Análisis predictivo sobre el consumo de alcohol en estudiantes

Sergi Fornés

Índice

Introducción

El objetivo de este trabajo es estimar un buen modelo que sea capaz de predecir la cantidad de alcohol que consume un estudiante de secundaria según su información demográfica. Para realizar esto se han usado los datos de una encuesta a estudiantes de matemáticas y de portugués de secundaria. Se pueden encontrar los datos en el siguiente *link*.

Hay un total de 674 alumnos con las siguientes variables:

- `school`: Instituto del estudiante.
- `sex`: Sexo del estudiante.
- `age`: Edad del estudiante.
- `address`: Tipo de zona en la que vive el estudiante. Zona urbana (`U`) o zona rural (`R`).
- `famsize`: Tamaño de la familia del estudiante. Menos o igual de 3 (`LE3`) o más de 3 (`GT3`).
- `Pstatus`: Convivencia de los padres del estudiante. Los padres conviven (`T`) o no (`A`).
- `Medu`: Educación de la madre del estudiante. Ninguna (0), cuarto de primaria (1), tercero de la ESO (2), bachillerato (3) o educación superior (4).
- `Fedu`: Educación del padre del estudiante. Ninguna (0), cuarto de primaria (1), tercero de la ESO (2), bachillerato (3) o educación superior (4).
- `Mjob`: Trabajo de la madre del estudiante. Maestra (`teacher`), relacionado con el cuidado (`health`), funcionaria (`services`), tareas domésticas (`at_home`) u otro (`other`).
- `Fjob`: Trabajo del padre del estudiante. Maestro (`teacher`), relacionado con el cuidado (`health`), funcionario (`services`), tareas domésticas (`at_home`) u otro (`other`).
- `reason`: Razón de haber elegido la escuela del estudiante. Cerca de casa (`home`), reputación del instituto (`reputation`), preferencia en la formación (`course`) u otra (`other`).
- `guardian`: Tutor del estudiante. Madre (`mother`), padre (`father`) u otro (`other`).

- `traveltime`: Duración en horas del trayecto desde la casa del estudiante hasta el instituto.
- `studytime`: Horas semanales de estudio.
- `failures`: Número de faltas de asistencia a clase.
- `schoolsup`: Apoyo educativo adicional.
- `famsup`: Apoyo educativo familiar.
- `paid`: Clases de repaso de la asignatura.
- `activities`: Actividades extraescolares.
- `nursery`: Asistió a educación preescolar.
- `higher`: Quiere estudiar educación superior.
- `internet`: Acceso a internet en casa.
- `romantic`: Tiene una relación romántica.
- `famrel`: Calidad de la relación familiar. De peor (1) a mejor (5).
- `freetime`: Tiempo libre después del instituto. De nada (1) a mucho (5).
- `goout`: Sale con los amigos. De nada (1) a mucho (5).
- `Dalc`: Consumo de alcohol entre semana. De nada (1) a mucho (5).
- `Walc`: Consumo de alcohol los fines de semana. De nada (1) a mucho (5).
- `health`: Estado de salud actual. De muy mala (1) a muy buena (5).
- `absences`: Número de ausencias escolares.
- `G1`: Nota en el primer trimestre de la asignatura. De 0 a 20.
- `G2`: Nota en el segundo trimestre de la asignatura. De 0 a 20.
- `G3`: Nota final de la asignatura. De 0 a 20.

