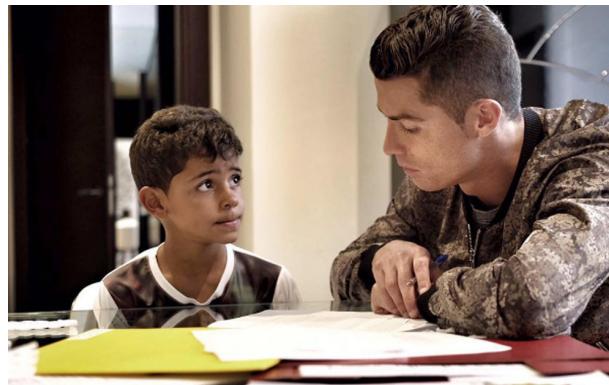


# Aprendizaje Estadístico TP N° 2: Rendimiento académico en Portugal



2º Cuatrimestre, 2022

Apellido y Nombre	Email
Sabatino, Gonzalo	gsabatino@fi.uba.ar
Pacheco, Federico	fpacheco@fi.uba.ar

## Índice

<b>1. Introducción</b>	<b>2</b>
<b>2. Análisis inicial de variables</b>	<b>2</b>
2.1. Descripción de variables . . . . .	2
2.2. Correlación entre variables . . . . .	5
2.3. Análisis de clusters con KMeans . . . . .	6
<b>3. Validación de supuestos</b>	<b>12</b>
<b>4. Modelo lineal</b>	<b>22</b>
4.1. Búsqueda de modelos con todos los predictores . . . . .	22
4.2. Búsqueda de modelos sin $G_1$ y $G_2$ . . . . .	29
4.3. Modelos escogidos . . . . .	32
<b>5. Árboles y Ensambles</b>	<b>34</b>
5.1. Búsqueda de modelos con todos los predictores . . . . .	34
5.2. Búsqueda de modelos sin $G_1$ y $G_2$ . . . . .	40
<b>6. Conclusiones y análisis de resultados</b>	<b>46</b>
<b>A. Apéndice A – Análisis de Outliers</b>	<b>48</b>
A.1. Matemáticas . . . . .	48
A.2. Portugués . . . . .	49

## 1. Introducción

En el presente trabajo práctico final se busca estudiar el rendimiento académico de los estudiantes en matemática y portugués, analizando cuáles son las diferentes características que impactan en la nota final de los cursos, y además se busca conseguir el mejor modelo que nos ayude a predecir la nota final de un alumno.

## 2. Análisis inicial de variables

### 2.1. Descripción de variables

Durante el estudio de este trabajo se trabajó con dos datasets (matemáticas y portugués), ambos con la misma cantidad de covariables, y la misma variable *G3* a predecir. Se cuenta con un total de 32 covariables, siendo un total de 33 variables. Realizando una breve descripción de variables, se cuenta con 13 variables binarias (pertenece a una categoría o no), 4 variables categóricas (más de dos categorías) y 16 variables numéricas:

Variables binarias:

- **school** - escuela del estudiante (binario: 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira)
- **sex** - género del estudiante (binario: 'F' - femenino o 'M' - masculino)
- **address** - tipo de dirección de la casa del estudiante (binario: 'U' - urbana o 'R' - rural)
- **famsup** - apoyo educativo familiar (binario: sí o no)
- **paid** - clases extra pagadas dentro de la asignatura (matemáticas o portugués) (binario: sí o no)
- **activities** - actividades extracurriculares (binario: sí o no)
- **nursery** - asistió a guardería (binario: sí o no)
- **higher** - quiere seguir estudios superiores (binario: sí o no)
- **internet** - acceso a Internet en casa (binario: sí o no)
- **romantic** - con una relación romántica (binario: sí o no)
- **schoolsup** - apoyo educativo adicional (binario: sí o no)
- **famsize** - tamaño de la familia (binario: 'LE3' - menos o igual a 3 o 'GT3' - más de 3)
- **Pstatus** - estado de convivencia de los padres (binario: 'T' - viviendo juntos o 'A' - separados)

Variables categóricas:

- **Mjob** - trabajo de la madre (nominal: 'teacher', relacionado con la atención médica, 'services' civiles (por ejemplo, administrativos o policiales), 'at home' u 'otro')
- **Fjob** - trabajo del padre (nominal: 'teacher', relacionado con la atención médica, 'services' civiles (por ejemplo, administrativos o policiales), 'at home' u 'otro')
- **reason** - razón para elegir esta escuela (nominal: cerca de 'casa', 'reputación' de la escuela, preferencia de 'curso' u 'otro')
- **guardian** - tutor del estudiante (nominal: 'madre', 'padre' u 'otro')

Variables numéricas:

- **age**: edad del estudiante (numérica: de 15 a 22)
- **Medu**: educación de la madre (numérica: 0 - ninguna, 1 - educación primaria (4to grado), 2 - de 5to a 9no grado, 3 - educación secundaria o 4 - educación superior)
- **Fedu**: educación del padre (numérica: 0 - ninguna, 1 - educación primaria (4to grado), 2 - de 5to a 9no grado, 3 - educación secundaria o 4 - educación superior)
- **traveltime**: tiempo de viaje de casa a la escuela (numérica: 1 - 1 hora)
- **studytime**: tiempo semanal de estudio (numérica: 1 - 10 horas)
- **failures**: número de fallos en clases anteriores (numérica: n si 1  $\neq$  n  $\neq$  3, de lo contrario 4)
- **famrel**: calidad de las relaciones familiares (numérica: de 1 - muy mala a 5 - excelente)
- **freetime**: tiempo libre después de la escuela (numérica: de 1 - muy poco a 5 - mucho)
- **goout**: salir con amigos (numérica: de 1 - muy poco a 5 - mucho)
- **Dalc**: consumo de alcohol durante la semana (numérica: de 1 - muy bajo a 5 - muy alto)
- **Walc**: consumo de alcohol durante el fin de semana (numérica: de 1 - muy bajo a 5 - muy alto)
- **health**: estado de salud actual (numérica: de 1 - muy malo a 5 - muy bueno)
- **absences**: número de ausencias (numérica: de 0 a 75)
- **G1**: nota del primer período (numérica: de 0 a 20)
- **G2**: nota del segundo período (numérica: de 0 a 20)
- **G3**: nota final (numérica: de 0 a 20, objetivo de salida)

El cual podemos observar que la variable que utilizaremos para predecir (G3) es numérica, tomando valores de 0 a 20:

Cuadro 1: Descripción de variable G3 para dataset de portugués y matemáticas

Dataset	n	Media	SD	Mediana	Mínimo	Máximo
Portugués	649	11.91	3.23	12	0	19
Matemáticas	395	10.42	4.58	11	0	20

El cual podemos observar, que para ambos datasets en promedio la nota final tiene un valor cercano a 10 y 11. Su distribución la podemos observar de manera mas clara en los siguientes histogramas:

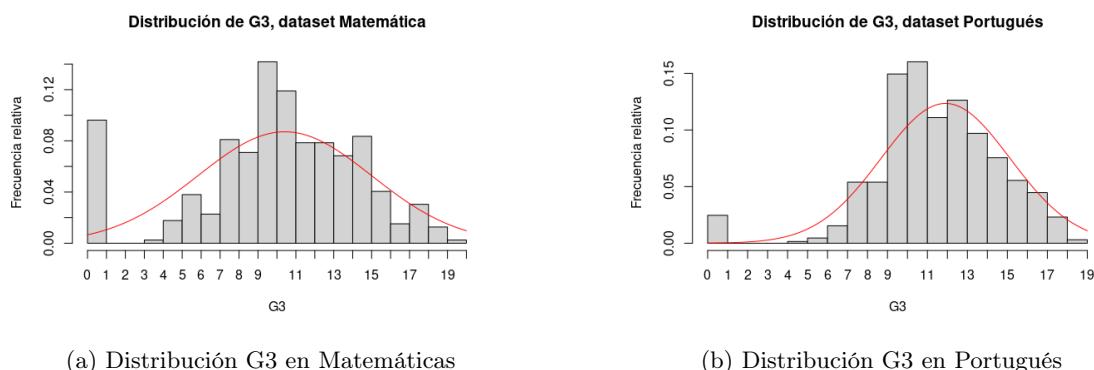


Figura 1: Distribuciones de G3

En el cual en ambos se observa lo obtenido en tabla descriptiva como también se puede observar que tienen como esta variable para ambos datasets tienen comportamiento a una distribución normal.

Podemos observar como se comporta  $G3$  en los datasets para las diferentes tipos de escuelas. Recordemos que un estudiante podría pertenecer a la escuela 'GP' (Gabriel Pereira) o 'MS' (Mousinho da Silveira).

Cuadro 2: Descripción de variable  $G3$  para dataset de matemáticas, por escuela

Escuela	n	Media	SD	Mediana	Mínimo	Máximo
GP	349	10.49	4.63	11	0	20
MS	46	9.85	4.24	10	0	19

Cuadro 3: Descripción de variable  $G3$  para dataset de portugués, por escuela

Escuela	n	Media	SD	Mediana	Mínimo	Máximo
GP	423	12.58	2.63	13	0	19
MS	226	10.65	3.83	11	0	19

Esta información la podemos detallar de forma más clara en un boxplot:

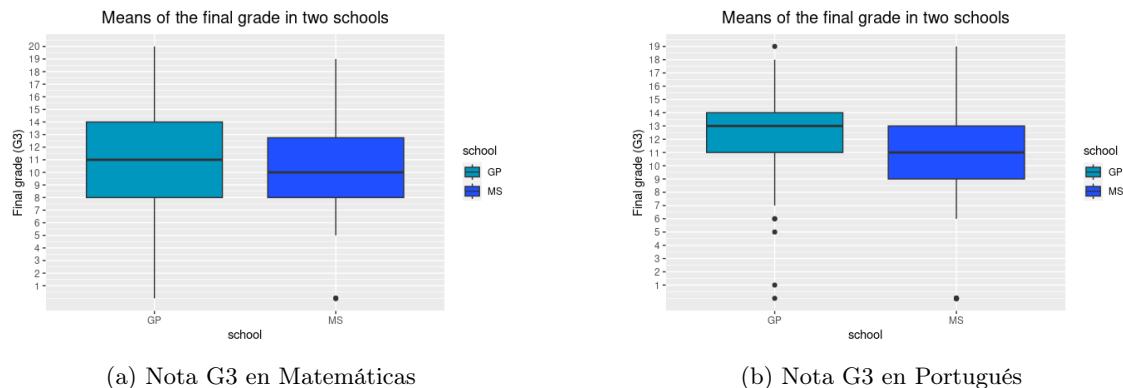


Figura 2: Nota  $G3$  en diferentes escuelas

Vemos que en el curso de matemáticas no hay una diferencia en las nota final ( $G3$ ) conseguida. En promedio parecen asemejarse. Si podemos observar una diferencia en el curso de portugués en el cual se obtiene una mayor nota en promedio en la escuela 'GP'.

Otra variable que nos parece interesante analizar es la covariante *absences* sobre las ausencias a los cursos:

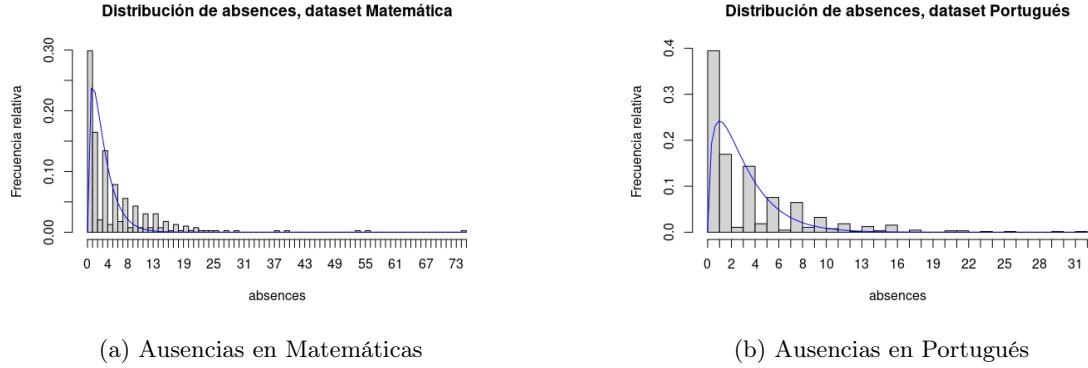


Figura 3: Ausencias en los diferentes cursos

Y la tabla descriptiva detallando mas información:

Cuadro 4: Descripción de variable absences para dataset de portugués y matemáticas

Dataset	n	Media	SD	Mediana	Mínimo	Máximo
Portugués	649	3.66	4.64	2	0	32
Matemáticas	395	5.71	8	4	0	75

En el cual, a partir de los gráficos y las tablas podemos observar como tiene un similar comportamiento a una distribución de gamma. En general podemos concluir que no suelen faltar demasiado a los cursos salvo en el curso de matematicas donde se registra una mayor cantidad de faltas en promedio.

## 2.2. Correlación entre variables

Veamos que ocurre con la estimación de las correlaciones entre las variables numéricas. Recorremos que esta correlación estimada también es conocida coeficiente de correlación de Pearson:

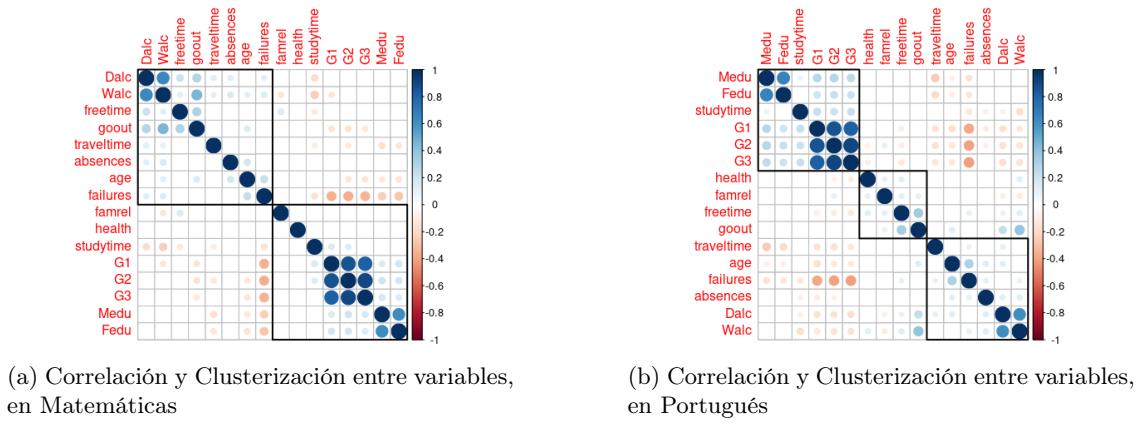


Figura 4: Correlación y Clusterización entre variables

Como se puede observar, además de buscar la correlación, aprovechamos a mostrar la clusterización jerárquica que nos ofrecía esta librería. En matemáticas se nota una separación de dos clusters, y en portugués sobre tres clusters.

A partir del gráfico anterior podemos identificar y concluir que:

- Entre  $G1$ ,  $G2$  y  $G3$  hay una alta correlación para tanto el curso de portugués y matemáticas. Esto seguramente puede ser relevante en los próximos modelos a estudiar.
- En ambos cursos, el par  $\{Dalc, Walc\}$  y  $\{Medu, Fedu\}$  tienen una alta correlación entre sí.
- Para la variable predictora de  $G3$ , ademas de las mencionadas  $G1$  y  $G2$ , vemos que hay una correlación moderada con  $\{failures\}$  para matemáticas,  $\{failures, Medu, Fedu, studytime\}$  para portugués.

### 2.3. Análisis de clusters con KMeans

En este apartado se realiza clustering utilizando KMeans para comprender ambos sets de datos. Se utiliza, en ambos casos, el método del codo para encontrar la cantidad de grupos ( $K$ ) óptima.

#### Matemática

En la siguiente imagen se reflejan los resultados de evaluar  $K$  para distintos tamaños.

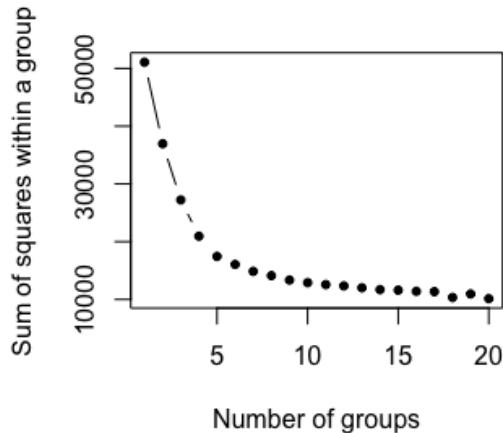


Figura 5: Método del codo para KMeans en el set de matemática.

Se aprecia que el  $K$  óptimo varía entre  $K = 4$  y  $K = 5$ . Analizando ambas divisiones de grupos, elegimos quedarnos con  $K = 4$  puesto que apreciamos una diferencia más representativa entre los grupos formados.

A continuación, por cada grupo generado, estimamos la media de cada variable (numérica y categórica con dos niveles) utilizando el promedio.

En las siguientes tablas, vamos a comparar subconjuntos de estas variables, analizando dónde radican las diferencias más relevantes entre los grupos. Se apreciará en cada tabla el nombre del grupo y el promedio de la nota final ( $G_3$ ).

En la siguiente tabla, comenzamos comparando por nota ( $G_1$ ,  $G_2$  y  $G_3$ ).

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>G1</i>	10.2692	8.823	13.7755	8.5263
<i>G2</i>	10.2051	8.4867	13.9252	7.5439
<i>G3</i>	9.9872	7.7168	14.1361	6.7544

Se aprecia:

- Un grupo ( $g_3$ ) con la mayor calificación en promedio.
- Dos grupos ( $g_2, g_4$ ) ampliamente desaprobados.
- Un grupo ( $g_1$ ) cuya nota tiende a 10, que es la estimación de la media para cualquier nota que va de 0 a 20.

Lo que intentaremos comprender es por qué  $g_3$  cuenta con mejores notas que el resto.

En la siguiente tabla apreciamos:

- $g_3$ : Los alumnos de este grupo suelen ir a una sola escuela (GP, con label 0); son menores en edad que el resto.
- Comparación  $g_1$  vs  $g_3$ : No se aprecian grandes diferencias.
- Comparación  $g_2$  vs  $g_3$ : Son muchas más mujeres (con label 0) en el segundo grupo.
- Comparación  $g_4$  vs  $g_3$ : Son muchos más hombres (con label 1) en el último grupo; son mucho más grandes en edad que el resto; se reparten equitativamente entre ambas escuelas.

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>school</i>	0.1026	0.0442	0.068	0.4035
<i>sex</i>	0.6795	0.115	0.5374	0.7368
<i>address</i>	0.7821	0.7788	0.898	0.4561
<i>famsize</i>	0.359	0.2212	0.2857	0.3333
<i>age</i>	16.7821	16.4779	16.3946	17.7895
<i>G3</i>	9.9872	7.7168	14.1361	6.7544

Comparando la educación de los padres (en la siguiente tabla) concluimos que la educación de los padres impacta en la nota final. Aunque tendremos que observar cuál es la diferencia entre  $g_1$  y  $g_3$ , porque tienen similar educación los padres pero con distintas notas.

