Aprendizaje Estadístico Automatizado.
Semestre 2021-1
Tarea 5
Modelos de Árboles

Luna Gutiérrez Yanely
Ortiz Silva Ana Beatriz

17 Diciembre 2020

1 Ejercicio 1.

2 Ejercicio 2.

En este ejercicio trabajamos con la base de datos `fgl` del paquete `MASS`, la cual contiene 214 observaciones de 10 variables correspondientes a la composición de fragmentos de vidrio recolectados durante trabajo forense. Nueve de estas variables son numéricas, mientras que la variable `type` es categórica y corresponde al tipo de vidrio, el cual puede ser uno de los seis que se enumeran en el Anexo 3.2.

Con la finalidad de predecir el valor de la variable `type` respecto a las otras variables ajustamos un modelo de árbol de decisión (clasificación). Usando todas las variables predictoras obtenemos el árbol que se muestra en la Figura 1 y observamos que la variable `Ba` es decisiva para clasificar una observación a la categoría `Head` mientras que para diferenciar entre las categorías `Con` y `Tabl` es muy importante el valor de la variable `Na`.

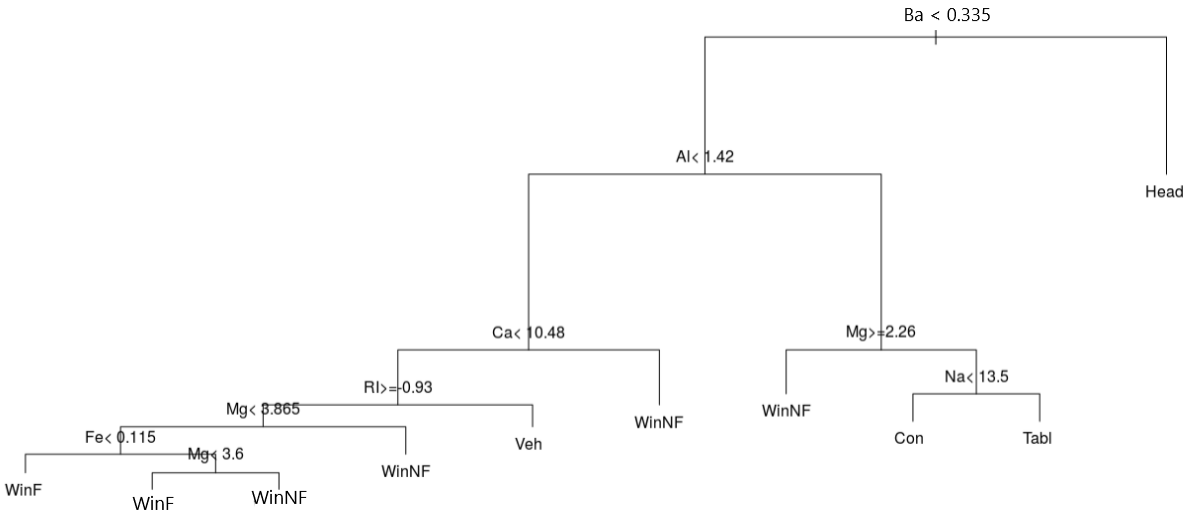

Figure 1: **Árbol de clasificación.** *Usando las 9 variables predictoras.*

Posteriormente usamos el método *bagging* para ajustar el modelo con 1000 árboles (usando muestras bootstrap). Elegimos esta cantidad de árboles porque observamos que con esta el error de clasificación es estable. (Figura 3 en el Anexo 3.2.).

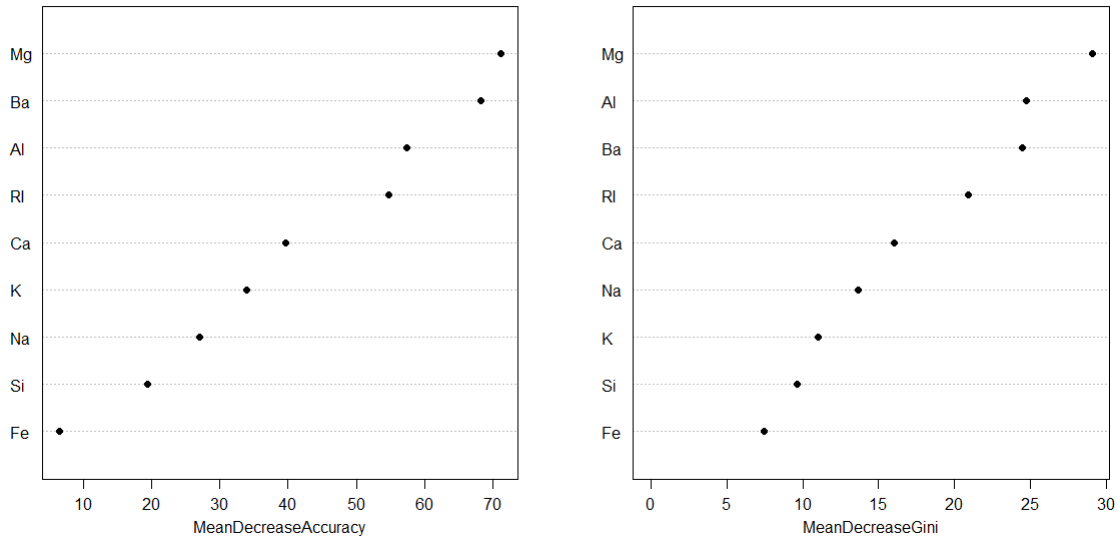

Figure 2: **Importancia de las variables predictoras** *en el modelo obtenido con bagging, ntrees=1000.*