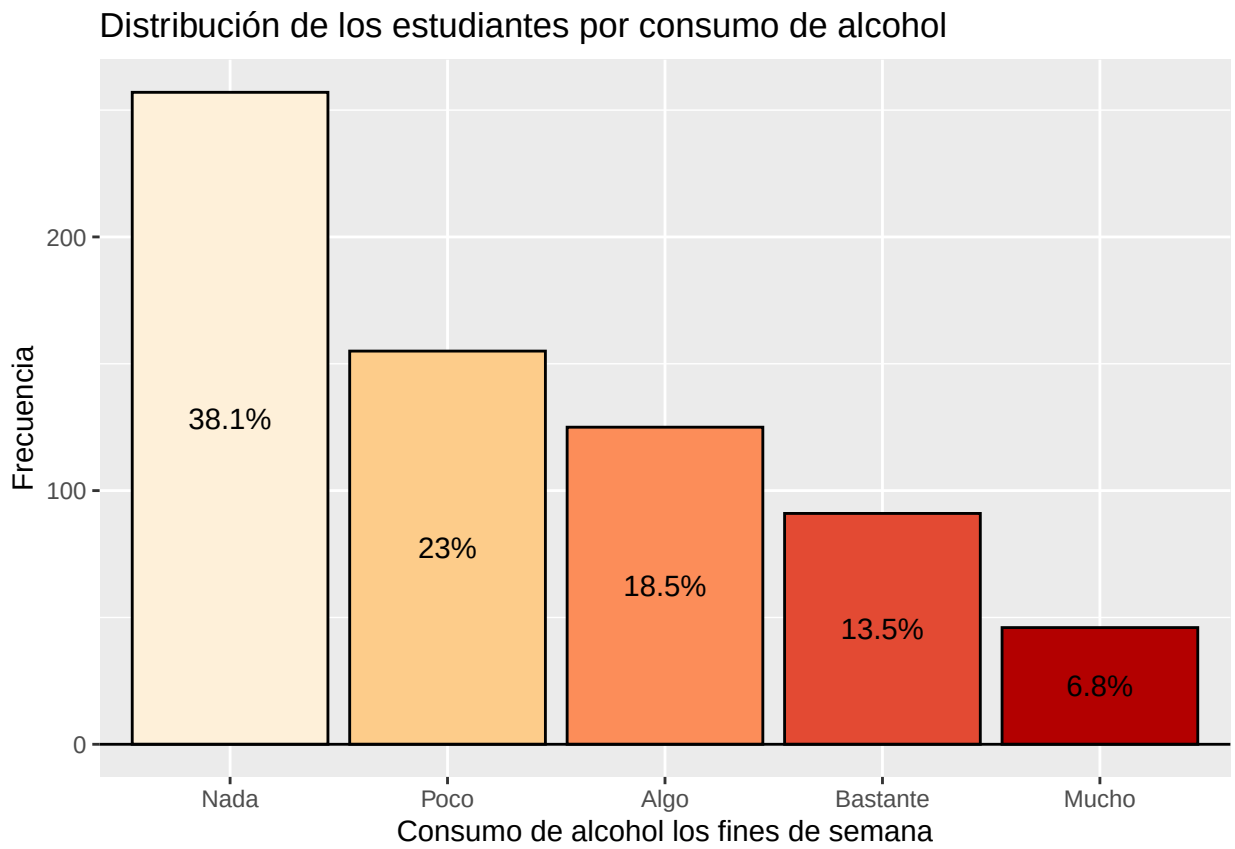
Hay dos variables que contienen el consumo de alcohol de los estudiantes, `Dalc` y `Walc`. Se predecirá la variable `Walc` debido a que el consumo los fines de semana seguramente es más significativo que el consumo entre semana. Las demás variables se usarán para estimar la predicción, exceptuando `Dalc`, ya que precisamente se quiere estimar el consumo de alcohol, así que no se añadirá como variable explicativa. Tampoco se añadirá la variable `school` puesto que se quiere hacer una predicción general, y en los datos únicamente hay dos institutos.

Primero de todo se arreglarán los datos para poder trabajar correctamente con ellos. Después se realizará un pequeño análisis descriptivo para observar la distribución de las variables de interés. Una vez hecho esto, se usarán distintas técnicas y se compararán sus resultados para elegir el mejor modelo.