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>Medu</i>	3.3333	2.3628	3.1361	1.7193
<i>Fedu</i>	3	2.2478	2.7687	1.7719
<i>G3</i>	9.9872	7.7168	14.1361	6.7544

En el tiempo de estudio es donde se aprecian diferencias significativas, puesto que:

- $g_1$  y  $g_4$  son los grupos que menos estudian y más tiempo libre tienen.
- $g_2$  es el grupo que más estudia, y menos tiempo libre tiene (que este tiempo libre también está relacionado con actividades extra).
- $g_3$  es un grupo con alto estudio, y bajo tiempo libre.

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>studytime</i>	1.8077	2.2832	2.1701	1.5088
<i>traveltime</i>	1.5256	1.3363	1.3197	1.8947
<i>freetime</i>	2.7179	1.9027	2.1905	2.3509
G3	9.9872	7.7168	14.1361	6.7544

Comparando los hábitos de cada grupo:

- $g_1$  es un grupo que consume mucho alcohol (en semana y fines de semana), y salen de su casa mucho más que el resto.
- $g_2$  y  $g_3$  mantienen niveles moderados de alcohol y de salidas de su casa.

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>Dalc</i>	1.5385	0.0973	0.1293	0.7018
<i>Walc</i>	2.7436	0.6283	0.7823	1.9298
<i>goout</i>	2.9872	1.8938	1.7823	2.1754
G3	9.9872	7.7168	14.1361	6.7544

En la siguiente tabla apreciamos:

- $g_3$  tiene el menor índice de fallas y de ausencias.
- $g_4$  es el único grupo que no quiere continuar con una educación superior en su totalidad. A su vez, presentan el mayor índice de fallas.
- $g_1$  presentan muchas más ausencias que el resto.
- $g_2$  presenta mayor índice de fallas que  $g_3$  y mayor ausencias.

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>failures</i>	0.2949	0.3628	0.0544	1.0526
<i>absences</i>	9.8462	5.2478	4.1497	4.9825
<i>higher</i>	1	1	1	0.6491
G3	9.9872	7.7168	14.1361	6.7544

Por último:

- $g_4$  tiene el menor acceso a internet de los cuatro grupos.
- $g_2$  presenta una mayor necesidad de apoyo escolar.

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>
<i>health</i>	2.859	2.469	2.3197	2.9123
<i>internet</i>	0.9487	0.7522	0.9252	0.5965
<i>schoolsup</i>	0.1282	0.3363	0.0136	0.0175
<b>G3</b>	<b>9.9872</b>	<b>7.7168</b>	<b>14.1361</b>	<b>6.7544</b>

### Análisis generales: Matemática

- $g_1$  es un grupo con alta educación de ambos padres, que en su soledad le garantizan una mayor nota que a los otros grupos desaprobados. Pero, este grupo tiene los peores hábitos (a nivel de alcohol, salidas y estudio), por lo que obtienen una nota final mucho menor en promedio que el mejor grupo.
- $g_2$  es un grupo conformado mayoritariamente por mujeres, con una menor educación de ambos padres. Este grupo presenta los mejores hábitos (a nivel de alcohol, salidas y estudio), pero con mayores fallas y peores notas que el mejor grupo.
- $g_3$  es el mejor grupo, que presenta una alta educación en ambos padres y con grandes hábitos. A su vez, presenta el menor índice de fallas y de ausencias.
- $g_4$  es un grupo conformado mayoritariamente por hombres. Es el único grupo que asiste a matemática que no quiere perseguir educación superior. Sus padres presentan el índice de educación más bajo. Sus hábitos no son comparables con los dos mejores grupos en este aspecto. Son el único grupo tan repartido entre ambas escuelas.

Por lo tanto, en este primer estudio, el éxito en el curso de matemática parece estar indicado por la mejor educación de los padres, buenos hábitos generales y menores fallas y ausencias. Esto se ve reflejado en la nota  $G_1$ ,  $G_2$  y, posteriormente, en la nota final  $G_3$ .

### Portugués

En esta subsección volvemos a realizar un estudio por vecinos más cercanos del data set de portugués.

En este caso, generamos distintos tamaños de grupos. Si bien el tamaño de grupo ( $K$ ) óptimo era 3 (obtenido con el método del codo), al escoger  $K = 5$  obtuvimos las relaciones más interesantes para diferenciar los grupos.

En la siguiente tabla, escogemos las variables más significativas:

	<b>g1</b>	<b>g2</b>	<b>g3</b>	<b>g4</b>	<b>g5</b>
<i>school</i>	0.8824	0.5889	0.2172	0.25	0.2441
<i>studytime</i>	1.6471	1.6944	2.0271	1.7308	2.2992
<i>failures</i>	0.7647	0.4444	0.0317	0.4038	0.0157
<i>Dalc</i>	1.1765	0.5056	0.3846	1	0.2047
<i>absences</i>	0.1176	2.3944	1.7149	11.9904	2.4882
<b>G1</b>	<b>7.1176</b>	<b>9.0167</b>	<b>12.1357</b>	<b>10.1154</b>	<b>15.1181</b>
<b>G2</b>	<b>4.2941</b>	<b>9.3556</b>	<b>12.2217</b>	<b>10.2596</b>	<b>15.622</b>
<b>G3</b>	<b>0.4118</b>	<b>9.7944</b>	<b>12.6923</b>	<b>10.6731</b>	<b>16.0787</b>

Se aprecia que hay dos casos muy distintos:

- $G_1$  y  $G_2$  en este caso son predictores muy fuertes de la nota final  $G_3$ .
- El tiempo de estudio y las fallas obtenidas son un factor diferencial para los alumnos con mayor promedio (en este set de datos importa en menor medida la educación superior de los padres). A su vez, beber menos en días de semana impacta positivamente en la nota.
- Pero, la mayor diferencia que observamos es si los alumnos asisten o no a determinado colegio.

Estudiando la diferencia entre ambos colegios, obtuvimos la siguiente tabla:

	0	1	5	6	7	8	9	10
MS	14	0	0	1	7	21	25	44
GP	1	1	1	2	3	14	10	53

Lo que apreciamos es que el colegio MS suele poner notas menores que GP. A su vez, la mayor parte de las notas 0 están concentradas en el colegio MS.

Es decir, es más probable estar en la media en el colegio GP y, con buenos índices obtenidos en la tabla anterior (sobre todo en tiempo de estudio y pocas fallas) se garantizan mejores notas. El colegio MS parece presentar mayor rigurosidad.

A continuación, estudiamos de forma separada lo que ocurre en el colegio MS. Volvemos a generar tres clusters utilizando vecinos más cercanos.

En la siguiente tabla, se aprecia que se generan tres grupos:

- Un grupo con las mejores notas: Poseen la mayor cantidad de horas de estudio y la menor cantidad de fallas. A su vez, consumen menos alcohol; tienen menos tiempo libre; por último, son los que mayores intenciones de perseguir educación superior tienen.
- Un grupo que se acerca al promedio: En este caso, apreciamos la mayor cantidad de ausencias, un mayor consumo de alcohol y mayor tiempo libre.
- Un grupo donde la nota es cercana a 0 final: La mayor particularidad de este grupo es que  $G_1$  y  $G_2$  no son 0, pero son notas relacionadas a un insuficiente (de 1 a 3 en escuelas argentinas). Son los alumnos mayores en edad, con menores intenciones de continuar con una educación superior, con menor tiempo de estudio y mayor consumo de alcohol.

	g1	g2	g3
age	16.9078	16.6571	17.8
higher	0.7872	0.9714	0.6667
studytime	1.6738	2	1.6667
failures	0.4255	0.0143	0.6667
freetime	2.303	1.9857	2.6667
Dalc	0.6383	0.3	1.0667
Walc	1.461	0.9143	1.6667
absences	3.5481	1.2857	0.1333
$G_1$	8.9149	13.8714	6.6667
$G_2$	9.4184	14.1429	3.6
$G_3$	9.7589	14.6429	0.4

### Análisis generales: Portugués

Luego de un breve estudio inicial, obtenemos como primeras relaciones que:

- El éxito en el curso de portugués parece estar indicado por las horas de estudio y por las fallas en los exámenes.

- Los colegios tienen distinta forma de calificación, sobre todo para los alumnos con peores notas.
- Es muy importante que las notas  $G_1$  y  $G_2$  sean superiores a un mínimo en el colegio MS, sino la nota final del alumno suele ser 0 (que podemos entender como que el alumno no regularizó el curso).

### 3. Validación de supuestos

El modelo ajustado para analizar y validar los supuestos es el que contiene todos los predictores.

$$\hat{G}_3 = \beta_0 + \beta_1 \times age + \cdots + \beta_p \times G_2$$

Los supuestos del presente modelo lineal son:

1. (Linealidad) Sea  $\epsilon_i$  el error aleatorio para  $G_{3i}$ ,

$$E[\epsilon_i] = 0, i = 1, \dots, n$$

2. (Homocedasticidad)  $VAR[\epsilon_1] = \dots = VAR[\epsilon_n] = \sigma^2$ .

3. Los errores aleatorios son independientes entre sí y no correlacionados con las covariables  $X_i$ .

4.  $\epsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \dots, n$ .

En *Cuadro 1* se presenta un resumen del modelo ajustado. Además, se presenta un intervalo de confianza para las covariables  $G_1$  y  $G_2$  que son las que presentan la mayor correlación con la variable a predecir.

	Mat	Por
G1	0.038** (0.001, 0.076)	0.178*** (0.115, 0.240)
G2	0.435*** (0.396, 0.474)	0.758*** (0.693, 0.823)
MSE	0.2	1.45
p-value	0.01	0.01
Observations	285	507
R-squared	0.941	0.911
Adjusted R-squared	0.931	0.903
Residual standard error	0.406 (df = 243)	0.843 (df = 465)
F statistic	94.620*** (df = 41; 243)	115.632*** (df = 41; 465)

Notes: \*\*\*p < .01; \*\*p < .05; \*p < .1

Cuadro 5: Regresión completa: Matemática y Portugués

Se observa que ambas regresiones son significativas ( $p - \text{valor} = 0,01$ ). Además, la estimación del error cuadrático medio es mayor para el set de datos de matemática que para el de portugués.

#### Supuesto de linealidad

La estimación de los errores aleatorios se realizan con los residuos de cada modelo ajustado. A su vez, la estimación de la media de los residuos se obtiene a partir del promedio.

En la Figura 6 se aprecia un gráfico de los residuos para cada modelo, indicando el promedio de los mismos.

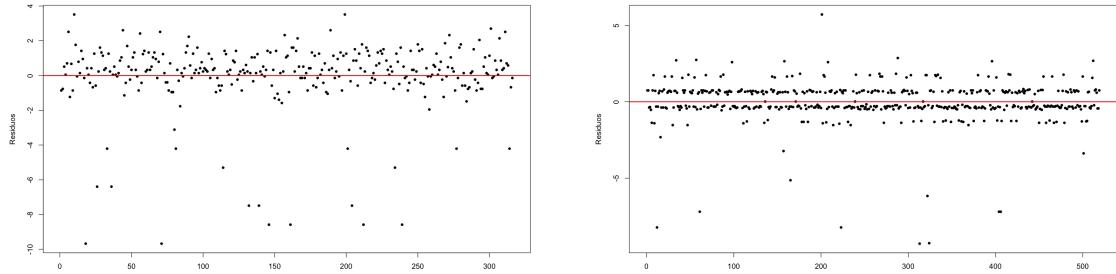


Figura 6: Gráficos del promedio de los residuos (estimación de la esperanza de los errores aleatorios) para *a.* Matemática, *b.* Portugués.

Puesto que ambos promedios tienden a 0, se valida el primer supuesto.

### Homocedasticidad

Para poder continuar utilizando los test de hipótesis de significancia de la regresión y los intervalos (de confianza, predicción) se debe validar el supuesto de homocedasticidad.

Para ello, como se aprecia en la Figura 7 se grafican los valores estimados de  $G_3$  contra los residuos.

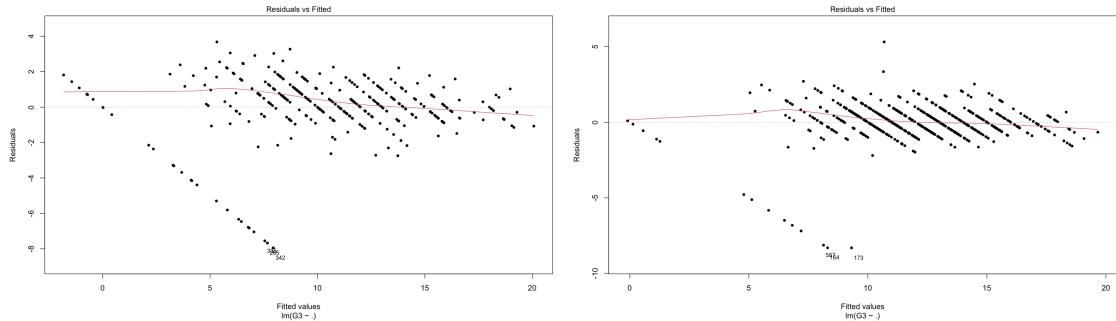


Figura 7: Valores estimados de  $G_3$  Vs. residuos para el ajuste realizado para *a.* Matemática, *b.* Portugués.

Puesto que el gráfico no muestra ninguna tendencia (por ejemplo: que la varianza aumente a medida que aumenta el valor de  $G_3$ ) se concluye que el supuesto es válido.

### Multicolinealidad

Se asume que los datos son independientes entre sí (porque en otro caso no podríamos realizar estudio). Luego, verificamos que las covariables no presenten multicolinealidad para que no afecten las estimaciones.

Para validar este supuesto realizamos como test: Dublin-Watson (cuya hipótesis nula es que las variables no presentan autocorrelación). En la Tabla 6 se observa el resultado del test para cada set de datos.

Ambos tests realizados no tienen información suficiente para rechazar la hipótesis nula, por lo tanto se concluye que el supuesto es válido.

### Supuesto de normalidad

Como análisis de normalidad, se realizaron los siguientes estudios:

	Mat	Por
Autocorrelación	-0.011	0.028
DW statistic	2.012	1.938
<b>p-value</b>	<b>0.926</b>	<b>0.474</b>

Cuadro 6: Test DW

- Histograma de los residuos para ambos modelos (Figura 8). Se superpone la función de densidad normal estándar en el gráfico (color azul).
- Gráfico de normalidad, *QQ-plot* (Figura 9).
- Test de Shapiro-Wilk (Tabla 7).

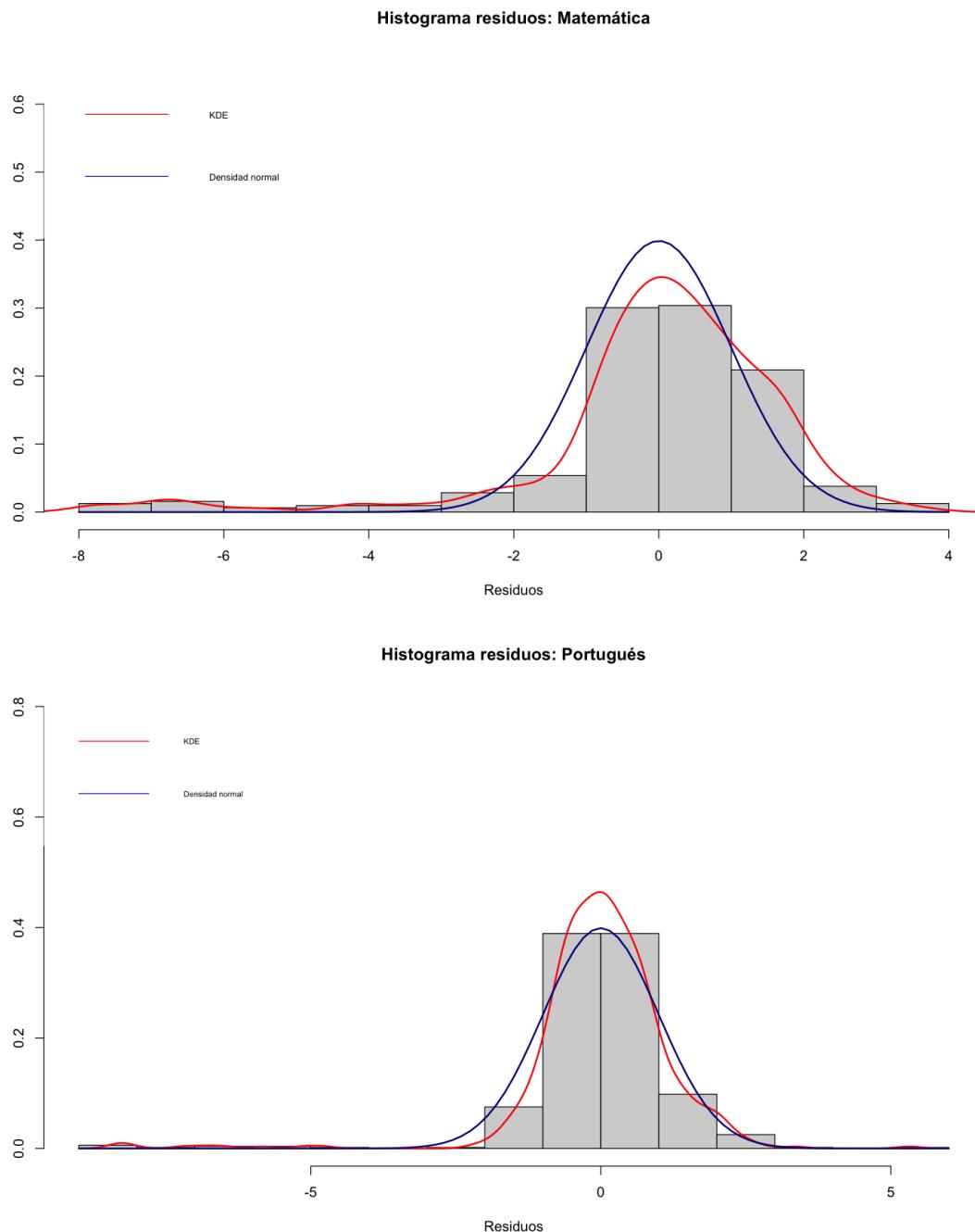


Figura 8: Estimación de distribución de residuos de la regresión ajustada para *a.* Matemática; *b.* Portugués.