En la Figura 2 observamos que la variable con más importancia en el modelo es `Mg` pues tanto en la precisión como en el índice de Gini es la variable tiene un valor mayor.

También usamos el método *random forest* para disminuir la correlación entre los árboles ajustados a las muestras bootstrap. Se ajustaron 1000 árboles considerando primero tres variables según la recomendación de la documentación en `R` ($\sqrt{p} = \sqrt{9} = 3$). En la Figura 4 del Anexo observamos que la variable `Mg` sigue siendo la más influyente pero ahora es seguida por las variables `Rl` y `Al` mientras que la variable `Ba` ya no figura entre las primeras como ocurrió empleando *bagging*.

Posteriormente variamos el número de predictores a considerar entre 1 y 9 para comparar las tasas de error de clasificación en el conjunto de observaciones *out of bag (oob)* y encontrar la mínima. La comparación de estas tasas se muestra en la Tabla 2 del Anexo. El modelo con 3 variables predictoras se encuentra entre los que tienen tasa de error global menor.

Para fines comparativos ajustamos también un modelo de regresión logística multinomial considerando todas las variables predictoras y con la categoría `WinF` como categoría de referencia. Notamos que para este modelo, todos los coeficientes correspondientes a los niveles `Tabl` y `Head` son significativos al 5%, lo que nos dice que hay una separación marcada entre cada uno de estas categorías con la categoría de referencia.

Para comparar los modelos ajustados evaluamos la tasa de error de clasificación aparente y de prueba. Para el modelo de regresión logística multinomial usamos el método de remuestreo de *training-test* con un conjunto entrenamiento del 80% de los datos, mientras que para *bagging* y *random forest* (considerando tres predictores) registramos las tasas de error de clasificación en el conjunto de observaciones *oob*. También incluimos las tasas de error aparente del primer árbol ajustado (Figura 1).

	Global	WinF	WinNF	Veh	Con	Tabl	Head
Aparentes Árbol 1	21	20	17	59	15	44	10
Aparentes RLM	26	29	28	76	15	0	0
Test RLM	38	37	38	85	38	32	17
Aparentes Bagging	0	0	0	0	0	0	0
oob-Bagging	24	16	28	65	15	22	14
Aparentes Random Forest	0	0	0	0	0	0	0
oob-Random Forest	19	10	21	59	23	22	10

Table 1: **Tasas de error de clasificación** *aparentes y de prueba para cada uno de los modelos ajustados, expresadas en porcentaje.*

En la Tabla 1 observamos que en el modelo de regresión logística multinomial (RLM) las tasas de error de prueba aumentan en promedio un 10% para cada categoría así como la global. Sin embargo, este modelo tiene tasas de error aparentes mayores a las que se observan en el primer árbol ajustado para 3 de las 6 categorías y en la tasa global.

Por otro lado, para *bagging* como para *random forest* las tasas de error de clasificación aparentes son cero, sin embargo, las tasas en el conjunto de observaciones *oob* toman valores cercanos a las de los otros modelos, siendo *random forest* el método con menores tasas (5% menor en la tasa global, lo que representa una disminución relativa del 21%) por lo que optaríamos por este modelo como modelo predictivo.

3 Anexo

3.1 Ejercicio 1.

3.2 Ejercicio 2.

Niveles de la variable categórica `type` de la base `fgl` y número de observaciones correspondiente:

- `WinF`: Vidrio flotado proveniente de ventanas. 70 observaciones.
- `WinNF`: Vidrio no flotado proveniente de ventanas. 76 observaciones.
- `Veh`: Vidrio de ventanas de vehículos. 17 observaciones.
- `Con`: Vidrio proveniente de contenedores. 13 observaciones.

- `Tab`: Vajillas. 9 observaciones.
- `Head`: Faros de vehículos. 29 observaciones.