Los datos vienen en dos tablas distintas, una con la encuesta de los estudiantes de matemáticas y la otra con la encuesta de los estudiantes de portugués. Algunos alumnos se encuentran en las dos tablas, por lo que únicamente queremos contarlos una vez. Se supone que un alumno se encuentra en ambas tablas si se encuentran dos observaciones con todas las variables iguales exceptuando los valores de `failures`, `paid`, `absences`, `G1`, `G2` y `G3`, ya que estas variables son específicas de cada clase, por ejemplo, un mismo alumno puede tener diferentes calificaciones en clase de matemáticas y de portugués. Para estas variables mencionadas anteriormente se tienen muchos valores `NA` debido a que muchos estudiantes no asisten a ambas clases, así que se colapsarán los valores de ambas asignaturas para evitar este problema. Las variables `failures`, `absences`, `G1`, `G2` y `G3` son numéricas, por lo que nos podemos quedar con el valor medio de las dos clases para los alumnos que asisten a ambas. Por otro lado `paid` es categórica, de modo que podemos redefinirla como que un estudiante recibe alguna clase particular, ya sea de matemáticas o de portugués.

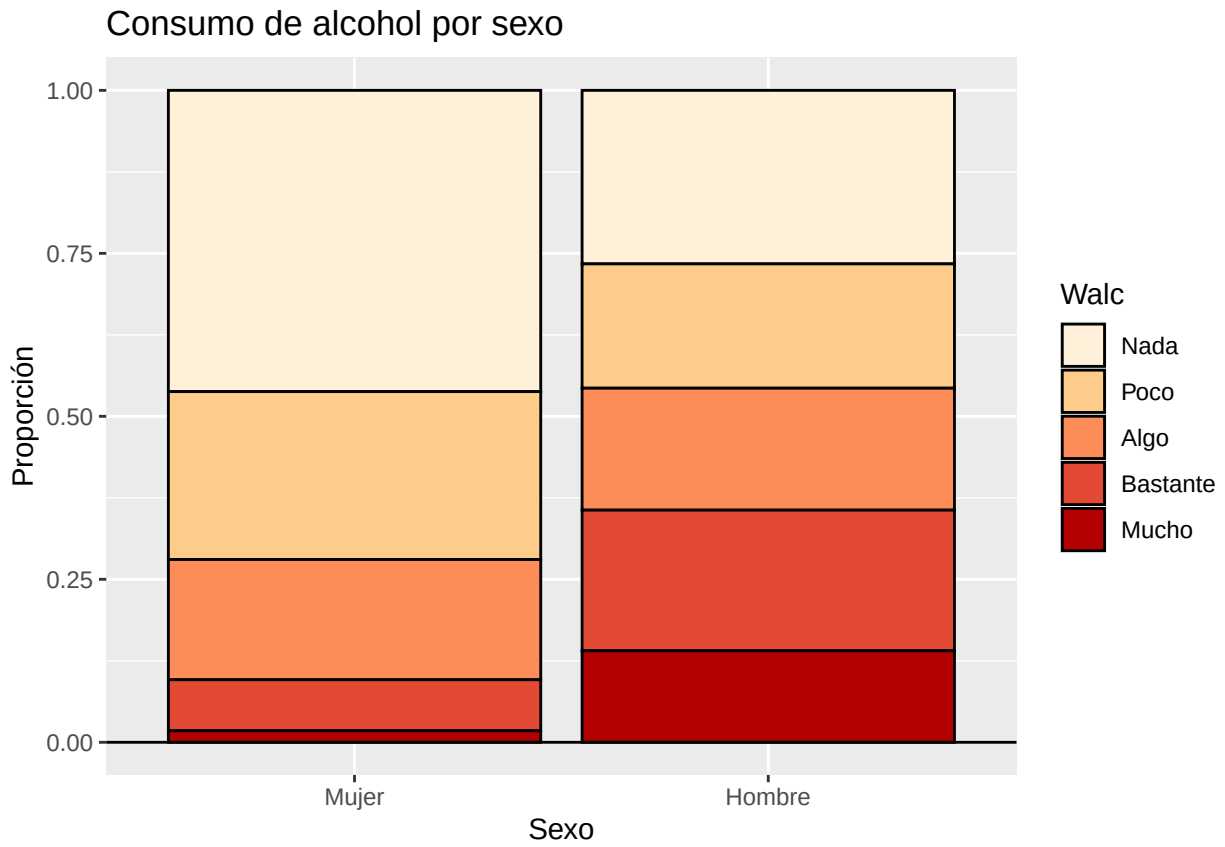
Análisis Exploratorio

La variable de interés `Walc` es cualitativa y tiene cinco valores posibles. La mayoría de los estudiantes no consumen mucho alcohol pero igualmente los datos no están muy desequilibrados, la categoría con menos estudiantes contiene más del 5 % de los estudiantes. Por lo tanto, se buscará el modelo que mejor consiga clasificar estudiantes en esta dimensión.