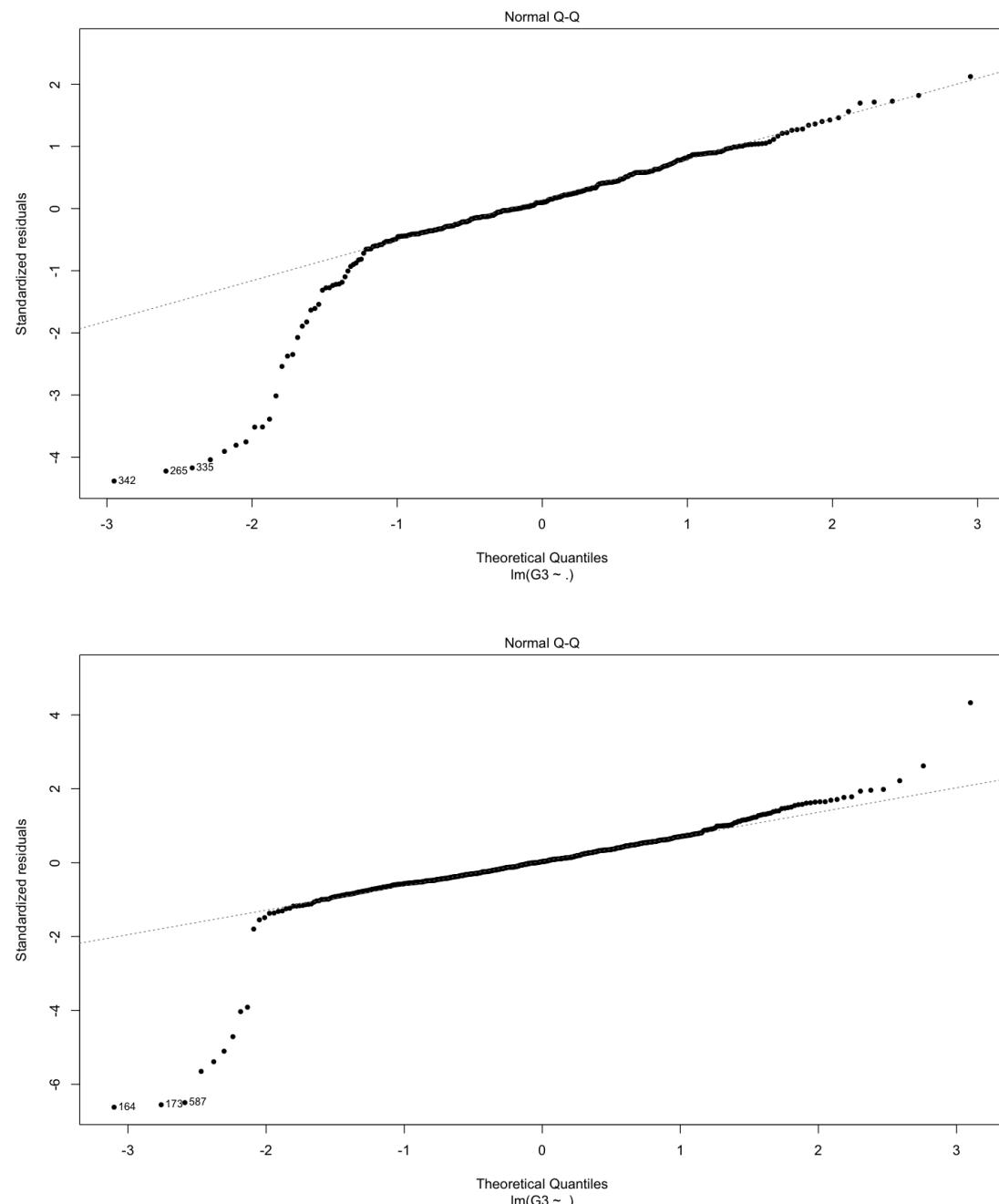


Figura 9: *QQ-plot* de residuos de la regresión ajustada para *a.* Matemática; *b.* Portugués.

	Mat	Por
W	0.827	0.777
p-value	<b>0.01</b>	<b>0.01</b>

Cuadro 7: Test Shapiro-Wilk

Analizando los gráficos y el test realizado, podemos concluir que no se cumple el supuesto de normalidad, por lo tanto deberemos proceder a identificar valores atípicos y realizar una transformación de nuestra variable a predecir, de forma de garantizar la normalidad.

### Identificación de datos atípicos

Realizamos dos estudios de outliers:

1. Eliminar uno por uno de los outliers utilizando resudios estudentizados y distancia de Cook.
2. Entender quiénes son estos outliers.

(1) trae el problema de no entender qué tipo de alumno se está quitando. Puesto que el objetivo del trabajo es entender qué features llevan a los alumnos a peores o mejores notas, se decidió descartar (1) y escoger (2).

Para esto, se analiza qué ocurre cuando se ajusta un modelo utilizando únicamente el predictor con mayor correlación ( $G_2$ , que presenta una correlación prácticamente lineal con  $G_3$ ).

El modelo ajustado para cada set de datos es:  $\hat{G}_3 = \beta_0 + \beta_1 \times G_2$ .

En la Figura 10 se aprecia que los datos atípicos se producen cuando  $G_3 = 0$  y  $G_2 \neq 0$ .

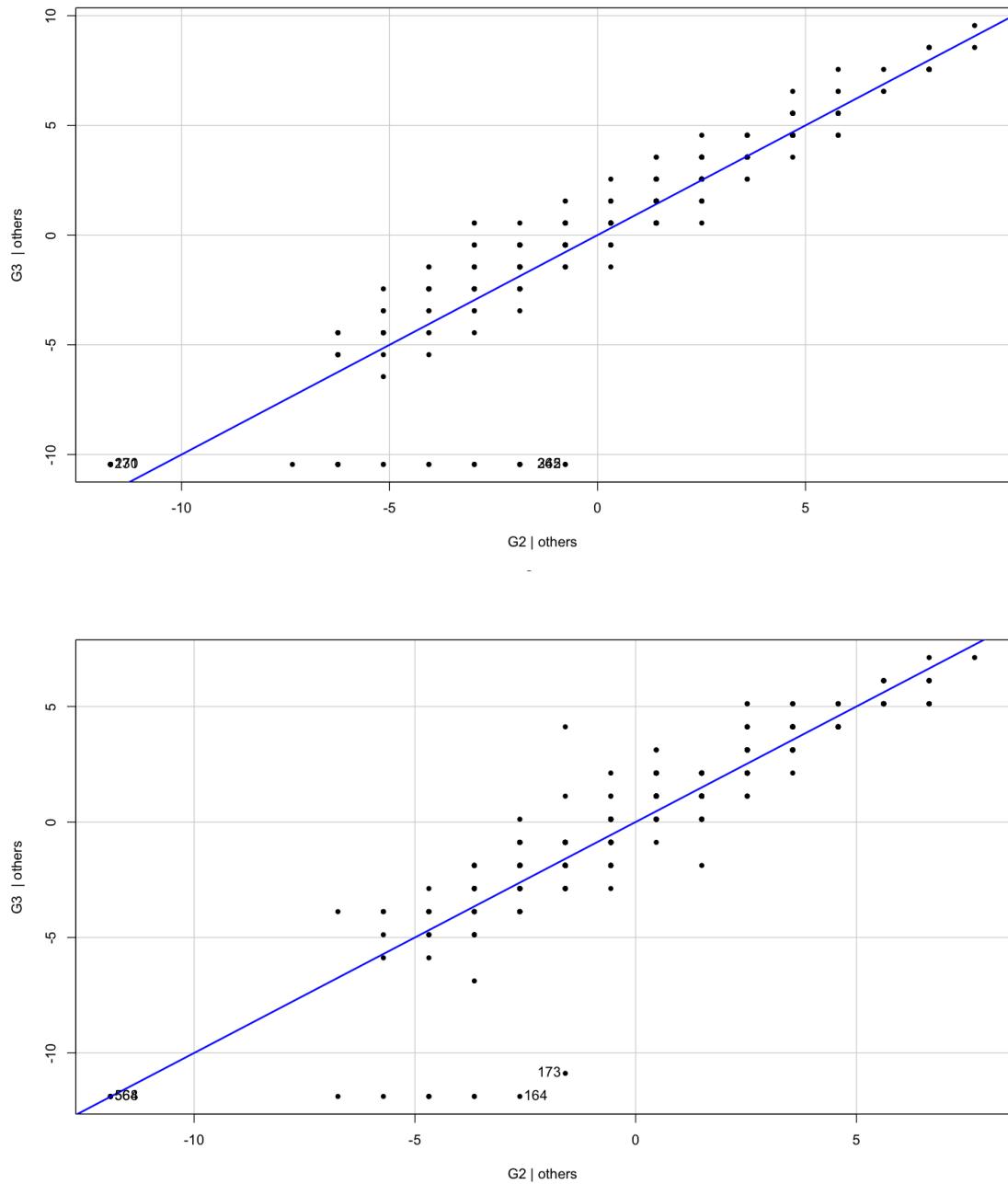


Figura 10: Regresión ajustada para *a.* Matemática; *b.* Portugués.

Analizando ambos set de datos, estos alumnos son aquellos que tienen notas en primer y segundo término distintas de 0, pero la nota final es 0. Es decir, a su vez estos alumnos no presentan características diferentes que los demás en cuanto a ausencias, fallas en exámenes o demás covariables.

En Apéndice A se justifica por qué este grupo pertenece a un grupo de alumnos atípicos. En conclusión, son removidos del set de datos aquellos alumnos con  $G_3 = 0$  y  $G_2 \neq 0$ .

### Transformación de Box-Cox

Una vez identificados los datos atípicos, se realizó una transformación de Box-Cox (Figuras 11 y 12) para asegurar el supuesto de normalidad de los datos.

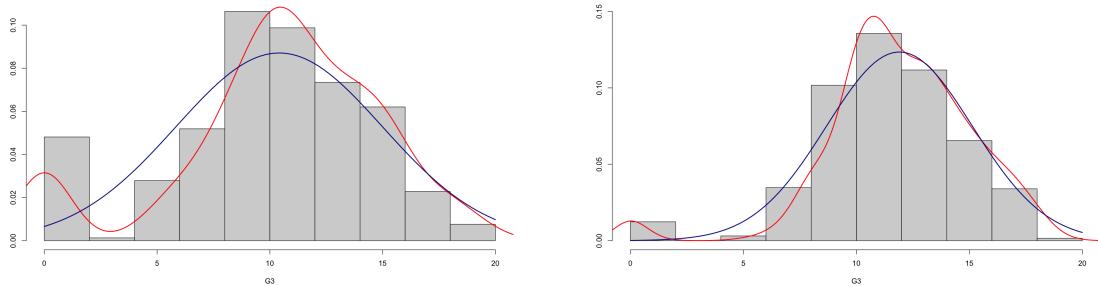


Figura 11: *a.* Set de Matemática sin transformar y con outliers; *b.* Set sin outliers y transformado.

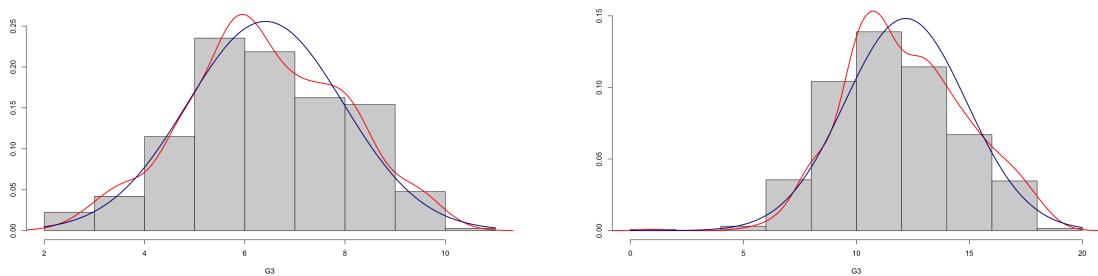


Figura 12: *a.* Set de Portugués sin transformar y con outliers; *b.* Set sin outliers y transformado.

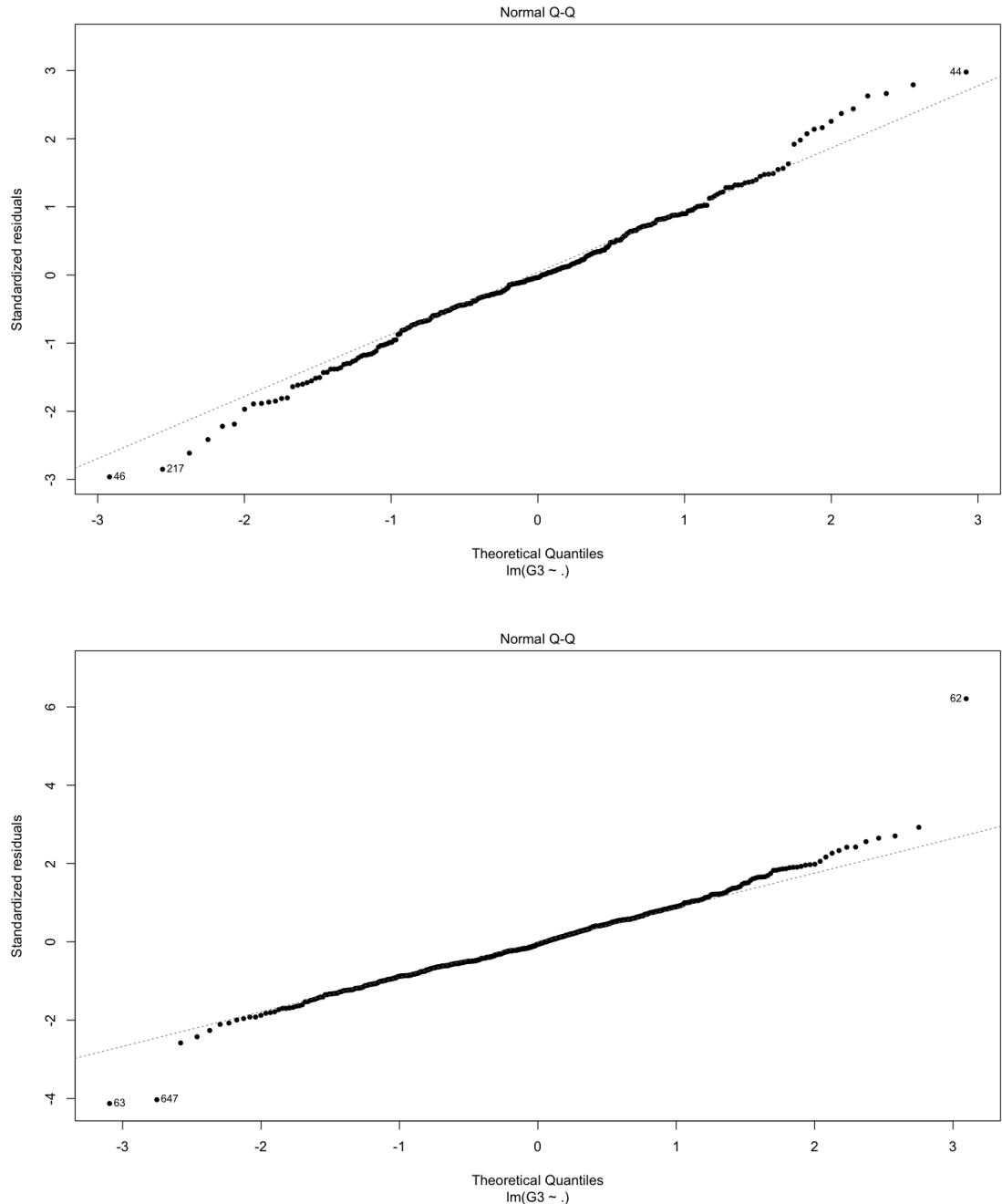


Figura 13: *QQ-plot* para Set de *a.* Matemática; *b.* Portugués. Ambos sin outliers y con la transformación de Box-Cox aplicada.

Luego de realizar las transformaciones, analizando el gráfico de normalidad (Figura 13) se aprecia que se sigue contando con datos atípicos, aunque esta vez con colas menos pesadas que en el primer análisis. Puesto que en la siguiente sección se estudia el modelo lineal y se realizan escalamientos sobre distintos predictores significativos, se decidió dejar estos datos, entendiendo que son alumnos con relaciones importantes para la conclusión final.

En la Tabla 8 se aprecia un nuevo test de Shapiro-Wilk. En este caso, el set de matemática valida el supuesto de normalidad, pero el set de portugués no si  $p - valor < 0,02$ . De igual forma,

seguiremos adelante, pese a que algunos datos atípicos en el set de portugués no garantizan la normalidad.

	Mat	Por
W	0.995	0.993
<b>p-value</b>	<b>0.255</b>	<b>0.02</b>

Cuadro 8: Test Shapiro-Wilk

## 4. Modelo lineal

En esta sección se utiliza el modelo de regresión lineal para predecir la nota final ( $G_3$ ), con el objetivo de construir modelos que ayuden a entender cuáles son las variables significativas para predecir matemática y portugués.

Puesto que, como se observó en la sección de análisis inicial,  $G_1$  y  $G_2$  tienen una relación cercana a la lineal con  $G_3$ , se utilizan dos configuraciones de búsqueda de modelos (para cada sets de datos):

1. Búsqueda de modelos con todos los predictores.
2. Búsqueda de modelos sin  $G_1$  y  $G_2$ .

Se utilizaron desde el primer momento ambos sets de datos sin los datos atípicos de aquellos alumnos sin nota final, así como la transformación realizada a la variable a predecir  $G_3$ .

### 4.1. Búsqueda de modelos con todos los predictores

En la Tabla 9 y 10 se presenta un resumen de los modelos ajustados para ambos sets de datos. Se presentan como medida de comparación:

- $ECM$  estimado.
- $R^2$  ajustado.
- Cantidad de predictores con  $p - valor < 0.01$ .

Cabe destacar que el primer modelo en ambos sets de datos corresponde al promedio de la variable a predecir  $G_3$ , para tener una medida de comparación del peor modelo posible (el cual sí o sí el resto tendrá que mejorar para poder ser considerado).

	Modelo	$ECM$	$R^2adj$	Predictores ( $p - value < 0.01$ )
1	Promedio	2.337	-	0
2	Solo $G_2$	0.167	0.927	1
3	Completo	0.190	0.941	1
4	Backward selection	0.176	0.938	3
5	Forward Selection	0.174	0.937	2
6	Backward con AIC	0.176	0.938	3
7	Lasso	0.179	0.938	2

Cuadro 9: Comparación de modelos ajustados para set de datos: Matemática

	Modelo	$ECM$	$R^2adj$	Predictores ( $p - value < 0.01$ )
1	Promedio	8.679	-	0
2	Solo $G_2$	0.824	0.862	1
3	Completo	0.890	0.887	4
4	Backward selection	0.732	0.869	2
5	Backward con AIC	0.818	0.881	3
6	Lasso	0.857	0.885	4

Cuadro 10: Comparación de modelos ajustados para set de datos: Portugués

Se muestra de forma resumida los resultados más importantes de cada modelo.

### Ajuste solo con $G_2$ como predictor

Las Figuras 14 y 15 muestran un resumen del modelo ajustado dado, para ambos sets de datos, por:

$$\hat{G}_3 = \beta_0 + \beta_1 \times G_2$$

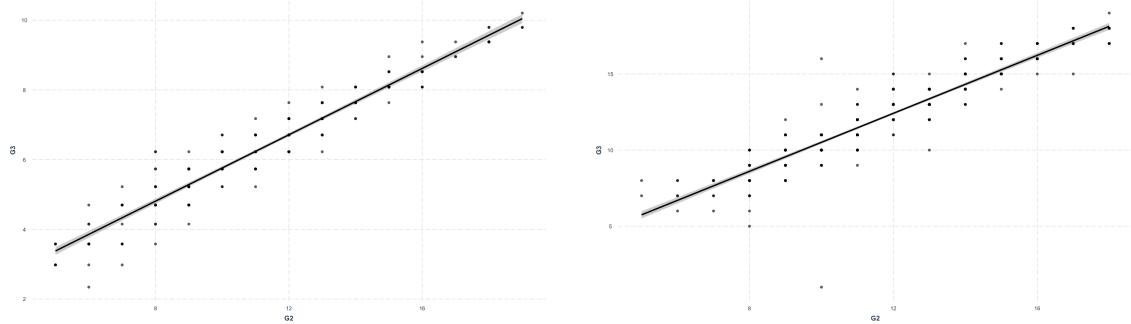


Figura 14: Ajuste realizado únicamente con  $G_2$ , junto con sus intervalos de confianza (95 %) para a. Matemática; b. Portugués.

