
2. **Importancia de las Variables:**
   * La función importance(modelo\_rf\_ebriedad) te mostrará qué tan importantes son las variables predictoras en la clasificación. Las variables con valores más altos en importancia son las que tienen más influencia sobre las predicciones.
3. **Predicciones:**
   * Las predicciones del modelo son el resultado de aplicar el modelo entrenado sobre el conjunto de datos de entrada (multas). Puedes revisar head(predicciones\_rf\_ebriedad) o head(predicciones\_rf\_area\_geo) para ver las primeras predicciones realizadas.

**Conclusión:**

* El algoritmo **Random Forest** es robusto y proporciona predicciones más estables que un solo árbol de decisión. Utilizando múltiples árboles, el modelo es capaz de generalizar mejor los patrones en los datos.
* **La importancia de las variables** te permite comprender cuáles son los factores más influyentes en la predicción.

Si necesitas realizar más análisis o ajustes, puedes modificar parámetros como ntree (número de árboles) o mtry (número de variables por árbol) para mejorar el rendimiento de los modelos.

¿Te gustaría realizar alguna otra predicción o necesitas más detalles sobre cómo mejorar los modelos?