Técnicas de Inteligencia Artificial – Actividad 1 **Nombre:** Shirley Claudette Martínez Cerrato

Las conclusiones de este informe se han hecho en base al archivo jupyter notebook adjunto.

Descripción del Dataset

Atributos de Entrada

- Número de Atributos de Entrada: 6
- Significado:

Atributo	Significado	Tipo
Buying	Se refiere a la frecuencia con la cuál el auto es comprado. Sus distintos valores son: low, med high y vhigh	Categórico Ordinal
Maintenance	Se refiere al nivel de mantenimiento que el carro necesita para que pueda funcionar correctamente. Sus distintos valores son: low, med high y vhigh	Categórico Ordinal
Doors	Se refiere al número de puertas que posee el carro. Sus distintos valores son: 2,3,4y 5more	Categórico Ordinal
Person	Hace referencia al número de personas que caben dentro del carro. Sus distintos valores son: 2,4 y more.	Categórico Ordinal
lug_boot	Significa el tamaño que posee el baúl del carro. Sus distintos valores son: small, med y big	Categórico Ordinal
safety	Se refiere al nivel de seguridad que provee el carro. Sus distintos valores son: low, med, high	Categórico Ordinal

Atributo de Salida

- Número de Clases: 4
- Número de Instancias que pertenecen a cada clase:

Clase	Núm. Instancias	Significado Clase
acc	390	El carro es aceptable
good	75	El carro se considera bueno
unacc	1215	El carro es inaceptable
vgood	70	El carro se considera muy bueno

- Número total de instancias: 1750
- ¿Existen atributos con valores desconocidos? = todos los valores del dataset son conocidos

Estadísticas de los datos

Muestra de los resultados obtenidos

A continuación, se muestra una tabla resumen la cuál contiene información general sobre los valores de datos dentro del dataset

	Buying	Maintenance	Doors	Person	lug_boot	safety	class

count	1750	1750	1750	1750	1750	1750	1750
unique	4	4	4	3	3	3	4
top	vhigh	low	2	4	big	high	unacc
freq	443	447	444	587	585	590	1215

También se muestra una visualización de la distribución de los datos

Distribución de Frecuencias – Valores de Atributos de Entrada

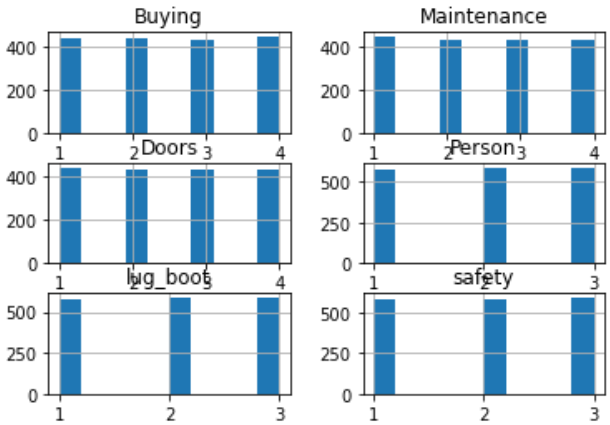
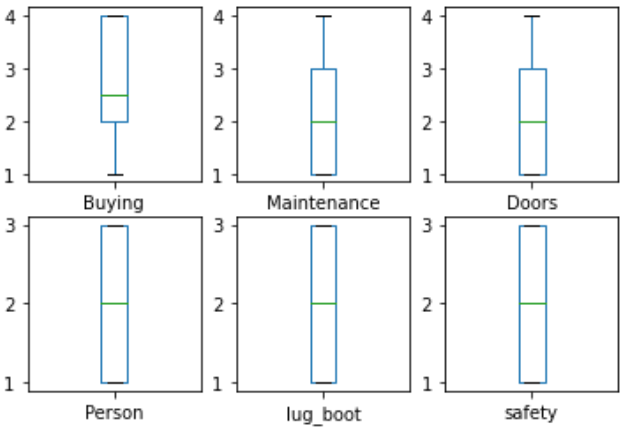


Gráfico de Cajas – Atributos de Entrada



Conclusiones de los datos Obtenidos

En base a la información anterior, podemos concluir que el conjunto de datos posee una distribución muy uniforme de sus valores para todos los atributos de entrada, tomando en cuenta que todos los atributos de entrada son de tipo categórico ordinal, la media de valor que en términos cualitativos sería equivalente a características ni muy buenas, ni muy malas encontradas en un automóvil.

Descripción de los Modelos Predictivos

Descripción de los modelos Escogidos

1. **Modelo Predictivo 1 – CART:** Este modelo fue escogido por su potente procesamiento para trabajar con variables cualitativas, al ser un árbol binario, realiza una predicción muy precisa al momento de clasificar las instancias, pero esto también implica que el conjunto de reglas para predecir cada clasificación será mayor.
2. **Modelo Predictivo 2 – Random Forest:** Este es un modelo que fue elegido por combinar el poder de la clasificación de árboles de decisión con múltiples iteraciones, lo cual lo hace un método muy preciso de clasificación, al realizar más iteraciones que el modelo anterior también requiere de un mayor tiempo para ser generado

Descripción de Parámetros Importantes

A criterio personal, listo los siguientes parámetros como importantes, ya que determinarán en gran media el nivel de precisión con el cuál los modelos generados clasifiquen futuras instancias.

