Trabajo Práctico:
Aprendizaje Automatizado

Fecha Entrega: 6/07/21

Integrantes:

Litmanovich, Ignacio
Marotte, Damian Ariel

Análisis de datos

El dataset que ha sido otorgado para este trabajo es el dataset de glass. En este se presentan 214 instancias.

Se compone por 11 columnas:

1. Número de identificación
2. RI: Índice de refracción
3. Na: Sodio
4. Mg: Magnesio
5. Al: Aluminio
6. Si: Silicio
7. K: Potasio
8. Ca: Calcio
9. Ba: Bario
10. Fe: Hierro
11. Tipo de vidrio que van de (1,2,3,4,5,6,7)

Para la clasificación decidimos descartar la columnas de “Número de identificación” dado que no es un atributo de los vidrios sino que es un índice para indexar los datos.

Ninguna de las instancias se corresponde con el tipo de vidrio 4.

Metodología

Hemos realizado los entrenamientos utilizando más de 50 semillas para separar los datos y computar el promedio de las certezas. Al dataset lo dividimos en 2 grupos, el grupo de entrenamiento que tiene un 80% del dataset y el grupo de test que tiene el siguiente 20%.

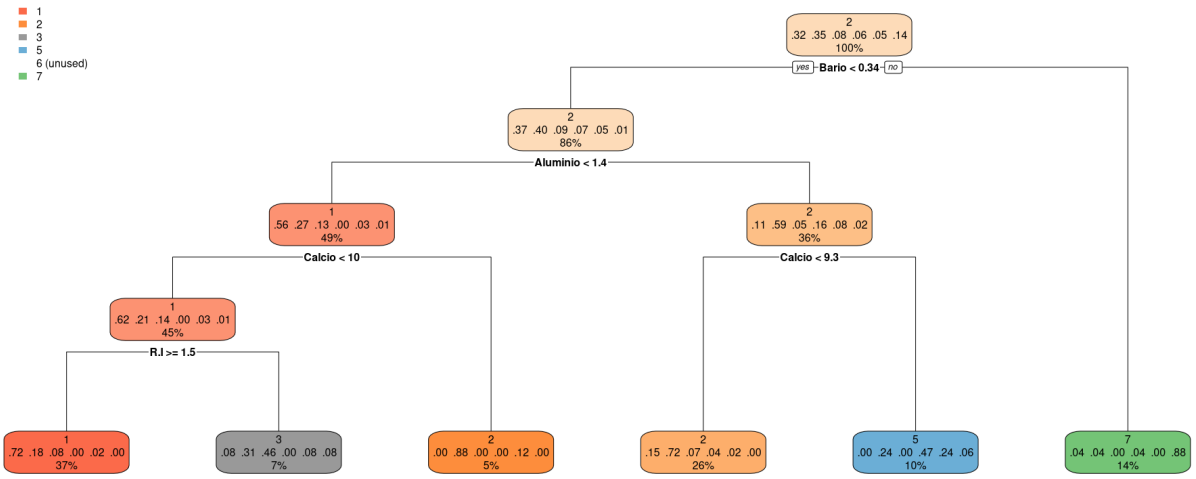
A partir de los datos pudimos observar que la cantidad de aluminio es una variable importante para la clasificación del vidrio, principalmente para la clasificación de tipo 1 y tipo 2. Se ha obtenido una accuracy de 0.5932629 con métrica de Gini en el grupo de testeo. Solo con una variable obtenemos aproximadamente el 60% de certeza, lo cual genera bastante optimismo para obtener mejores resultados.

También hemos podido percibir que el índice de refracción resulta importante para la clasificación y en conjunto con Aluminio, se puede alcanzar una certeza de 0.6671694 con métrica de Ginni.

Otra variable que muestra una mejora ante la certeza es el Bario, que con el cual se puede alcanzar el 0.7063851 con métrica Gini.