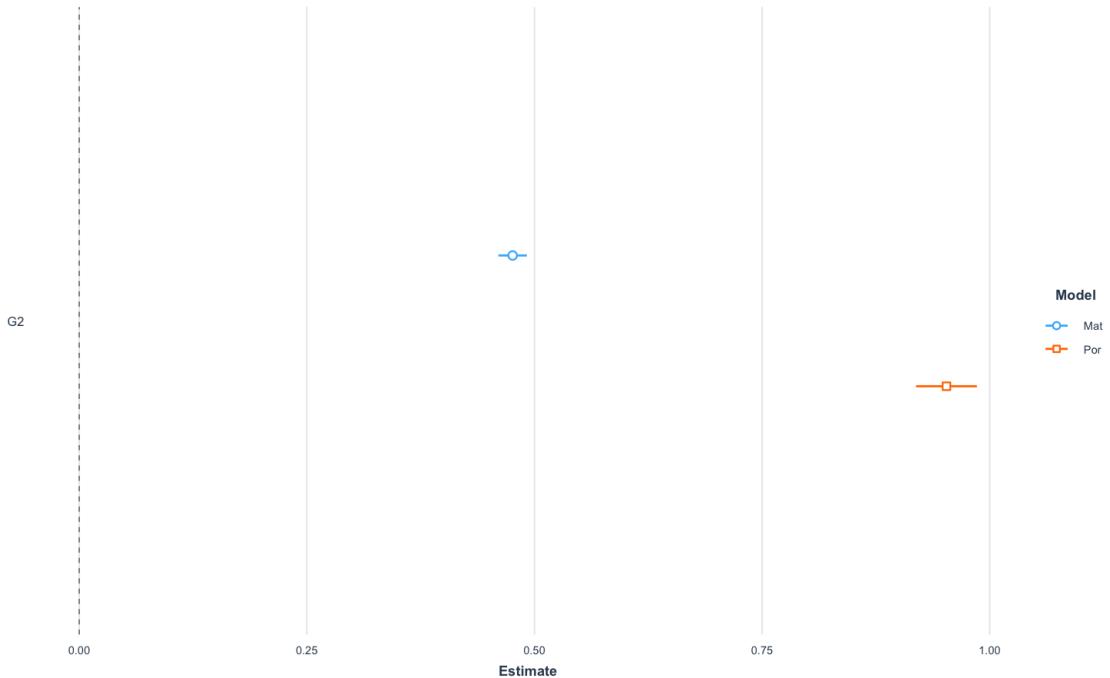


Figura 15: Estimación de  $\beta_1$  con su intervalo de confianza (95 %) a. Matemática; b. Portugués.

En la Figura 15 se observa la estimación de  $\beta_1$  en cada set de datos, junto con su intervalo de confianza. Este método será de utilidad para entender cómo se están prediciendo distintos predictores en los siguientes modelos, y para poder comparar modelos entre sí.

En este caso se aprecia que  $G_2$  en el set de portugués es cercano a 1, y en matemática a 0.5, ambos con intervalos de confianza acotados.

### Ajuste realizado con todos los predictores

El ajuste realizado (que se resume en las Figuras 16 y 17) es:

$$\hat{G}_3 = \beta_0 + \beta_1 \times \text{age} + \cdots + \beta_p \times G_2$$

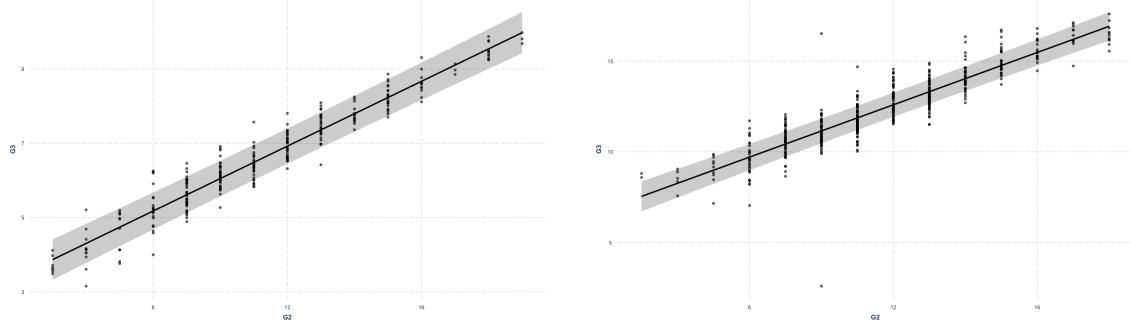


Figura 16: Ajuste con todos los predictores, junto con sus intervalos de confianza (95 %) para *a.* Matemática; *b.* Portugués.

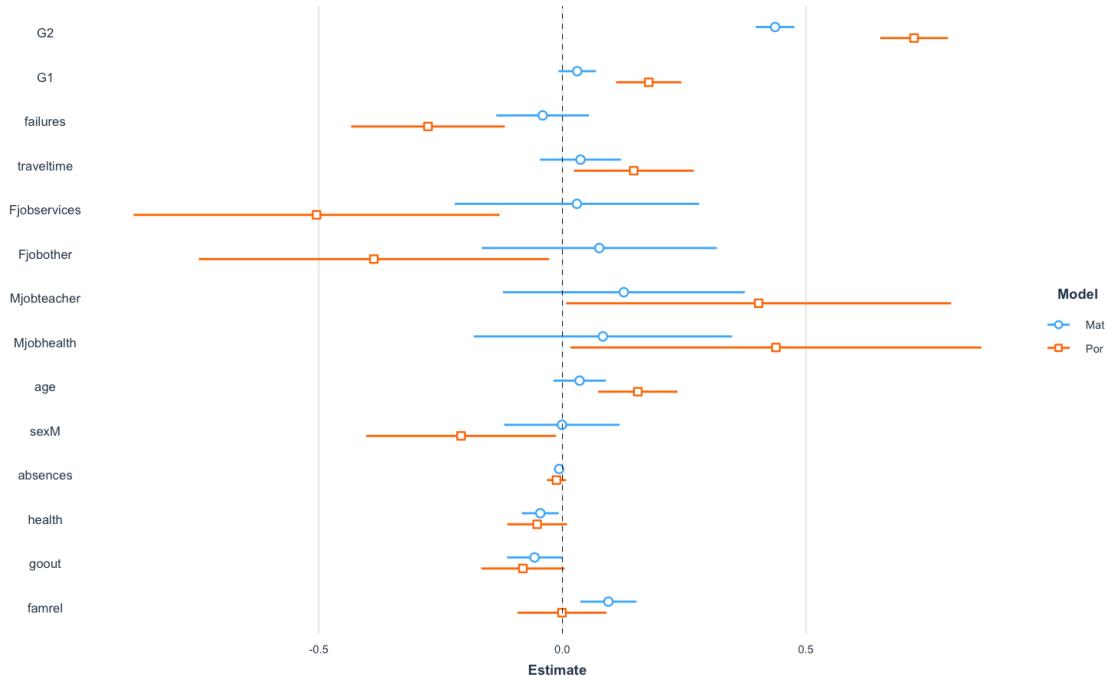


Figura 17: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués.

Se aprecia que para el set de datos de matemática la mayor parte de los predictores son cercanos a 0 (con una confianza del 95 %), a excepción de  $G_2$ .

En el set de portugués la estimación de los  $\beta_i, i = 1, \dots, p$  ocupa distintos puntos, pero con intervalos de confianza mucho más amplios.

### Backward selection

Utilizando selección de modelos hacia atrás con todos los predictores (utilizando el  $RMSE$  con validación cruzada como medida para quitar un predictor), se tienen que los modelos obtenidos (se aprecia la cantidad de features en la Figura 21) son:

Para matemática, los predictores escogidos son: *address, Mjob, nursery, famreal, goout, health, absences, G<sub>1</sub> y G<sub>2</sub>*. Para portugués: *G<sub>1</sub> y G<sub>2</sub>*.

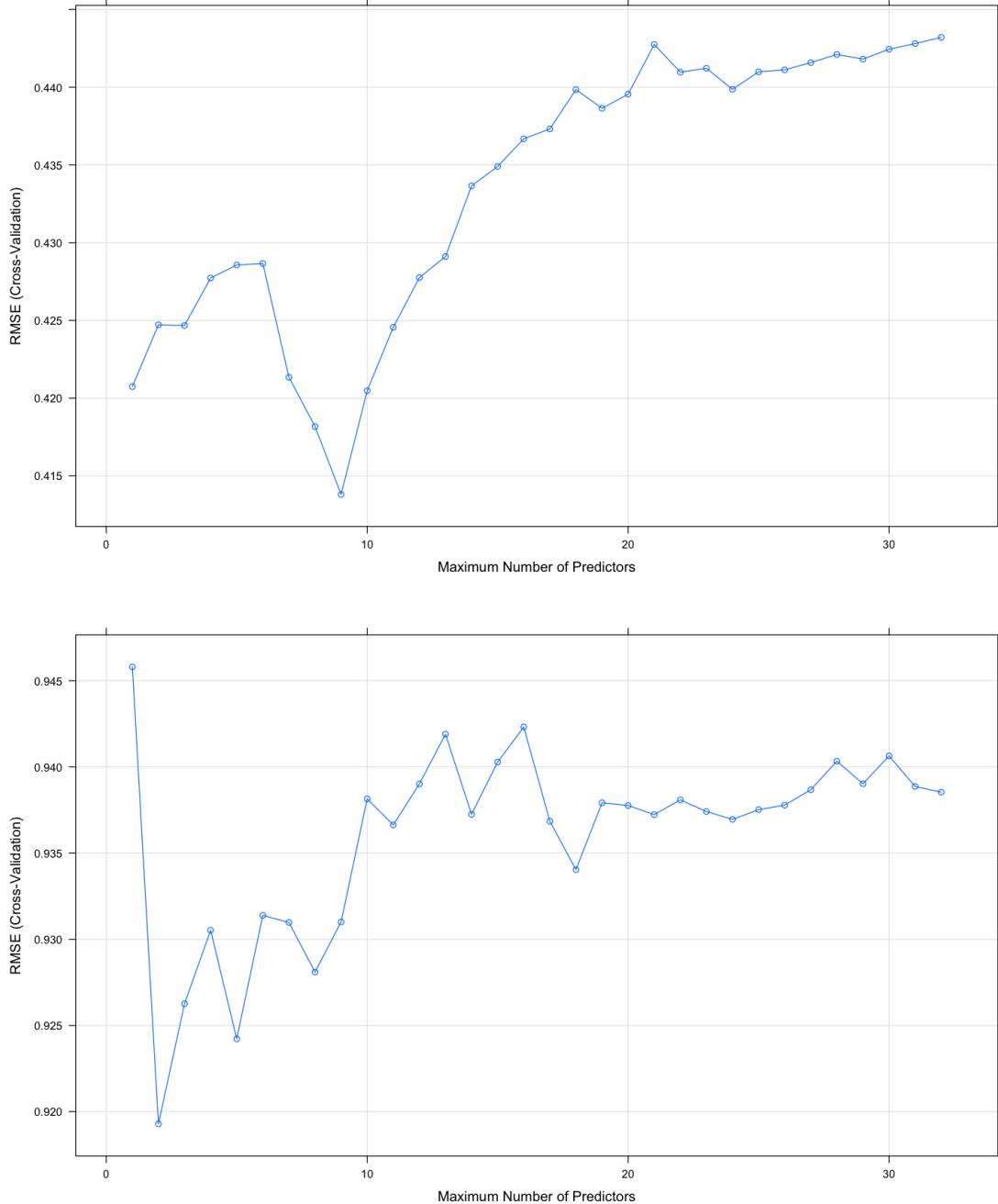


Figura 18: Comparación del tamaño de predictores óptimo para Backward Selection, usando  $RMSE$ , para a. Matemática; b. Portugués.

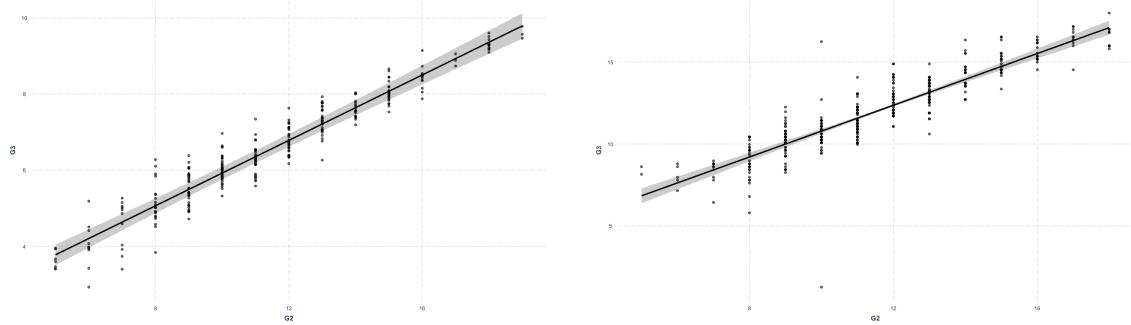


Figura 19: Ajuste del modelo de Backward Selection, junto con sus intervalos de confianza (95 %) para *a.* Matemática; *b.* Portugués.

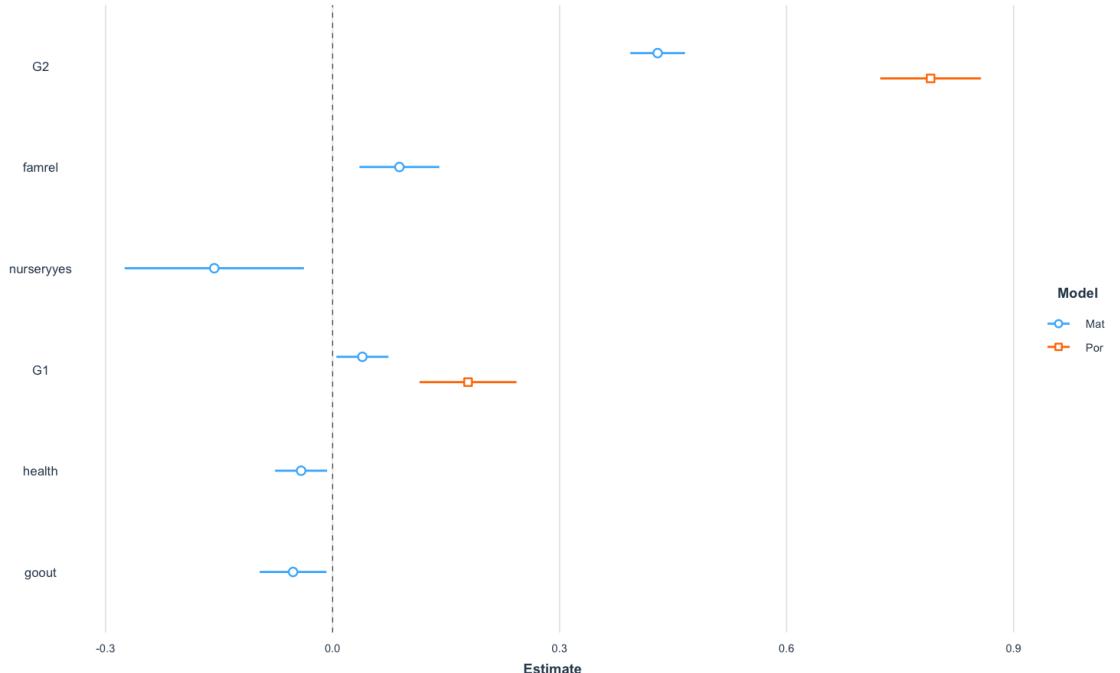


Figura 20: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués.

### Backward selection

Utilizando selección de variables hacia atrás con todos los predictores (utilizando el *RMSE* con validación cruzada como medida para quitar un predictor), se tienen que los modelos obtenidos (se aprecia la cantidad de features en la Figura 21) son:

- Para matemática, los predictores escogidos son: *address*, *Mjob*, *nursery*, *famreal*, *goout*, *health*, *absences*, *G<sub>1</sub>* y *G<sub>2</sub>*.
- Para portugués: *G<sub>1</sub>* y *G<sub>2</sub>*.

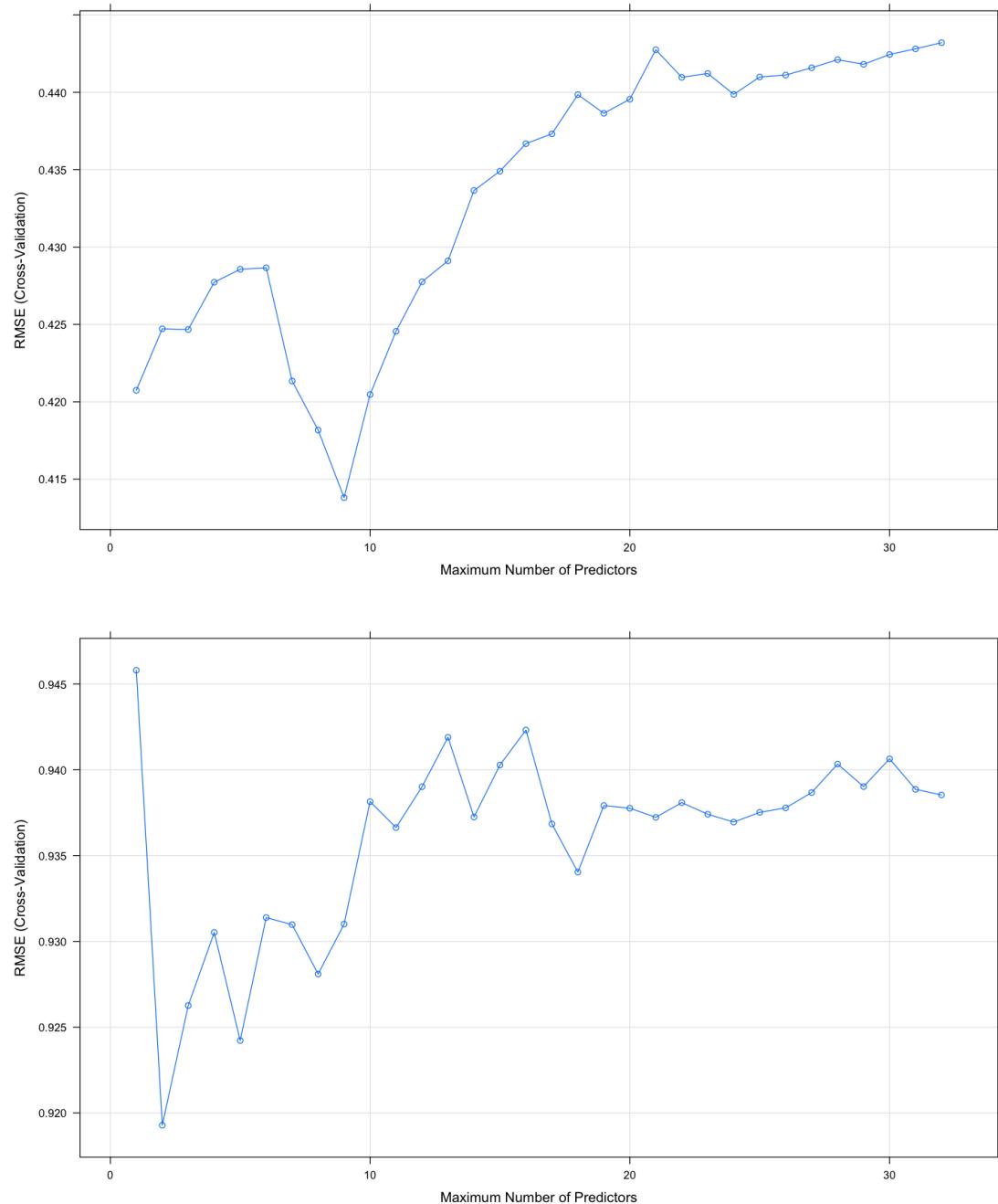


Figura 21: Comparación del tamaño de predictores óptimo para Backward Selection, usando RMSE, para *a.* Matemática; *b.* Portugués.

