unir LA UNIVERSIDAD EN INTERNET

Informe Actividad 1 – Técnicas de Inteligencia Artificial

Presentado Por: Henry Alejandro Gerena Ricardo

Fecha 09/12/2025

1. Descripción del problema

Nos enfrentamos a una tarea de clasificación y queremos predecir la aceptabilidad de un automóvil (clase objetivo) a partir de características como el precio de compra, el precio de mantenimiento, el número de puertas, la capacidad de asientos, el tamaño del maletero y la seguridad. La clase objetivo es nominal porque puede tomar cuatro valores posibles: vgood (muy bueno), good (bueno), acc (aceptable) y unacc (inaceptable), y los datos se obtuvieron del Repositorio de Aprendizaje Automático de la UCI.

El conjunto de datos es el conjunto de datos Car Evaluation (archivo Laboratorio_dataset_car.csv) porque se trata de un problema de aprendizaje supervisado y clasificación multiclase, por lo que la clase objetivo tiene más de dos valores posibles. Este tipo de modelo podría ser útil para decisiones de compra, motores de recomendación y estudios de políticas de productos dentro del sector automovilístico, por lo que tiene diversas aplicaciones prácticas.

2. Caracterización del dataset

- El dataset contiene 1,727 instancias y 7 variables, todas categóricas.
- No se detectaron valores nulos ni desconocidos en los atributos.
- Se observar 23 instancias duplicadas que podrían eliminarse.

Tabla 1. Variables del dataset.

Variable	Tipo	Descripción	Posibles valores
Buying	Categórica	Coste de compra del coche	vhigh (Muy alto), high (Alto), med (Medio), low (Bajo)
Maintenance	Categórica	Coste de mantenimiento del coche	vhigh, high, med, low
Doors	Categórica	Número de puertas del coche	2, 3, 4, 5more
Person	Categórica	Capacidad de personas en el coche	2, 4, more
lug_boot	Categórica	Tamaño del maletero del coche	small, med, big
safety	Categórica	Nivel de seguridad del coche	low, med, high
class	Categórica	Variable objetivo: Elegibilidad del coche	unacc (inaceptable), acc (aceptable), good (bueno), vgood (muy bueno)

Se observan que 1,215 instancias en la variable objetivo que están asignadas a la clase unacc, indicando una distribución desequilibrada.

Número de instancias por clase:

- unacc: 1209 (70.0%)
- acc: 384 (22.0%)
- good: 69 (4.0%)
- vgood: 65 (4.0%)

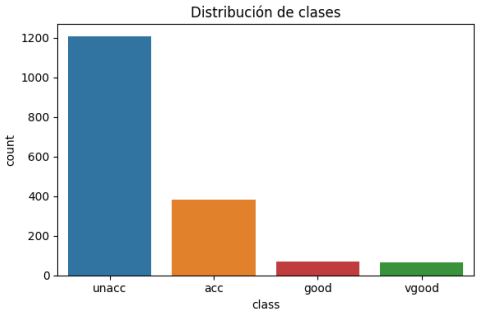

Figura 2. Distribución de clases en el dataset.

Se ha utilizado apoyo visual principalmente gráficos de barras de conteo (countplot), para visualizar la relación entre las variables categóricas y la clase, puesto que son especialmente adecuados para visualizar la distribución de variables categóricas, podemos observar los siguientes puntos:

- El número de coches aceptables tiende a aumentar a medida que el coste de compre disminuye.
- La capacidad tiene un impacto significativo en los coches con ocupación para solo 2 personas, en su mayoría se consideran inaceptables.
- La seguridad es un factor determinante para la aceptación del coche, los coches con seguridad baja casi siempre se clasifican como inaceptables.

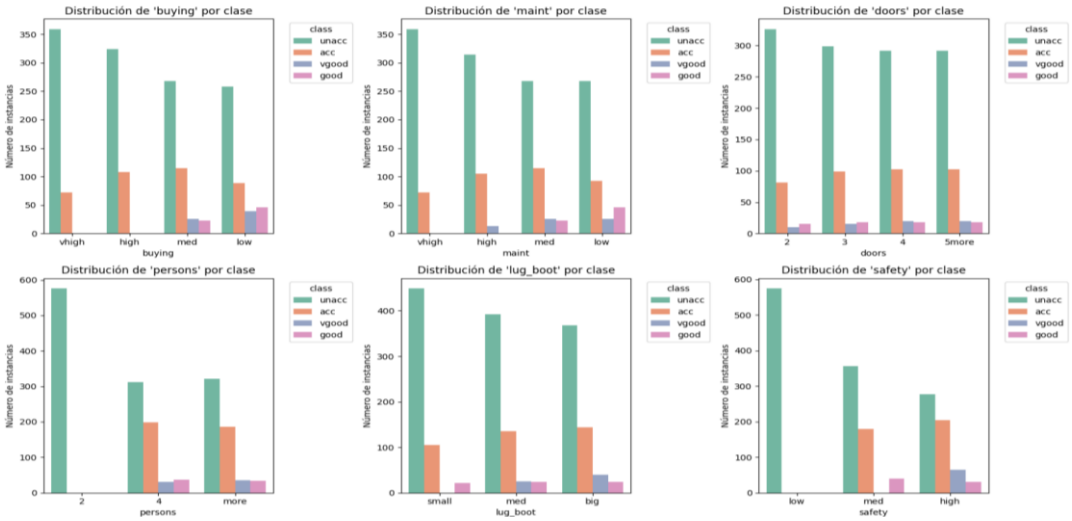

Figura 2. Distribución de clases en el dataset.