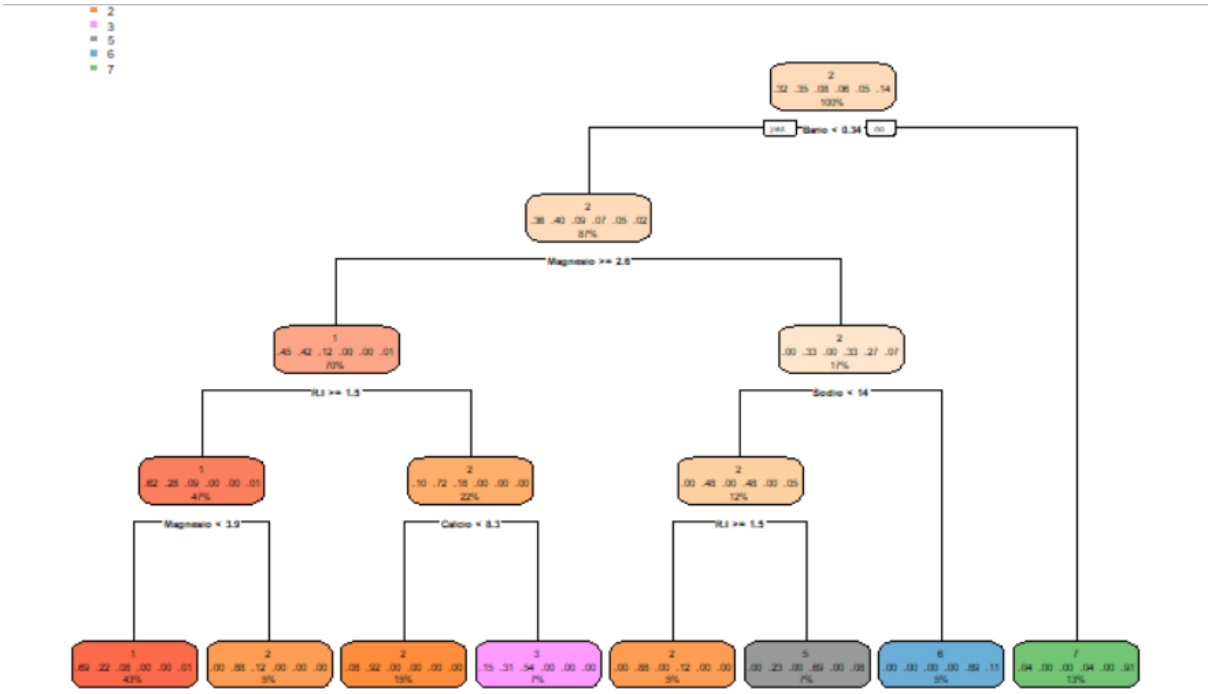
Finalmente, la última variable que mostró mayor accuracy fue el calcio, obteniendo una certeza de 0.7340372 con métrica de Gini.

De esta última combinación (Aluminio, Índice de Refracción, Bario y Calcio) y utilizando la semilla 44, pudimos obtener el siguiente árbol.



De entre todos los árboles que pudimos generar nos quedamos con este, que es uno de los más simples, al necesitar solamente 5 niveles.

Consideramos prudente recomendar otro árbol, uno que pueda clasificar a los vidrios de tipo 6 a costa de perder un poco de certeza.
Este árbol se genera con las variables, Bario, Calcio, Magnesio, Índice de refracción y Sodio. El promedio de certeza es de 0.7003519 utilizando la semilla 50.



Si bien el árbol posee más hojas y su promedio de certeza es menor, este no lo es significativamente, solo empeora alrededor de un 3%. Podría considerarse también un buen árbol para el modelo. Es preferible obtener un árbol que pueda clasificar a todos los tipos de vidrios posibles a costa de perder un poco de precisión.

Conclusiones

Como no tenemos ninguna instancia de vidrio tipo 4 esto genera un gran problema ya que no lo podemos clasificar. Tampoco con el primer árbol podemos clasificar el tipo 6 lo cual consideramos que no es lo suficientemente inteligente para representar un problema en la realidad.

Tal como puede observarse al correr el programa adjunto, el parámetro de poda no presenta ninguna mejora significativa, razón por la cual decidimos no podar el árbol.

Si bien aproximadamente un 73% de certeza sigue siendo una buena clasificación, los datos de entrenamiento son pocos. Una base de 214 instancias es un dataset muy pequeño

por lo tanto los datos observados no son suficientes. Para mejorar este árbol creemos necesario obtener más resultados para poder encontrar más patrones para la clasificación. Además consideramos necesario agregar instancias del tipo 4 porque como se ha dicho anteriormente con nuestro modelo es imposible que podamos clasificarlos.

En conclusión, un árbol de decisión podría ser un modelo apropiado para la clasificación de vidrios, pero sólo si se dispone de mayor cantidad de datos para entrenamiento.