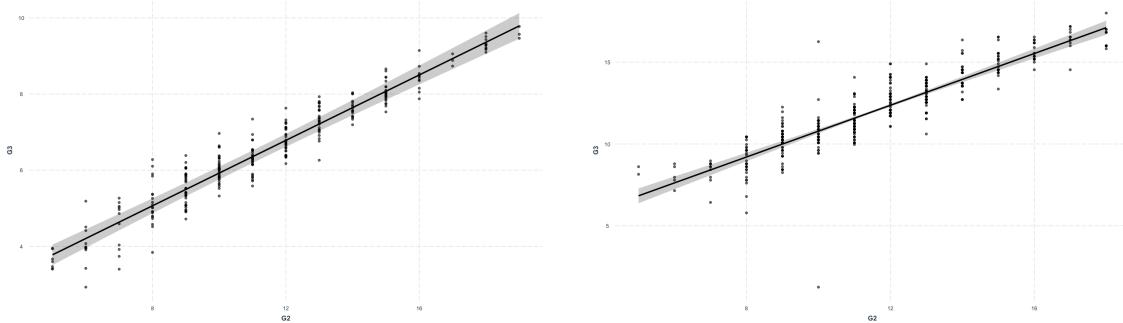


Figura 22: Ajuste del modelo de Backward Selection, junto con sus intervalos de confianza (95 %) para *a.* Matemática; *b.* Portugués.

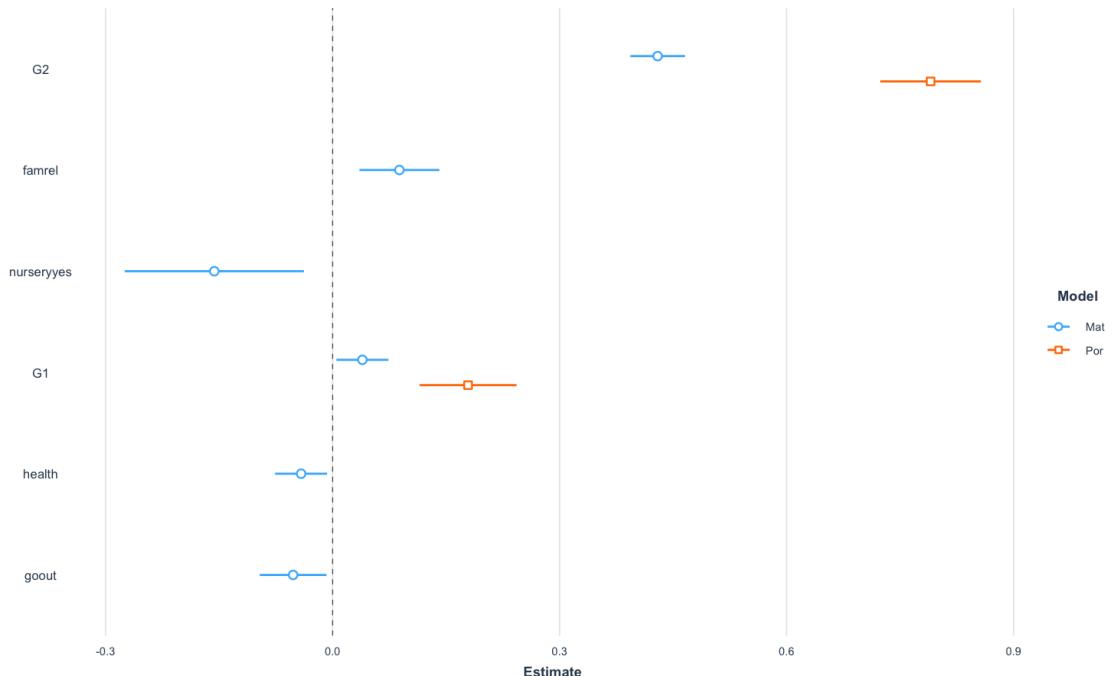


Figura 23: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués.

También se utilizó como selección de modelos: Forward-Selection y Backward-Selection cambiando la medida de bondad de ajuste (AIC), pero ambas obtienen resultados similares.

### Método de Regularización: Lasso

En las Figuras 24 y 25 se observa el resultado del mejor modelo obtenido por Lasso. Las variables que no son puestas en 0 para cada set de dato son:

- Matemática: *famrel*, *health*, *goout* y *G<sub>2</sub>*
- Portugués: *sex*, *age*, *Mjob*, *Fjob*, *reason*, *traveltime*, *failures*, *activities*, *higher*, *freetime*, *goout*, *Dalc*, *Walc*, *health*, *absences*, *G<sub>1</sub>* y *G<sub>2</sub>*.

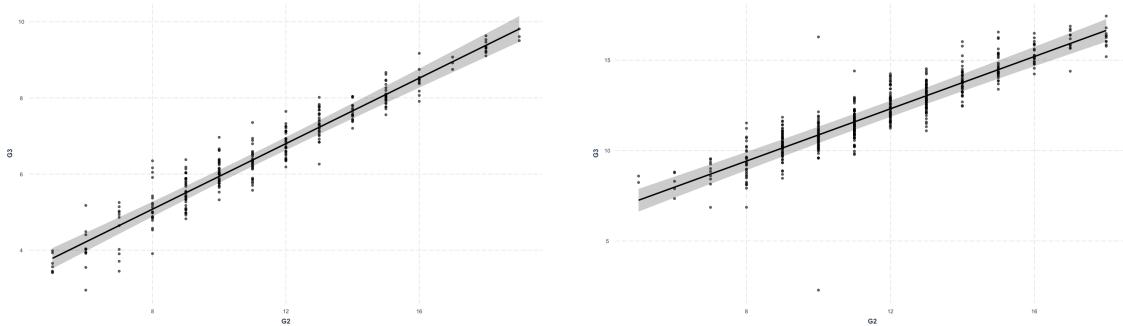


Figura 24: Ajuste del modelo de Lasso, junto con sus intervalos de confianza (95 %) para a. Matemática; b. Portugués.

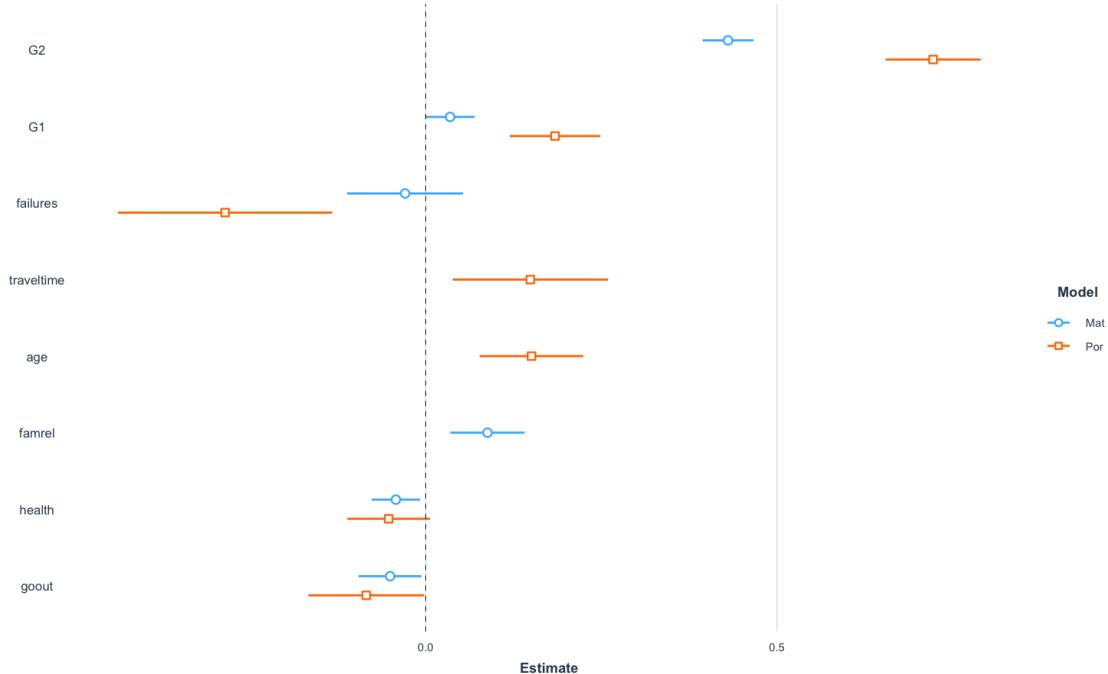


Figura 25: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués.

Se utilizó únicamente Lasso como método de regularización puesto que Ridge no permite dar cuenta de qué variables deben ser 0, y lo que se busca es comprender la influencia de los predictores.

#### 4.2. Búsqueda de modelos sin $G_1$ y $G_2$

En las Tablas 11 y 12 se resumen los modelos utilizados, quitando las notas para predecir a la nota final (ya que cuentan con una gran correlación lineal).

Se aprecia que:

- La cantidad de predictores significativos para predecir aumenta notablemente.
- El  $ECM$  estimado aumenta y la varianza explicada por cada modelo es mucho menor.

En las Figuras 26 a 29 se presenta un breve resumen de la información aportada por cada modelo construido.

	Modelo	$ECM$	$R^2 adj$	Predictores ( $p - value < 0.01$ )
1	Promedio	2.337	0.000	0
2	Completo	1.727	0.359	5
3	Backward selection	1.815	0.314	8
4	Backward con AIC	1.787	0.339	6
5	Lasso	1.729	0.358	5

Cuadro 11: Comparación de modelos ajustados para set de datos: Portugués

	Modelo	$ECM$	$R^2 adj$	Predictores ( $p - value < 0.01$ )
1	Promedio	8.679	0.000	0
2	Completo	6.298	0.394	6
3	Backward selection	6.287	0.394	6
4	Backward con AIC	6.214	0.383	7
5	Lasso	6.251	0.388	6

Cuadro 12: Comparación de modelos ajustados para set de datos: Portugués

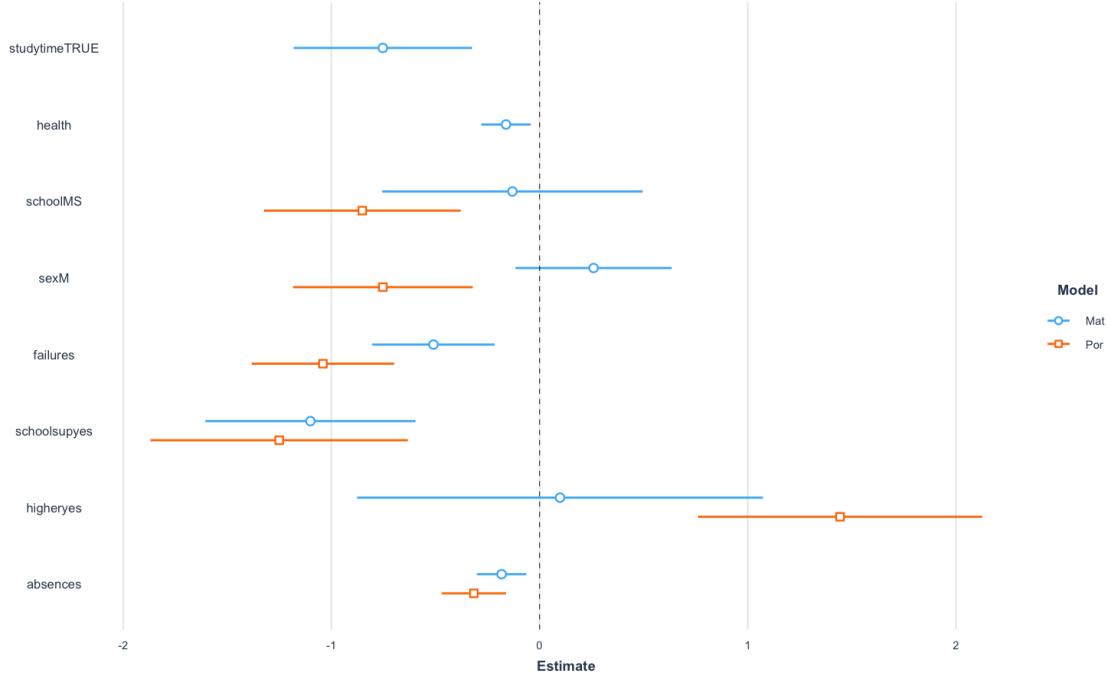


Figura 26: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués, para cada modelo construido sin  $G_1$  y  $G_2$ .

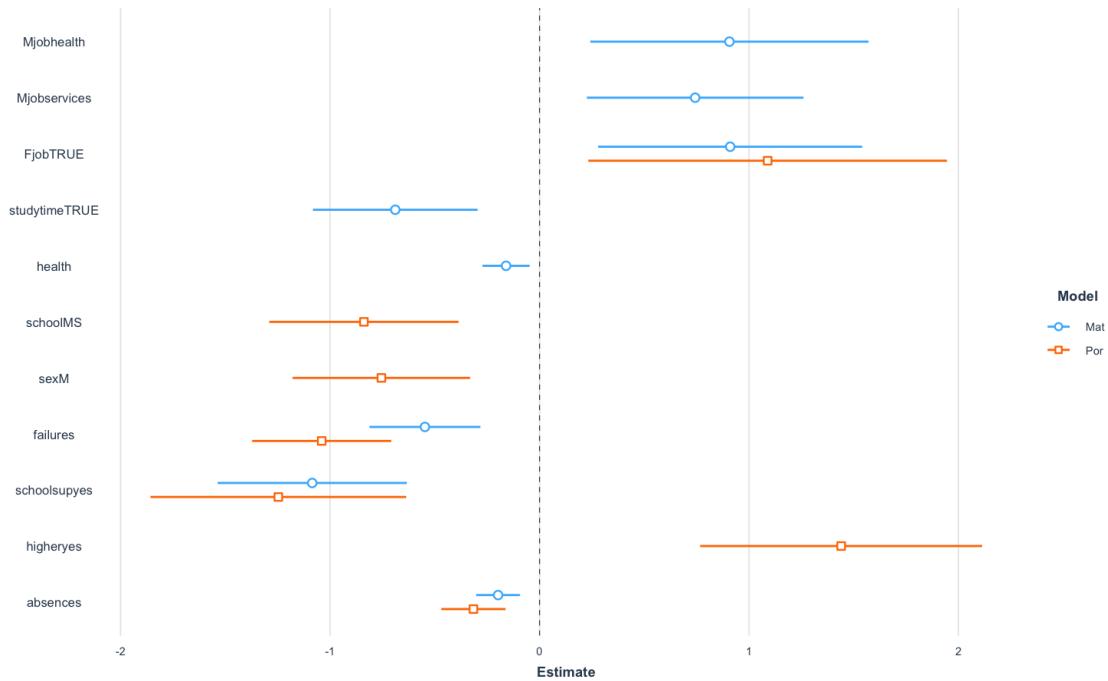


Figura 27: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués, para cada modelo construido sin  $G_1$  y  $G_2$ .

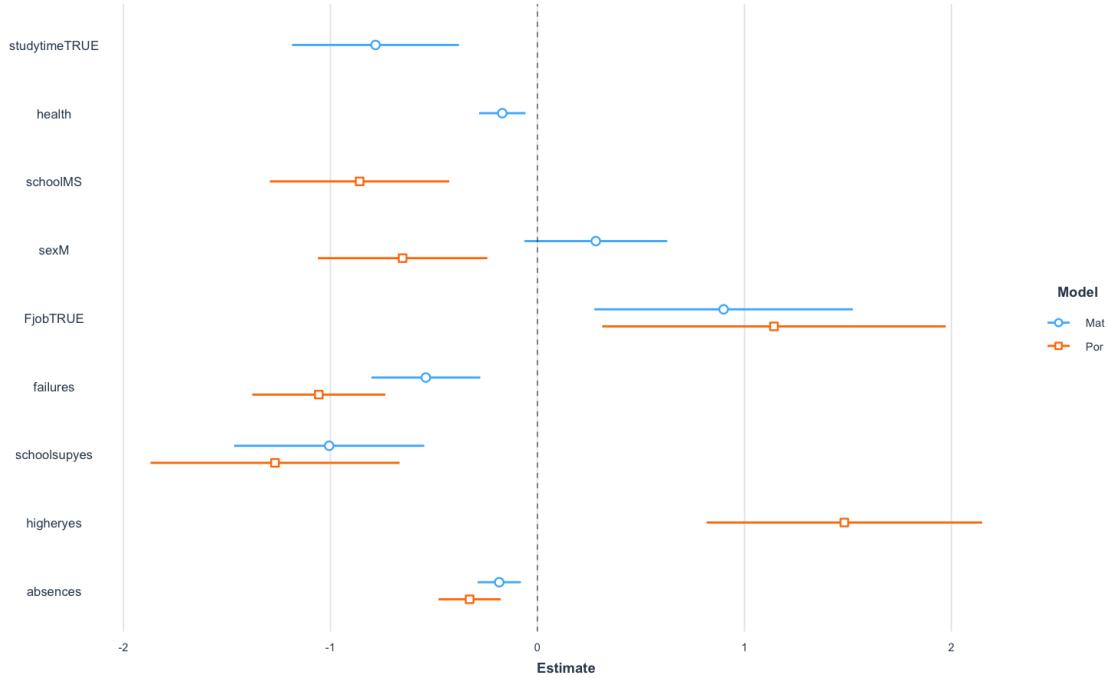


Figura 28: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués, para cada modelo construido sin  $G_1$  y  $G_2$ .

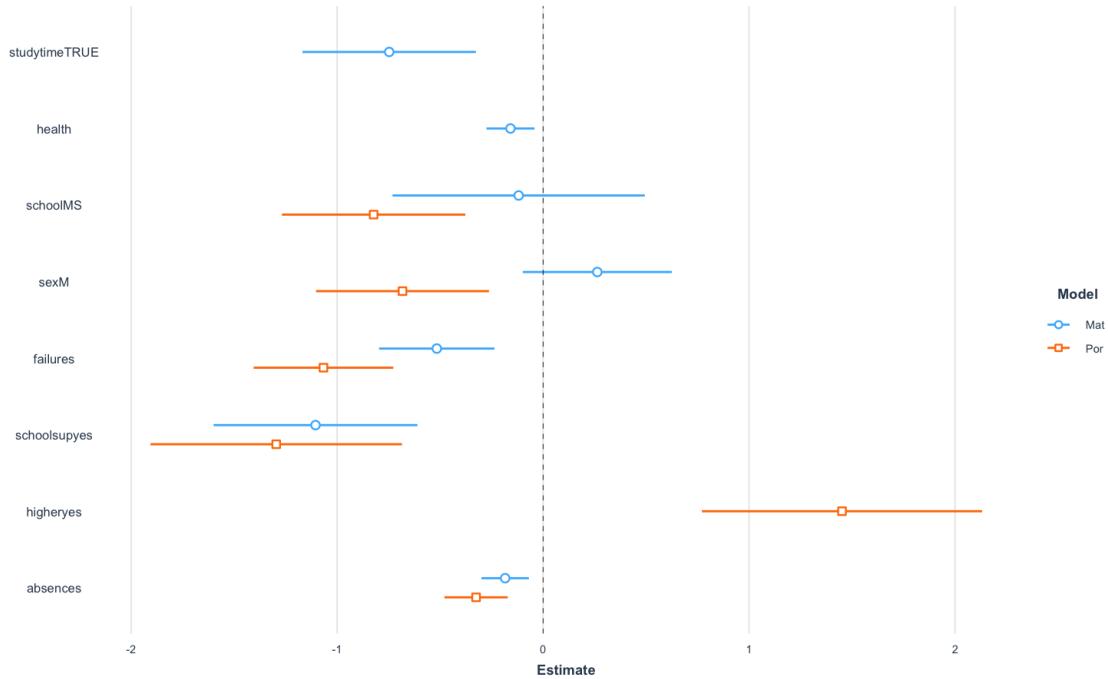


Figura 29: Comparación de la estimación de cada  $\beta_i$  significativo con su intervalo de confianza (95 %) entre Matemática y Portugués, para cada modelo construido sin  $G_1$  y  $G_2$ .