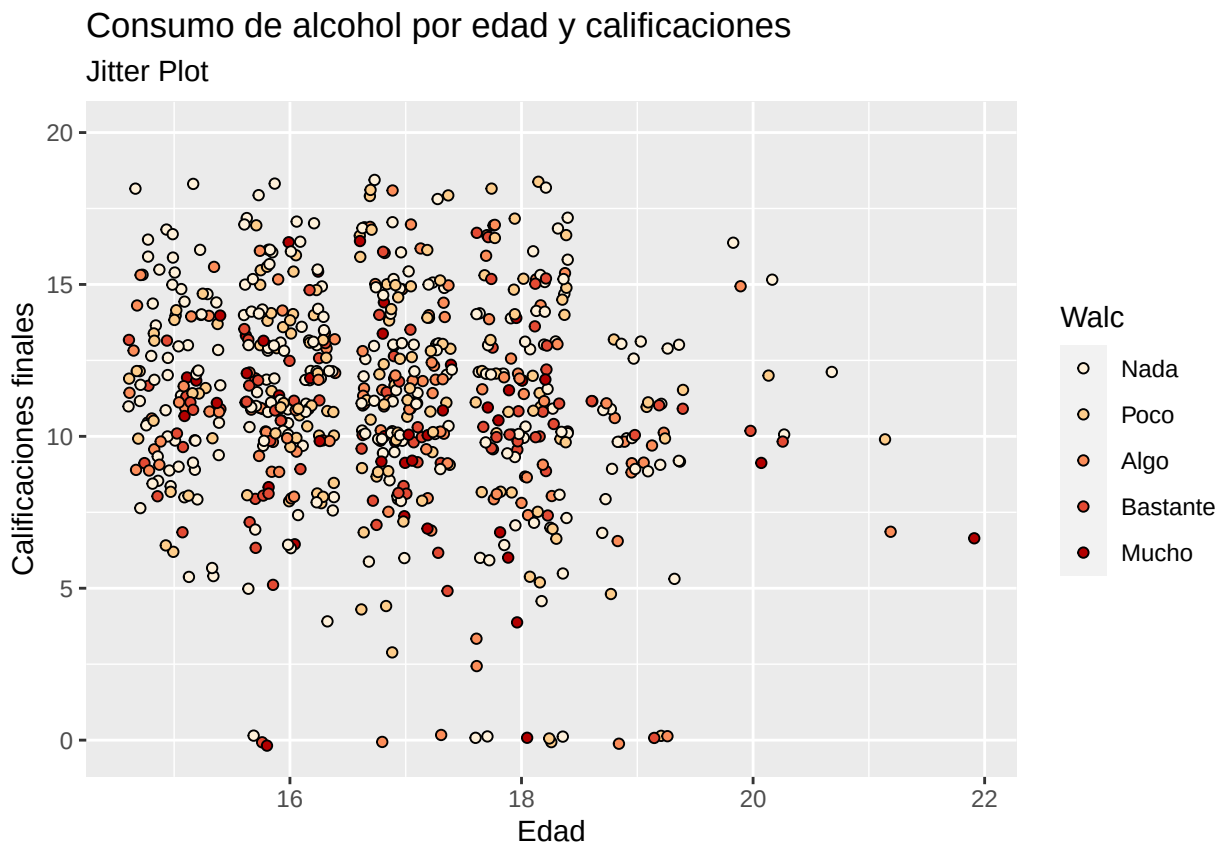


A continuación se observará la relación que tiene `Walc` con otras variables. Las variables más influyentes sobre el consumo de alcohol podrían ser el sexo del estudiante, la edad, las calificaciones, y si suele salir con sus amigos.