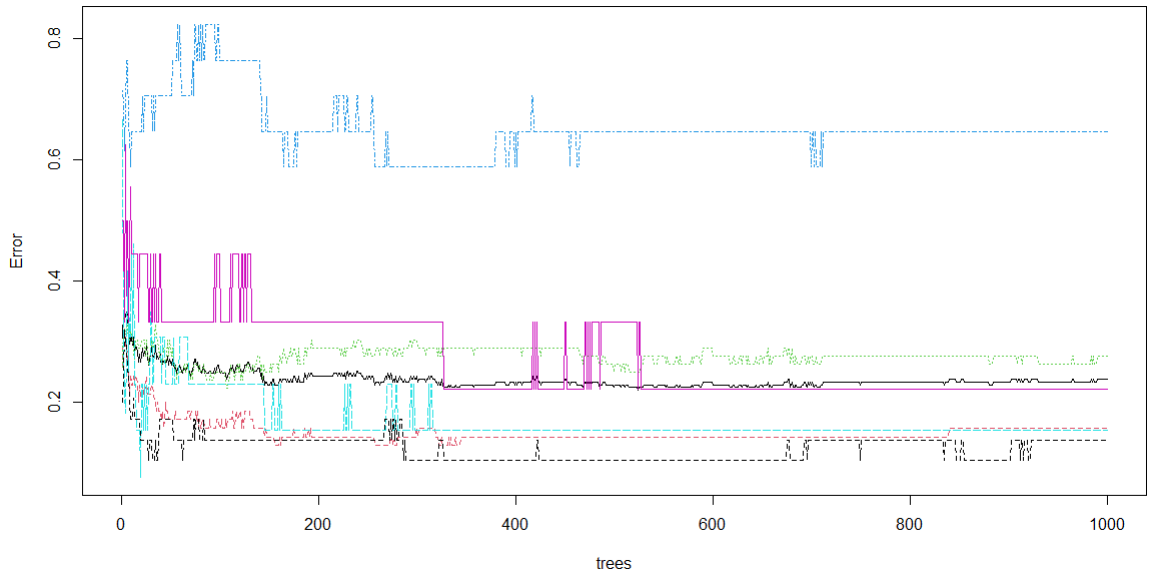

Figure 3: **Error de clasificación oob en el método bagging para cada categoría de** `type` *según el número de árboles ajustados. Observamos que 1000 árboles son suficientes para obtener valores estables del error.*

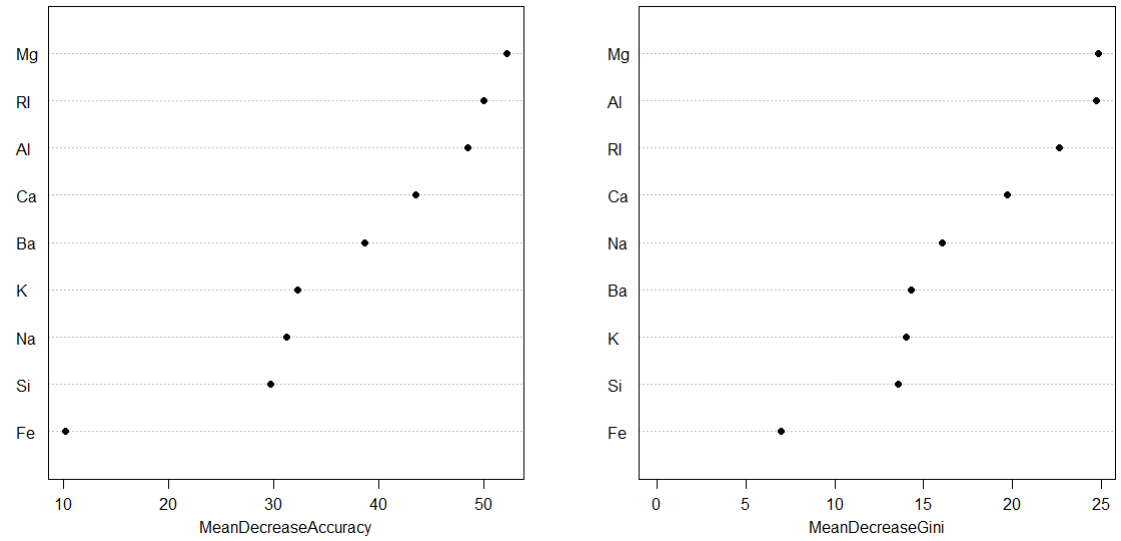

Figure 4: **Importancia de las variables predictoras usando random forest con 1000 árboles y 3 predictoras a considerar en cada uno.** *La variable* `Mg` *sigue siendo la más influyente pero ahora es seguida por las variables* `Rl` *y* `Al`.

	Global	WinF	WinNF	Veh	Con	Tabl	Head
1	22	11	18	82	46	11	14
2	20	10	18	65	31	22	14
3	20	11	20	59	31	22	10
4	20	13	20	53	23	22	14
5	21	14	25	53	23	22	10
6	21	14	22	59	31	22	10
7	23	19	25	53	31	22	10
8	23	16	28	59	23	22	10
9	23	17	26	53	23	22	14

Table 2: **Tasas de error de clasificación en el conjunto de observaciones *oob* de random forest** *expresadas en porcentaje variando el número de predictores a considerar. La tasa de error global varia entre 20% y 22% sin embargo no hay un modelo en el que todas las tasas de error para cada categoría sean mínimas.*