#### 4.3. Modelos escogidos

Los modelos escogidos (a partir de analizar quién minimiza el  $ECM$  estimado y maximiza  $R^2$  ajustado) son:

##### Matemática: Features completos

- Ajuste realizado solo con  $G_2$ .
- Backward-selection.

##### Matemática: Sin $G_1$ y $G_2$

- Backward-selection utilizando AIC.
- Lasso.

##### Portugués: Features completos

- Ajuste realizado solo con  $G_2$ .
- Backward-selection.

##### Portugués: Sin $G_1$ y $G_2$

- Backward-selection utilizando AIC.

Esto lo podemos resumir en la siguiente tabla:

Cuadro 13: Resumen sobre los mejores modelos

	Matemáticas				Portugués		
	Features completos		Sin G1 y G2		Features completos		Sin G1 y G2
	Solo G2	B.S.	B.S. AIC	Lasso	Solo G2	B.S.	B.S. AIC
absences							
famrel							
failures							
G1							
G2							
health							
higher							
Fjob							
nursery							
sex							
school							
schoolsup							
studytime							
<b>ECM</b>	0.167	0.176	1.787	1.729	0.824	0.732	6.214
<i>R</i> <sup>2</sup> adj	0.927	0.938	0.339	0.358	0.862	0.869	0.383

En el cual podemos observar aquellas celdas marcadas en **verde**, indicando las features mas significativas consideradas por los mejores modelos elegidos.

## 5. Árboles y Ensambles

En esta sección se desarrollan los modelos de regresión basados en árboles. De la misma forma que en la sección anterior, se tendrán dos configuraciones posibles por cada set de datos: Todos los predictores, y removiendo  $G_1, G_2$ .

### 5.1. Búsqueda de modelos con todos los predictores

En la Tabla 14 se aprecia la estimación de la estimación del  $ECM$  en el set de test, para todos los modelos construidos, en ambos sets de datos.

	Modelo	$ECM$ mat	$ECM$ por
1	Promedio	2.3367	8.6791
2	Árbol crecido	0.2722	1.2871
3	Árbol podado	0.3135	1.4485
4	Bagging	0.1819	0.8039
5	Random Forest	0.1727	0.8022
6	Boosting	0.2619	0.8886
7	BART	0.2113	0.8333

Cuadro 14: Comparación de modelos construidos usando árboles para ambos sets de datos.

Se observa que el modelo que mejor predice a  $G_3$  es Random Forest en ambos casos. Puesto que analizando el resto de los modelos se llegaron a conclusiones de features similares, se decidió desarrollar únicamente el modelo de Random Forest a continuación.

#### Random Forest

En las Figuras 30 y 31 se observa la profundidad mínima promedio de los features más importantes (utilizando la pureza del nodo como medida de importancia). Esta medida significa: dado  $X$ , en promedio, hasta qué profundidad del árbol hay que irse para que aparezca una partición de  $X$ .

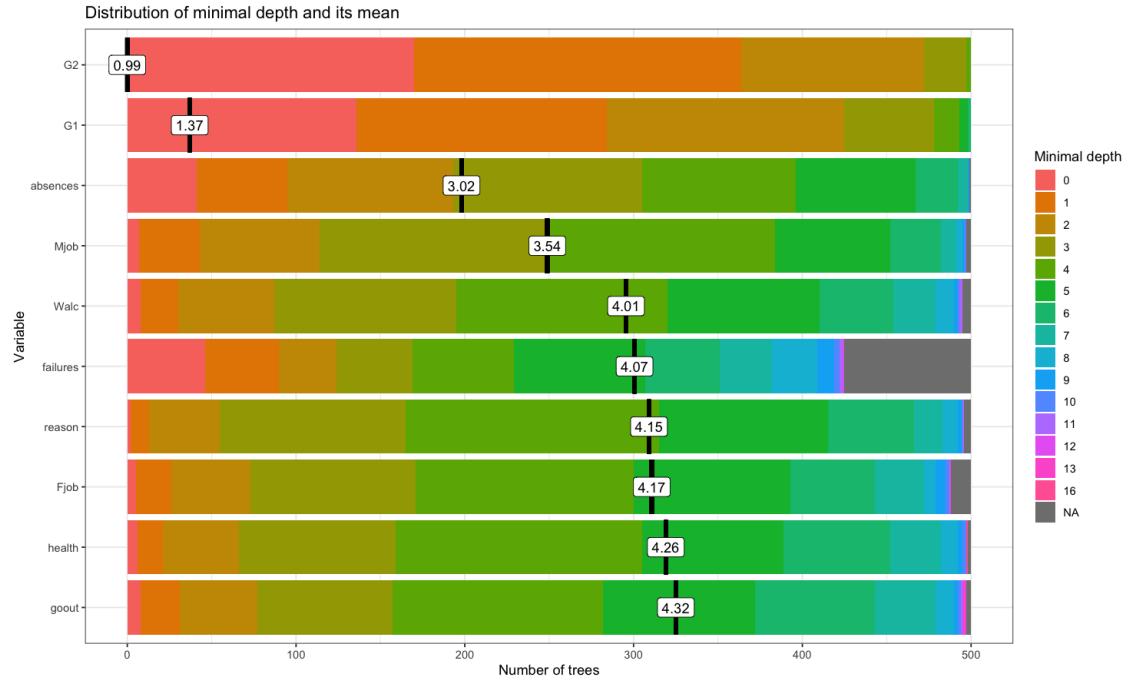


Figura 30: Mínima profundidad donde aparece cada feature en promedio. Set de Matemática

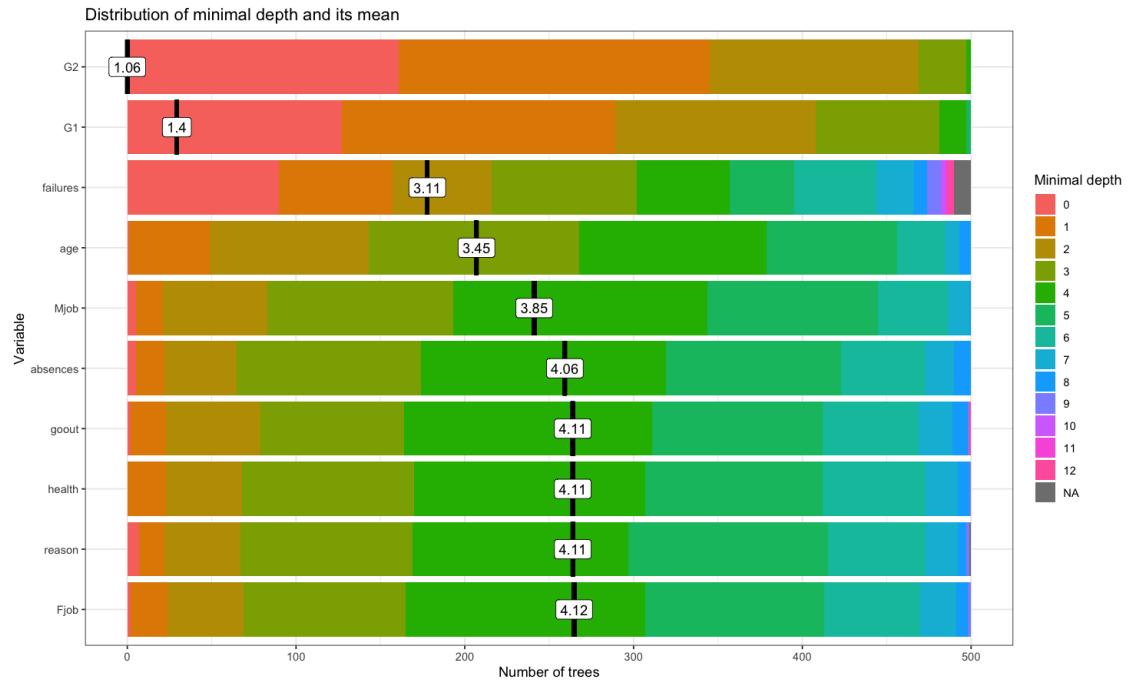


Figura 31: Mínima profundidad donde aparece cada feature en promedio. Set de Portugués

Se aprecia que las profundidades de  $G_1$  y  $G_2$  son mucho menores que el resto. Otra vez, utilizando estas variables, resultará más difícil encontrar relaciones interesantes para el resto de los features.

De igual forma, se aprecia que  $absences$ ,  $Mjob$ ,  $Walc$ ,  $failures$ ,  $reason$ ,  $Fjob$ ,  $health$  y  $goout$  son

predictores importantes para este modelo (aparecen, en promedio, antes de la profundidad 5).

Las Figuras 32 y 33 muestran otra forma de medir la importancia de los features en dos dimensiones:

- En el eje horizontal se cuenta con el incremento en el *ECM* estimado al remover ese predictor.
- En el eje vertical se aprecia la medida de impureza del nodo.

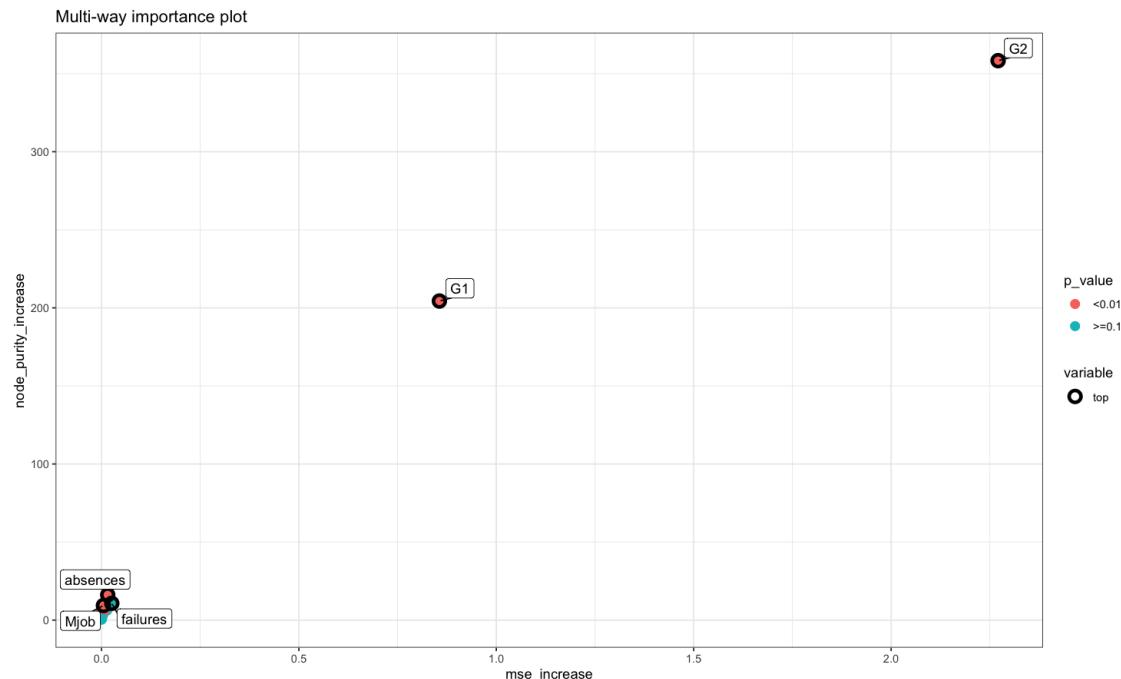


Figura 32: Mínima profundidad donde aparece cada feature en promedio. Set de Matemática

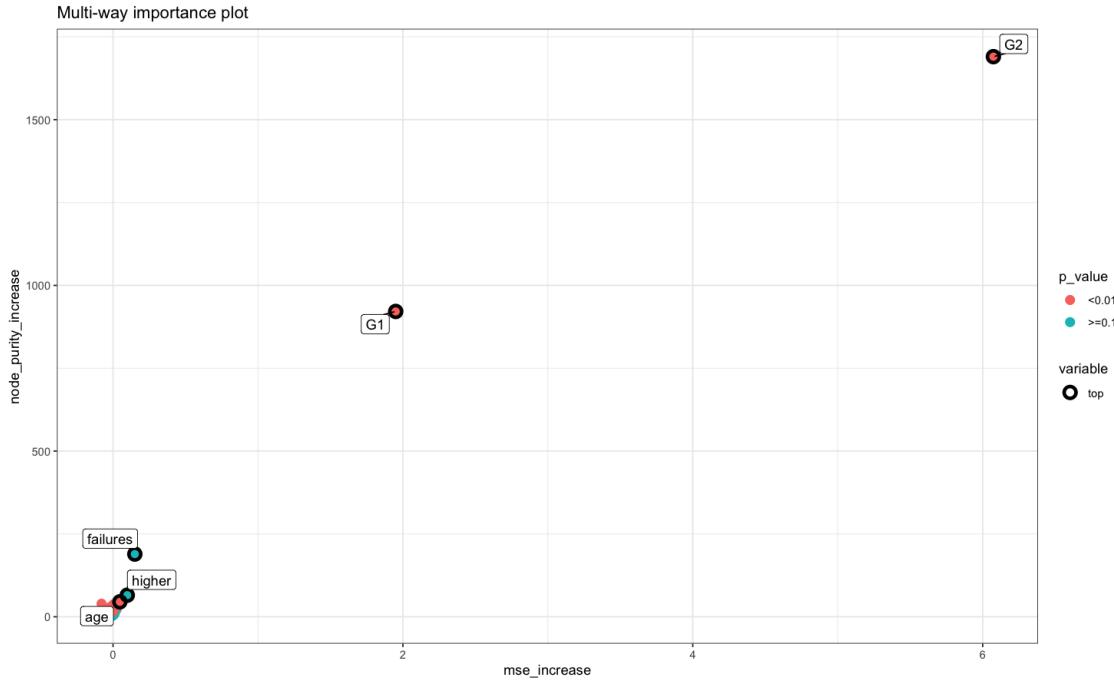


Figura 33: Mínima profundidad donde aparece cada feature en promedio. Set de Portugués

De esta forma, los predictores interesantes previamente mencionados quedan totalmente opacados por las dos variables de notas.

Puesto que estas dos variables son importantes en cada modelo, exploramos las relaciones que tienen los demás features con estas variables. Es decir, dado que el árbol tiene como raíz (primer nodo de partición) al predictor  $X$ , se busca entender la distancia mínima en promedio (con la misma noción que la profundida mínima en promedio) del predictor  $Y$  a  $X$ .

En las Figuras 34 y 35 se observa este análisis.

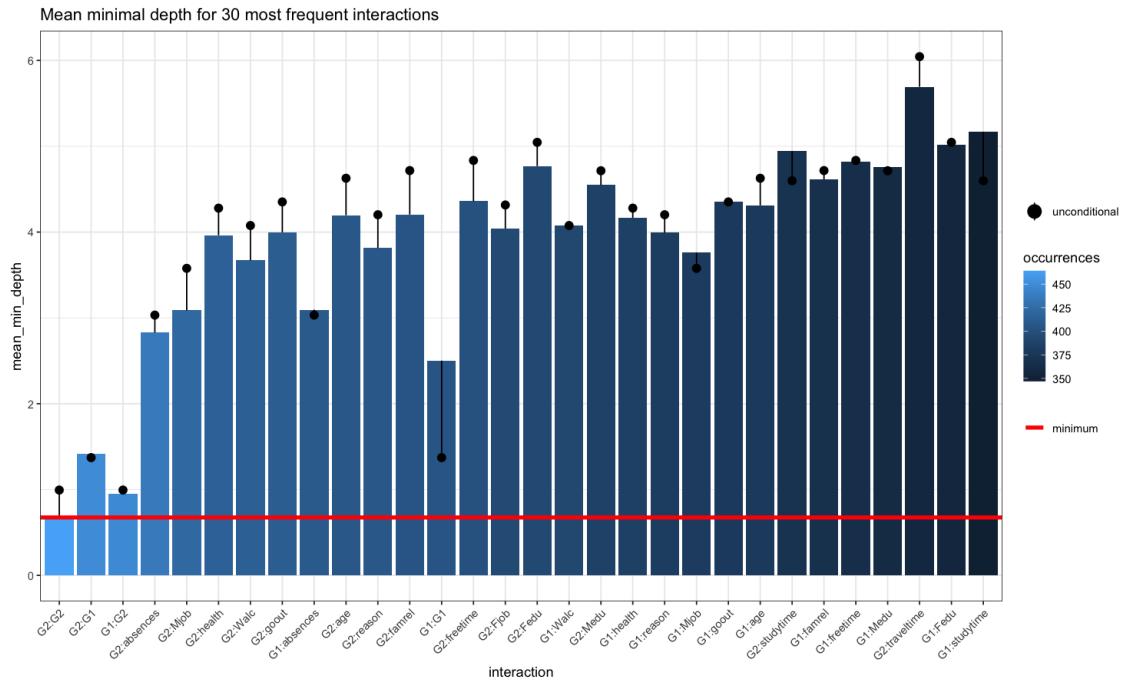


Figura 34: Mínima profundidad donde aparece cada feature en promedio. Set de Matemática

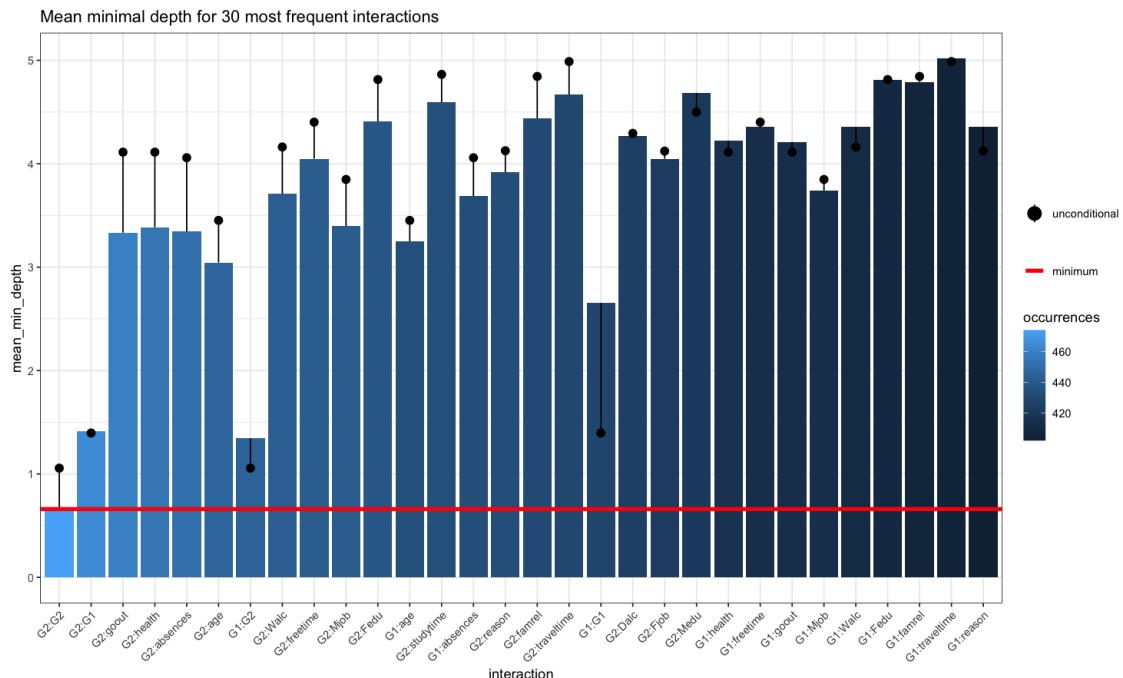


Figura 35: Mínima profundidad donde aparece cada feature en promedio. Set de Portugués

Era esperable que las menores distancias sean entre  $G_1$  y  $G_2$ . En las siguientes figuras se explora las relaciones entre algunos de estas relaciones.