3. Justificación de los modelos seleccionados

Se han elegido dos algoritmos de clasificación que cumplen con el requisito del enunciado para la actividad.

Modelo 1: Árbol de Decisión (CART)

- Justificación: Es un algoritmo interpretable, fácil de visualizar y adecuado para variables categóricas. Permite entender las reglas de decisión utilizadas en cada clase. Es bueno para problemas de clasificación multiclase como en este caso.
- Ventajas: Se puede interpretar, no requiere escalar los datos, maneja bien variables categóricas codificadas.
- Desventajas: Tendencia al sobreajuste (overfitting) si no se limita la profundidad por medio de la modificación de sus hiper parámetros, especialmente con clases desbalanceadas que es este caso.

Modelo 2: Random Forest (RandomForestClassifier)

- Justificación: Random Forest combina múltiples árboles de decisión entrenados con muestras aleatorias del dataset (bagging) y selección aleatoria de características. Esto reduce el sobreajuste y mejora la generalización.
- Ventajas: Mayor robustez que un árbol individual, mejor rendimiento en clases desbalanceadas, reduce varianza.
- Desventajas: Menor interpretabilidad que un árbol individual, y adicionalmente tiene un mayor coste computacional.

Ambos modelos se ajustan a problemas de clasificación multiclase con variables categóricas codificadas.

4. Estrategia implementada

- Se Utilizó la estrategia de división: (80% para entrenamiento y 20% para prueba).
- Se ha aplicado stratify = y para mantener las proporciones de clases equilibradas al momento de realizar la partición de datos, es importante teniendo en cuenta el desbalanceo entre clases.
- Se estableció una semilla con el valor 123 para asegurar la reproducibilidad en futuras ejecuciones.

5. Resultados obtenidos

Accuracy Árbol de Decisión: 94.51%

Accuracy Random Forest: 95.95%

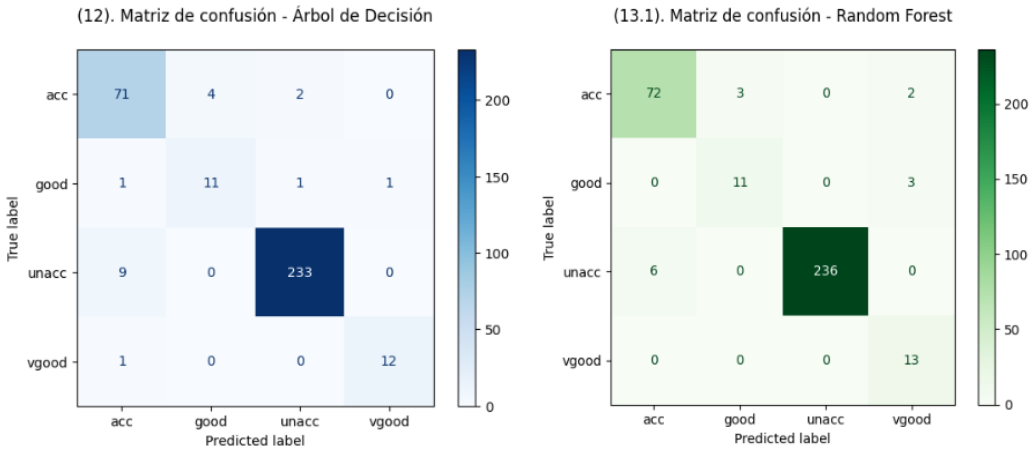

Figura 3: Matrices de confusión para ambos modelos.

6. Discusión y conclusiones

Ambos modelos presentan un buen desempeño, aunque Random Forest muestra una mayor precisión en la clasificación. El Árbol de Decisión es más interpretable, pero puede sufrir de sobreajuste si no se limita su profundidad. Para mejorar los resultados, se podrían ajustar hiperparámetros, usar técnicas de balanceo de clases o probar otros algoritmos ensemble.

En conclusión, la aplicación de técnicas de clasificación supervisada permite predecir con buena precisión la elegibilidad de un coche, lo que puede ser útil para sistemas de recomendación y apoyo a la decisión en el sector automotriz.

7. Propuestas de mejora

Técnicas de balanceo de clases: Oversampling de clases minoritarias** (SMOTE): generar instancias sintéticas de las clases good y vgood para equilibrar el dataset.

Optimización de hiperparámetros: Grid Search o Random Search**: realizar búsqueda exhaustiva de los mejores hiperparámetros (max_depth, min_samples_split, n_estimators).