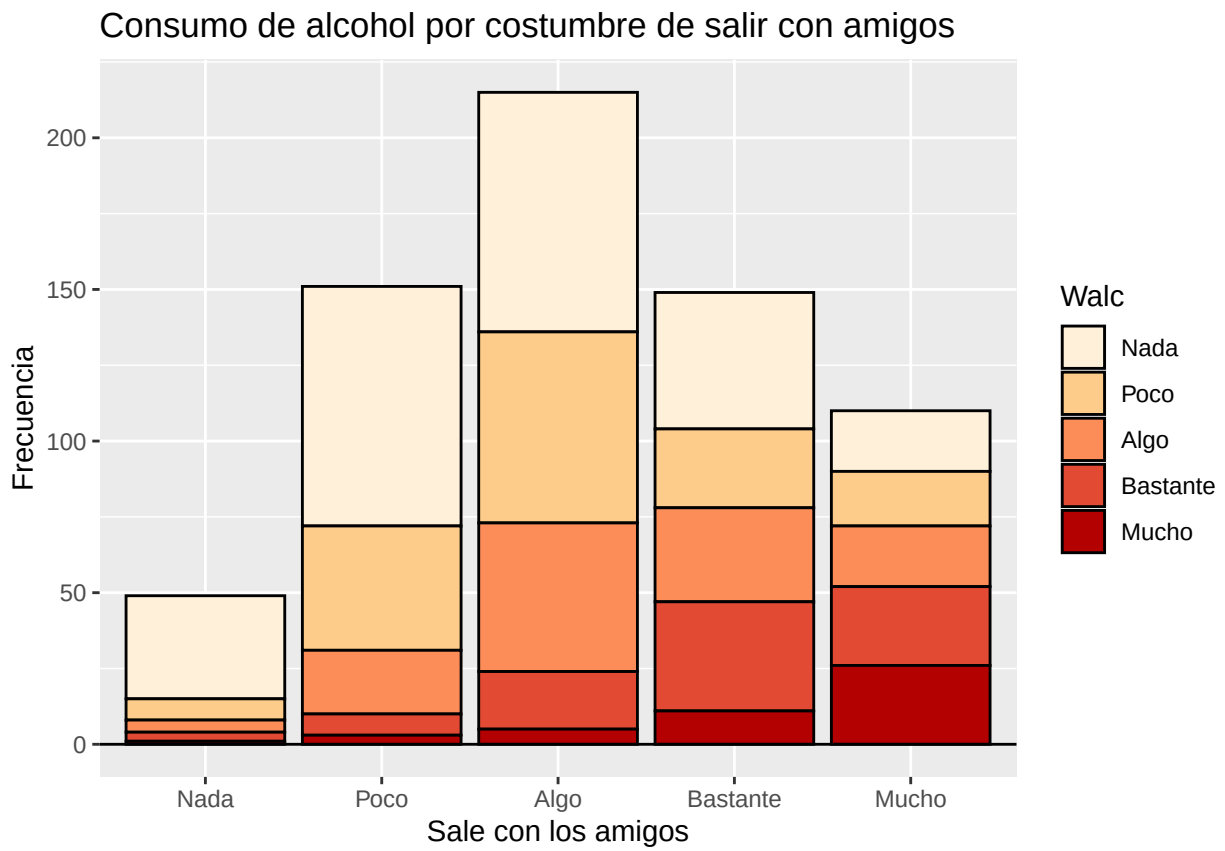
Efectivamente, el sexo del estudiante está muy relacionado con el consumo de alcohol. Los hombres consumen mucho más alcohol los fines de semana que las mujeres.