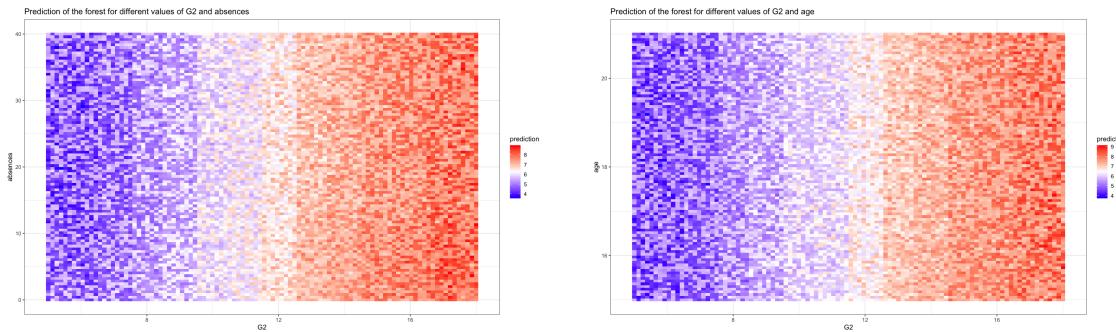


Figura 36: Relación entre  $G_2$  y: a. *absences*; b. *age* para predecir a  $G_3$ . Matemática

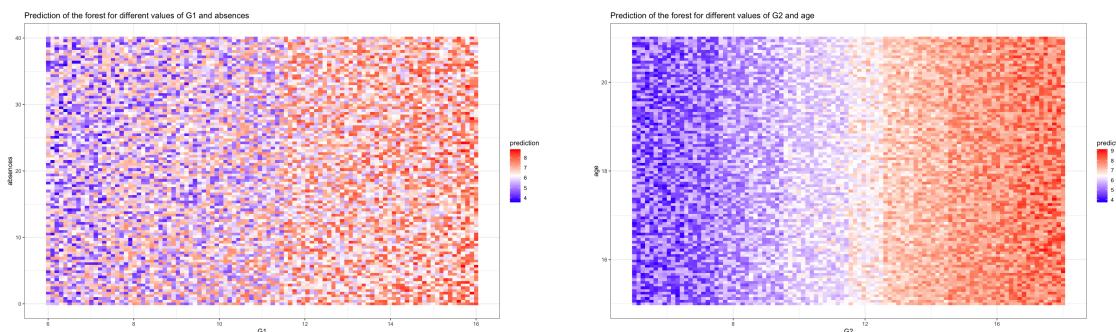


Figura 37: Relación entre  $G_1$  y: a. *absences*; b. *age* para predecir a  $G_3$ . Matemática

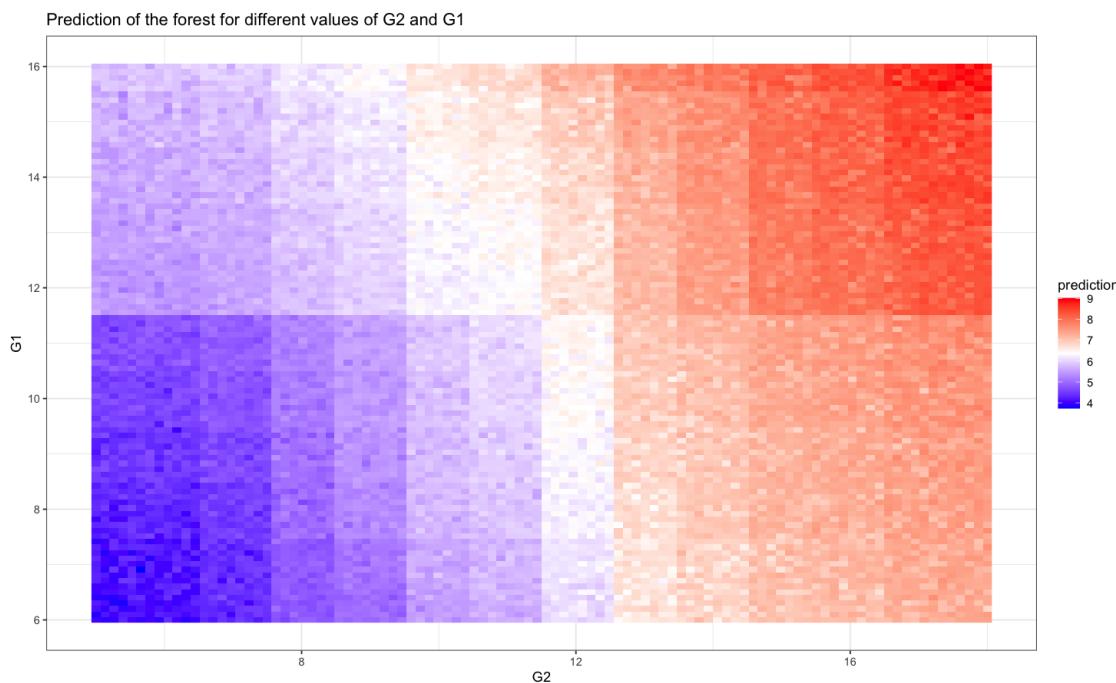


Figura 38: Relación entre  $G_1$  y  $G_2$  para predecir a  $G_3$ . Matemática

La relación entre  $G_2$  y los demás es clara: Al aumentar  $G_2$ , la nota aumenta, de forma inde-

pendiente del resto de las covariables.

En cambio, se observa que para  $G_1$ , si bien tiene la misma tendencia en relación al resto, no es tan determinante para la nota final.

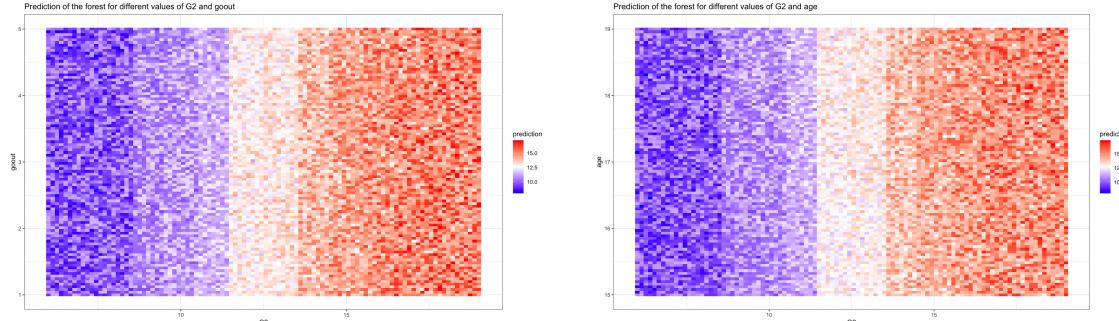


Figura 39: Relación entre  $G_2$  y: a. goout; b. age para predecir a  $G_3$ . Portugués

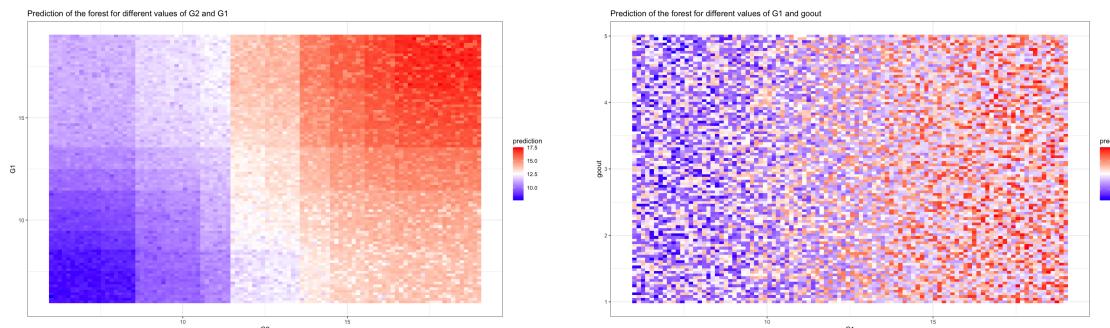


Figura 40: Relación entre  $G_1$  y: a.  $G_2$ ; b. goout para predecir a  $G_3$ . Portugués

## 5.2. Búsqueda de modelos sin $G_1$ y $G_2$

En la Tabla 15 se aprecia la estimación de la estimación del  $ECM$  en el set de test, para todos los modelos construidos, en ambos sets de datos.

	Modelo	$ECM$ mat	$ECM$ por
1	Nulo	2.3367	8.6791
2	Árbol crecido	3.1806	10.9471
3	Árbol podado	2.4730	6.9771
4	Bagging	1.8294	6.1466
5	Random Forest	1.8277	5.9995
6	Boosting	2.3059	6.8498
7	BART	1.7541	6.3986

Cuadro 15: Comparación de modelos construidos usando árboles para ambos sets de datos.

Nuevamente desarrollaremos los modelos construidos con Random Forest. Cabe destacar que en el set de matemática, el mejor modelo en términos del  $ECM$  estimado es *Bayes Additive Regression Tree (BART)*, pero utilizaremos en ambos sets de datos el mismo modelo para conservar consistencia en el análisis.

## Random Forest

En las Figuras 41 y 42 se aprecia que:

- Las menores profundidades en matemática están dadas por: *absences*, *failures*, *schoolsup*.
- En portugués: *failures*, *higher* son las más importantes, destacando la gran diferencia de *failures* con el resto.

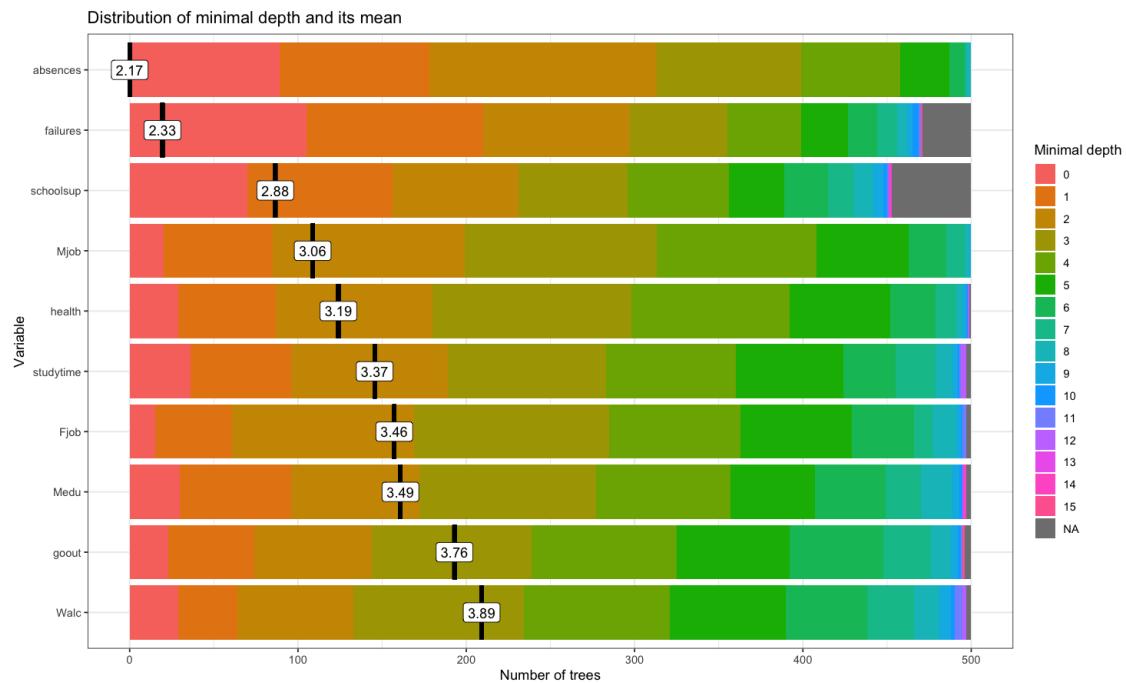


Figura 41: Mínima profundidad donde aparece cada feature en promedio. Set de Matemática

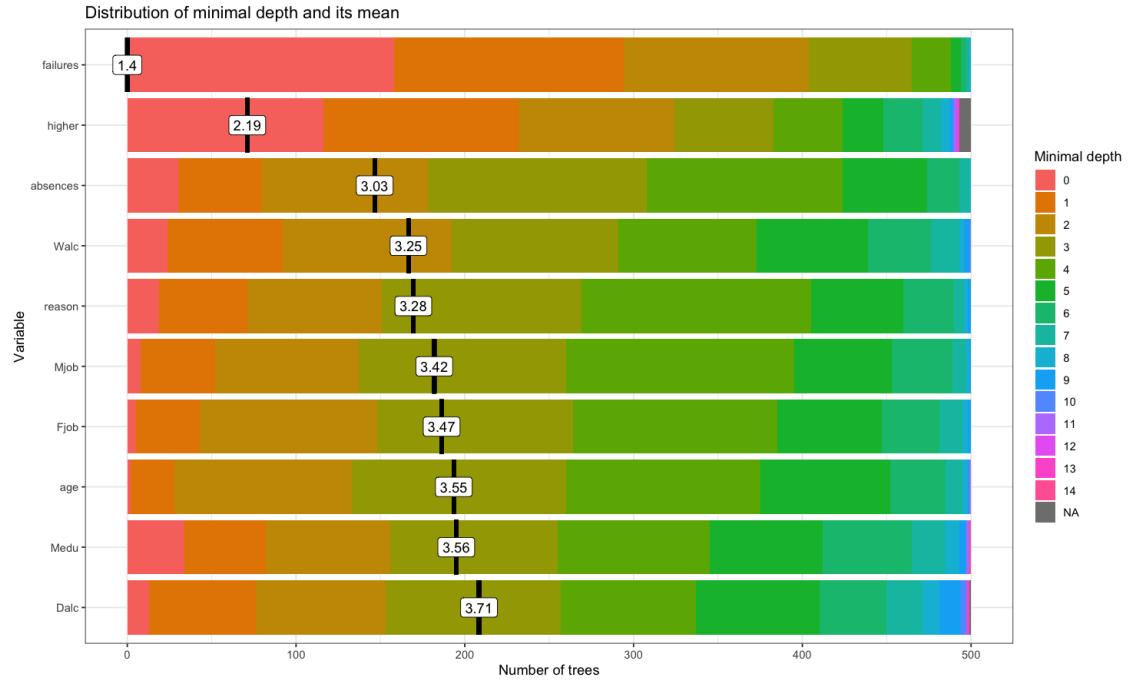


Figura 42: Mínima profundidad donde aparece cada feature en promedio. Set de Portugués

Las Figuras 43 y 44 vuelven a mostrar otra medida de los features.

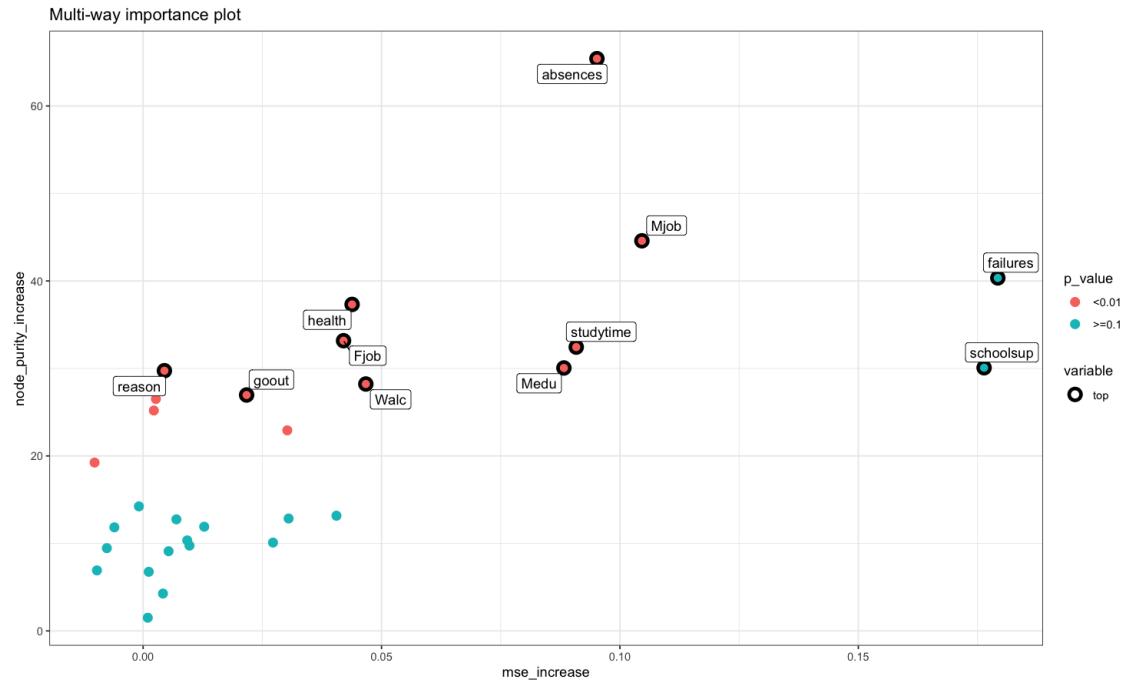


Figura 43: Mínima profundidad donde aparece cada feature en promedio. Set de Matemática

En el caso de matemática:

- *absences* es aquella que maximiza la pureza del nodo, pero no es el mejor predictor en términos del *ECM* estimado. Se puede apreciar que *Mjob* también entra en esta categoría.

- *failures* y *schoolsup* son muy importantes en términos de reducir el *ECM* estimado.