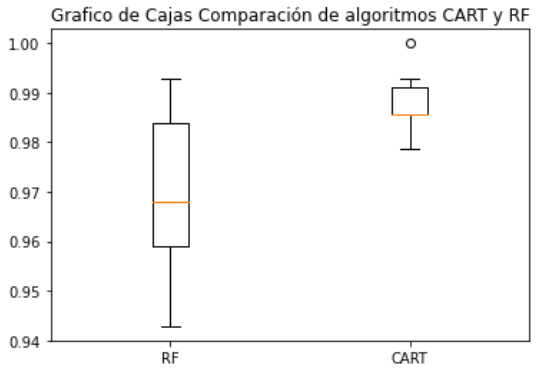
- **Valores de entrada de entrenamiento:** (X_train) constituye los datos de entrada para entrenar los modelos escogidos
- **Valores de salida de entrenamiento:** (Y_train) Este parámetro es el que dictará a los modelos escogidos en que clase clasificar a los datos de entrenamiento.
- **Valores de entrada de validación:** (X_validation) Este parámetro constituye el conjunto de datos de entrada que servirán para validar los modelos deseados
- **Valores de salida de validación:** (Y_validation) Dicho parámetro servirá para conocer si las predicciones efectuadas con los datos de validación son correctas.
- **Número de estimaciones del árbol aleatorio:** (n_estimators) Esta variable determina la profundidad de nuestro árbol, a mayor profundidad mayor precisión y mayor coste de tiempo.

Comparación entre modelos

Para comparar la efectividad y precisión de cada modelo se utilizó la técnica de validación cruzada con n_splits = 10, con una partición de los datos de 80% para datos de entrenamiento y 20% para datos de validación obteniendo los siguientes resultados:

Modelo	Media de Precisión Validación Cruzada
CART	0.987857
Random Forest	0.970000

De los resultados anteriores también se obtuvo la siguiente gráfica



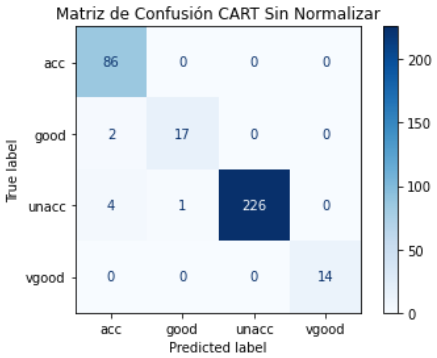
Como se puede observar en el gráfico de cajas, el algoritmo CART presenta una mayor precisión frente al algoritmo de Random Forest, al mismo tiempo también presenta un menor rango de dispersión en cuanto a la presión de las predicciones. En otras palabras, el algoritmo de CART es más preciso que el algoritmo RF para el conjunto de datos dado.

Desglose Modelo Predictivo 1 CART

Resultados Obtenidos

	Clases			
Indicador	unacc	acc	good	vgood
TPR	0.97835	1.0	0.89474	1.0
FPR	0.0	0.02273	0.00302	0.0

Matriz de Confusión Modelo Predictivo 1 CART



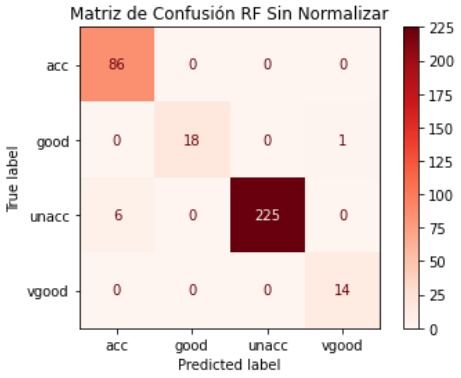
Descripción de Resultados Modelo Predictivo 1 CART

Como se puede observar en la matriz de confusión y a través de los diferentes valores de TPR y FPR para cada una de las clases, el modelo predictivo se comporta de forma muy precisa al momento de realizar las clasificaciones

Desglose Modelo Predictivo 2 Random Forest

	Clases			
Indicador	unacc	acc	good	vgood
TPR	0.97403	1.0	0.94737	1.0
FPR	0.0	0.02273	0.0	0.00298

Matriz de Confusión Modelo Predictivo 2 RF



Descripción de Resultados Modelo Predictivo 2 Random Forest

Se puede hacer una comparación entre los resultados obtenidos entre el modelo anterior y el modelo actual y concluir que ambos modelos realizan clasificaciones de forma muy precisa, pero el modelo creado con el algoritmo CART es un poco más efectivo al momento de realizar predicciones.

Conclusiones y Mejoras propuestas

Ambos modelos son muy precisos al momento de realizar una clasificación, el algoritmo CART obtuvo un porcentaje mayor pero no significativo con respecto al algoritmo de RF, esto se debe a que este algoritmo es mejor para clasificar datos con una distribución más homogénea, de hecho, tal como se menciona en el material de la clase, el algoritmo RF se desempeñará mejor con un conjunto de datos más inestable. Como mejoras propuestas, se propone entrenar los modelos con diferentes porcentajes de conjunto de datos de entrada y validación, observar como se comportan los modelos y en base a este indicador seleccionar las métricas de partición del dataset más propicias. Para el caso del algoritmo de Random Forest, se recomienda hacer pruebas para distintos valores de n_estimators y observar con cuál valor realiza predicciones más precisas.