En cuanto a la relación con la edad y las calificaciones, los estudiantes consumidores de alcohol parece que se encuentran por el centro de la distribución de notas y uniformemente distribuidos por edad. Por lo que a simple vista no se observa una fuerte relación.



Mientras tanto, el efecto de salir con los amigos si que parece ser muy fuerte. En proporción, los estudiantes que más salen con sus amigos beben más.



Se ha observado que los diferentes grupos de estudiantes según su consumo de alcohol no están claramente separados por las demás variables, y además los diferentes valores de `Walc` pueden resultar ambiguos, por lo que no se espera estimar un modelo que consiga clasificar correctamente la mayoría de los casos.

Métodos de Clasificación

El principal objetivo de este trabajo es la predicción, por lo que se ajustarán diversos modelos de clasificación y se obtendrá el porcentaje de observaciones correctamente clasificadas de cada uno de ellos, es decir, su precisión. La inferencia y la interpretación no serán de interés. Primero se realizará un Linear Discriminant Analysis, un modelo relativamente sencillo que suele dar buenos resultados a la hora de clasificar distribuciones multinomiales. Después se estimará un modelo más potente, un Random Forest, basado en árboles de decisión. Y por último ajustaremos un Extreme Gradient Boosting, otro algoritmo basado en árboles de decisión pero mucho más potente.

Linear Discriminant Analysis

Con el objetivo de que no haya sobreajuste en el modelo, este se ha estimado mediante Leave-One-Out Cross-Validation. Una vez ajustado, se han clasificado las observaciones y se ha obtenido la siguiente tabla de confusión:

	Walc Predicho				
	Nada	Poco	Algo	Bastante	Mucho
Nada	168	50	24	8	7
Poco	78	34	26	15	2
Algo	53	26	16	23	7
Bastante	21	13	19	23	15
Mucho	6	2	6	17	15

En la cual tenemos los valores actuales de la variable `Walc` en las filas y los valores predichos en las columnas. Con esta tabla se puede calcular que el porcentaje de observaciones correctamente clasificadas es del 37.98 %.

Random Forest

El modelo de Random Forest se ha ajustado con Nested Cross-Validation. Para ello, primero se han dividido las observaciones entre datos de entrenamiento (80 % de las observaciones) y datos de validación (20 % restante). Después se han dividido los datos de entrenamiento en 5 cajas de aproximadamente el mismo tamaño para realizar un 5-Fold Cross-Validation, con el objetivo de encontrar el hiperparámetro `mtry` del Random Forest óptimo. Para ello se realiza, para cada caja, una estimación del modelo Random Forest con diferentes valores de `mtry`. Se han elegido como posibles valores $\{2, 3, 4, 6, 9, 12\}$. De entre todos los modelos de una misma caja `k`, los valores con los que se estima el mejor modelo son:

	mtry
k1	9
k2	6
k3	9
k4	12
k5	9

El valor óptimo del hiperparámetro podría ser mtry = 9, por lo que se estima el Random Forest usando este valor y el total de datos de entrenamiento para así poder conocer la precisión real del algoritmo usando Cross-Validation. Finalmente se predicen los datos de validación y se comparan con los valores reales, obteniéndose un porcentaje de observaciones correctamente clasificadas del 41.04 %.

Extreme Gradient Boosting

Para realizar la estimación del modelo de Extreme Gradient Boosting se ha usado la misma metodología de Nested Cross-Validation que en el apartado anterior, pero en este caso se tienen que elegir los hiperparámetros del Extreme Gradient Boosting. En la siguiente tabla se observan los diferentes valores que se han probado para la optimización de los hiperparámetros:

Hiperparámetro	Valores
`nrounds`	{50}
`max_depth`	{3, 6, 10}
`eta`	{0.01, 0.1, 0.3}
`gamma`	{0.01}
`colsample_bytree`	{0.5, 1}
`min_child_weight`	{0, 1}
`subsample`	{0.5, 1}

Los mejores modelos de cada caja `k` tienen los siguientes hiperparámetros:

	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
k1	50	3	0.01	0.01	1.0	1	0.5
k2	50	3	0.01	0.01	0.5	0	0.5
k3	50	10	0.01	0.01	1.0	0	0.5
k4	50	3	0.01	0.01	0.5	1	1.0
k5	50	10	0.01	0.01	1.0	1	0.5

Se puede observar que los valores de la caja 1 son los más comunes entre las cajas, por lo que se usarán estos en la estimación del modelo final. Tras ajustar el modelo con todos los datos de entrenamiento y los anteriores hiperparámetros, se utiliza este Extreme Gradient Boosting para predecir los datos de validación. Finalmente se comparan las predicciones con los valores reales y se obtiene que un 39.55 % de las observaciones han sido correctamente clasificadas.

Conclusiones

Una vez se han realizado los modelos y conocemos su precisión, se pueden comparar para elegir el mejor. A la hora de compararlos también se debe tener en cuenta el modelo trivial que clasificaría todas las observaciónes al mayor grupo. En este caso, ese modelo predeciría que ningún alumno consume alcohol, por lo que clasificaría correctamente un 38.13 % de las observaciones.

Modelo	Precisión
Trivial	38.13 %
Linear Discriminant Analysis	37.98 %
Random Forest	41.04 %
Extreme Gradient Boosting	39.55 %

En la tabla anterior se puede observar como el modelo Linear Discriminant Analysis es muy malo, incluso el modelo trivial lo hace mejor. Los modelos basados en árboles predicen prácticamente con la misma precisión, y ambos lo hacen ligeramente mejor que el modelo trivial. Pero aunque la diferencia sea mínima, el Random Forest ha clasificado mejor que el Extreme Gradient Boosting. Por esto, se vuelve a ajustar un modelo Random Forest con mtry = 9 y todas las observaciones para conseguir predecir de la mejor manera el consumo de alcohol los fines de semana de alumnos de secundaria.

Una vez estimado el Random Forest, se puede obtener la disminución media de GINI de las variables explicativas. Esta es una medida de la importancia global de la variable y representa la disminución de la impureza de los nodos producida por la varable en cuestión. A mayor valor, mayor importancia tiene la variable a la hora de hacer la predicción. En la tabla siguiente se pueden observar estos índices.

Variable	Disminución Media de GINI
absences	35.93
goout	35.20
G2	34.85
G1	34.05
G3	32.89
Mjob	26.26
freetime	22.87
age	22.66
health	22.24
reason	20.68
famrel	20.25
Fjob	19.40
Fedu	19.05
Medu	18.37
studytime	17.26
sex	15.29
traveltime	13.87
failures	10.47
guardian	9.73
famsize	9.42
activities	7.85
famsup	7.67
paid	7.60
romantic	7.04
nursery	6.77
address	5.95
internet	5.02
schoolsup	4.85
Pstatus	4.04
higher	3.56

Las variables con mayor importancia son las ausencias escolares, la costumbre de salir con los amigos y las notas. Mientras que la intención de realizar estudios superiores, la convivencia de los padres y el apoyo educativo son las que menos influyen.

Igualmente hay que tener cuidado con estas conclusiones debido a la pobre precisión del modelo. Esto seguramente es debido a la poca cantidad de observaciones y a la naturaleza de los datos. Las encuestas con tantas respuestas cualitativas suelen estar sesgadas y además la variable dependiente tiene muchos valores posibles.