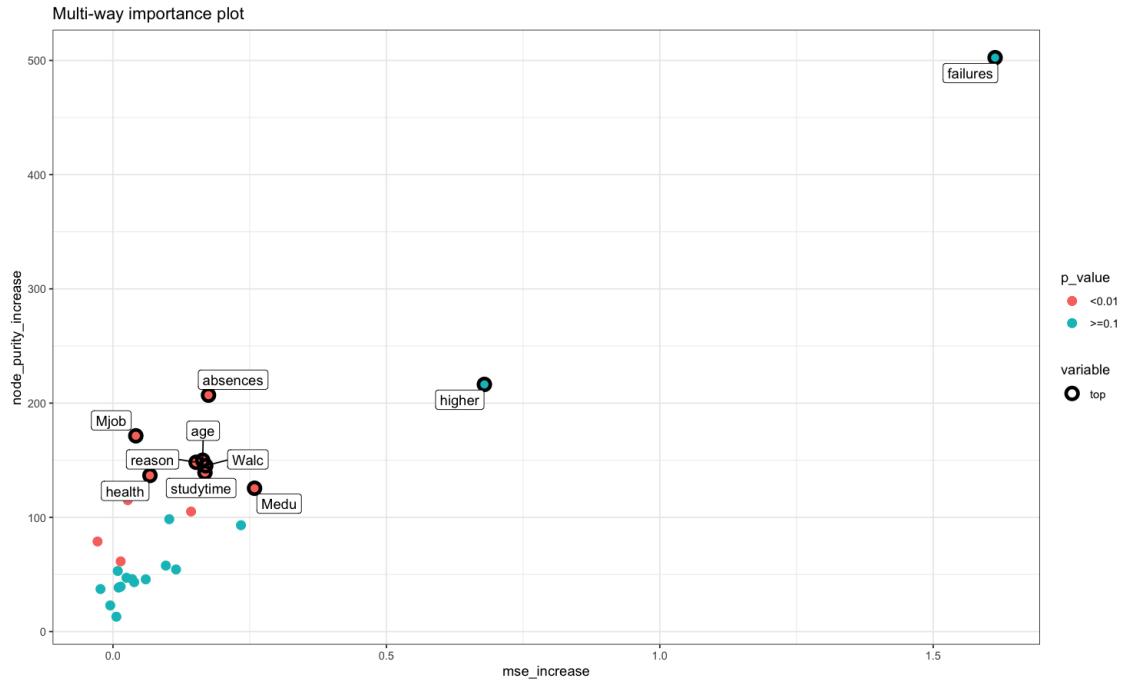


Figura 44: Mínima profundidad donde aparece cada feature en promedio. Set de Portugués

Para portugués: *failures* es la mejor variable para ambas medidas, seguida por *higher*. El resto de los features (más importantes) no muestran grandes diferencias entre sí.

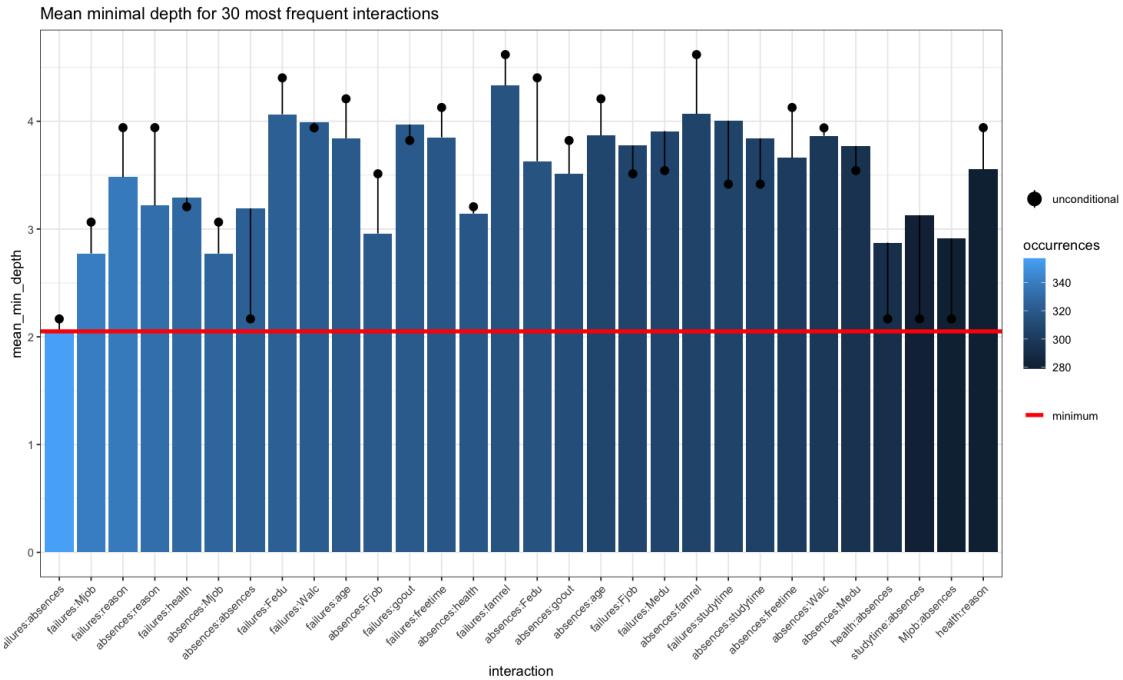


Figura 45: Relaciones de mínima profundidad donde aparece cada feature en promedio. Set de Matemática

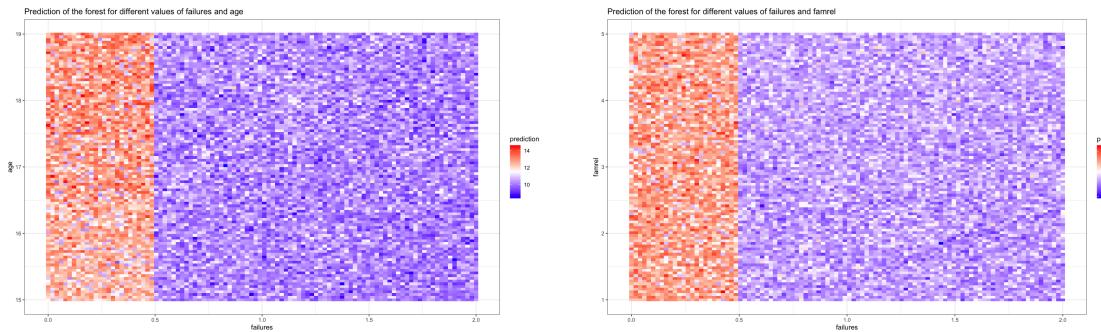


Figura 46: Matemática: Relación entre mínima profundida promedio entre a. *failures vs absences*; b. *failures vs health*.

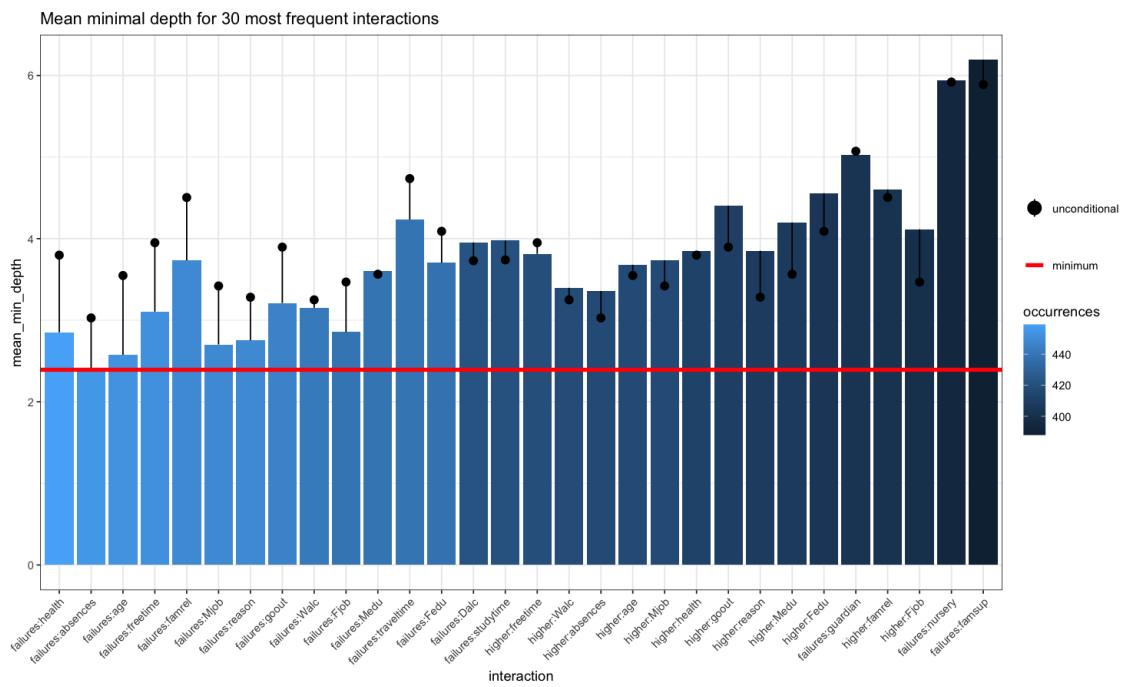


Figura 47: Relaciones de mínima profundidad donde aparece cada feature en promedio. Set de Portugués

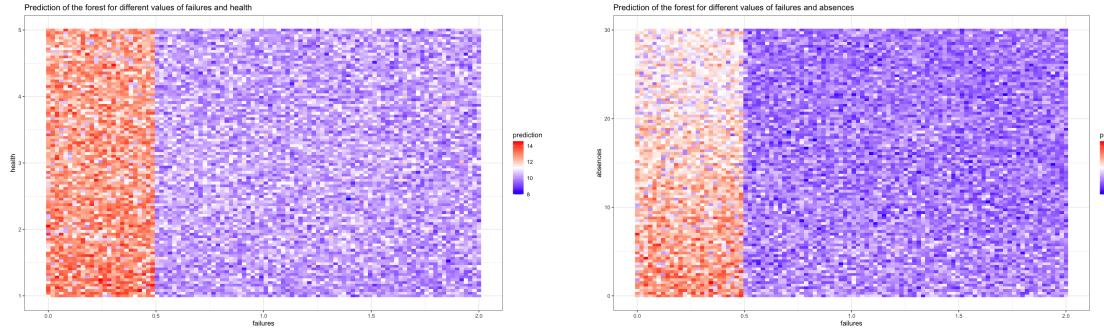


Figura 48: Portugués: Relación entre mínima profundida promedio entre a. *failures vs health*; b. *failures vs absences*.

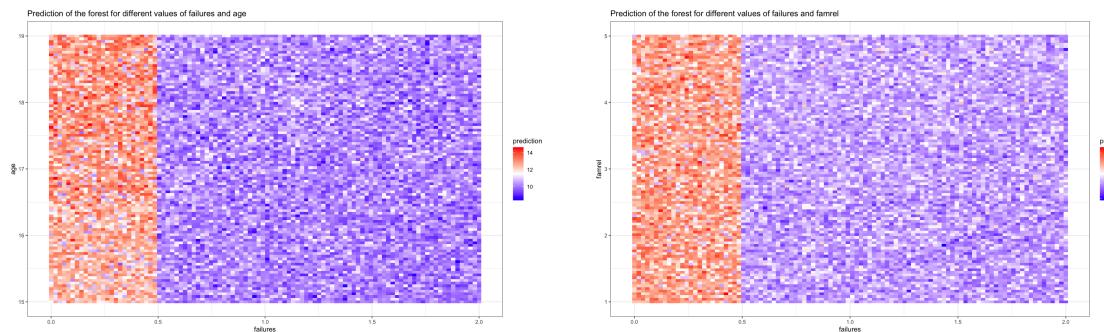


Figura 49: Portugués: Relación entre mínima profundida promedio entre a. *failures vs age*; b. *failures vs famrel*.

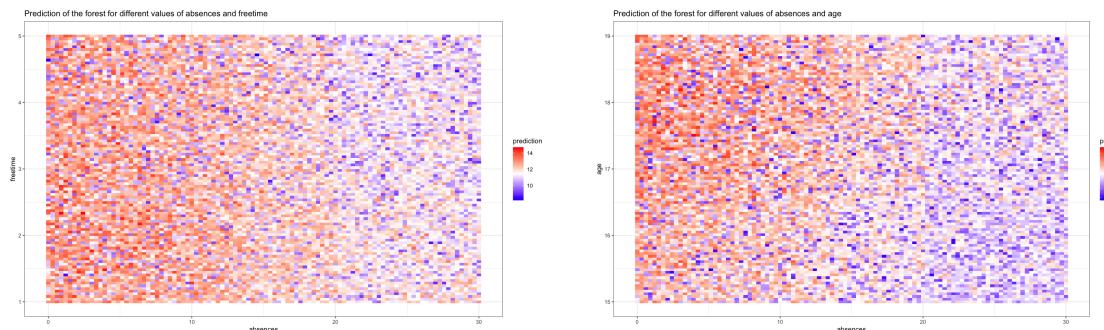


Figura 50: Portugués: Relación entre mínima profundida promedio entre a. *absences vs freetime*; b. *absences vs age*.

Se concluye que las fallas en el set de portugués tiene una mayor importancia en comparación al resto de las covariables, para predecir a  $G_3$ .

En el set de matemática hay otras variables a tener en cuenta, principalmente las ausencias.

## 6. Conclusiones y análisis de resultados

En la presente sección analizamos los resultados obtenidos en ambos modelos, para ambos sets de datos, con el objetivo de entender qué predictores son diferentes en ambas materias.

Es claro que contando con todos los predictores, las variables predominantes para predecir la nota final son las notas:  $G_1$  y  $G_2$ . A continuación, se analizan para ambas materias, cuáles son las características más importantes para una mejor o peor nota, obtenidas de ambos modelos (lineal y árboles).

### Qué produce una buena nota en matemática

- $G_2$  es el mejor predictor (significativo y distinto de 0 con confianza del 95 %).
- *famrel*, es decir, la calidad de las relaciones familiares, es una variable significativa y, como se observa en la Figura 23, se tiene una confianza del 95 % de ser distinta de 0. Por lo tanto, se concluye que es una variable significativa, a una mayor calidad de relación familiar, mayor es la nota final en matemática.
- *nursery*, de forma contraria, si el alumno fue a una guardería, la nota final en matemática tiende a ser menor.
- *studytime*: En la Figura 29 se aprecia que contamos con el predictor *studytimeTRUE*, que equivale a aquellos alumnos que estudiaron menos de 3 horas de forma semanal. Es decir, menor estudio (sobre todo menor a 3 horas por semana) produce una nota final de matemática menor.
- *schoolsup*: Si se cuenta con apoyo escolar, la nota en matemática se espera que baje en relación con otros alumnos sin apoyo escolar.
- *failures*: A mayor cantidad de fallas, la nota final es peor.
- *Fjob*: *FjobTRUE* equivale a aquellos padres que son profesores. Se concluye que si el trabajo del padre es profesor, el alumno tiene mayor probabilidad de una buena nota en matemática que el resto.
- *absences*: Los coeficientes asociados con las ausencias en el modelo lineal son cercanos a cero (con una confianza del 95 %). Para Random Forest, para la medida de impureza del nodo, las ausencias son el predictor más significativo (sin las notas). Es decir, los árboles consiguen partir las ausencias en intervalos representativos, que no se consiguen apreciar en el modelo lineal. Se concluye que es una variable significativa, entonces, a mayores ausencias, peor es la nota final en matemática.
- Otras variables importantes para Random Forest, pero no para los modelos obtenidos con regresión lineal: *Mjob*, *Medu*, *Fedu*, *goout*, *Walc*, *health*, *reason*.

### Qué produce una buena nota en portugués

- $G_2$  y  $G_1$  son los dos mejores predictores.
- *failures*: La cantidad de fallas vuelven a ser determinantes para una peor nota, en portugués.
- *higher*: Querer perseguir educación universitaria es un predictor de una mejor nota en portugués.
- *school*: La elección de colegio es importante para la nota final en portugués. Como se apreció en el análisis inicial, los alumnos que asisten al colegio Mousinho da Silveira tienen peores notas que los que asisten a Gabriel Pereira.

- *sex*: Un hombre tiene peores notas que las mujeres en portugués.
- *schoolsup*: Si se cuenta con apoyo escolar, la nota en portugués se espera que baje en relación con otros alumnos sin apoyo escolar.
- Se observa en el modelo construido con Random Forest sin  $G_1$  y  $G_2$ , las fallas y el querer perseguir educación superior son los dos predictores más importantes (sin contar los que se quitaron) en cualquier medida de las presentadas (para modelo lineal y árboles).

### Conclusión de ambos sets de datos

Se puede apreciar que para matemática, los hábitos (cuánto tiempo uno sale, su salud, el tiempo de estudio), su relación familiar (tanto para sí como con los estudios: Educación de los padres, trabajo, relación con el alumno) y las fallas y ausencias son factores determinantes para entender por qué unos alumnos tienen mejores notas que otros.

En portugués, los factores que determinan mejores notas finales están asociadas con la actualidad del alumno: Si quiere perseguir educación superior, a qué colegio va, el sexo, si recibe o no clases extra. Y, sobre todo, el factor determinante es la cantidad de fallas.

De esta manera, se descartan features que inicialmente se creían de mayor importancia (como la dirección donde un alumno vive, el acceso a internet, etc), y se reflejan relaciones más profundas entre los predictores.

En conclusión, quitando las notas de cursada: **la educación familiar (y su relación con el alumno), los hábitos relacionados con el estudio y el rendimiento en los exámenes (fallas) son determinantes para la nota final en matemática**. En cambio, **las ambiciones del alumno (educación superior) y su presente (escuela, sexo, clases de apoyo) y su rendimiento en los exámenes (fallas) son determinantes para la nota final en portugués**.

## A. Apéndice A – Análisis de Outliers

En este apéndice se busca mostrar el análisis realizado en el cual nos basamos para considerar a los puntos outliers a aquellos donde **G3 = 0**.

La idea principal de este análisis se basa en la utilización de métodos de clusterización (jerárquicos). Para empezar buscamos encontrar clusters sin tener en cuenta las variables  $\{G1, G2, G3\}$  para que a la hora de formar estos grupos no lo haga teniendo en cuenta ningún tipo de nota. Al utilizar clusters jerárquicos, buscaremos repartir en cierta cantidad las observaciones en diferentes grupos de tal forma quede distribuido. Es decir, formando inicialmente entre entre 7 y 8 clusters.

Una vez formado estos clusters, asignaremos su respectiva etiqueta de clustering en un nuevo dataset y volvemos a tener en cuenta a  $G3$ . A partir de ahí faremos una descripción de como se comporta esta variable  $G3$  en los diferentes tipos de grupos formados (que recordemos que han sido formados sin tener en cuenta ninguna nota, ni siquiera  $G3$ ).

Lo que queremos tener en cuenta de esta descripción es saber en qué grupos caen los puntos en el cual **G3 = 0**. Si observamos que hay puntos donde **G3 = 0** se mantienen en un mismo grupo entonces podemos decir hay un patrón o característica aparente que se puede tener en cuenta para cuando los puntos  $G3$  adoptan ese valor. En definitiva, consideraremos como puntos outliers cuando los valores donde **G3 = 0** se distribuyan en diferentes grupos, sin presentar lograr obtener tipo de patrón. Es decir, observando que son puntos atípicos al dataset.

### A.1. Matemáticas

Empezamos el análisis con el dataset de matemáticas, la clusterización obtenida fue de 8 grupos:

Cuadro 16: Descripción de G3 para dataset de matemáticas, por grupo

Grupo	n	Media	SD	Mediana	Mínimo	Máximo
1	91	11.69	3.40	11.0	5	20
2	<b>225</b>	9.95	5.29	11.0	<b>0</b>	19
3	54	10.35	2.86	11.0	5	16
4	10	10.70	4.62	10.0	4	18
5	2	9.50	2.12	9.5	8	11
6	10	9.80	3.49	8.5	5	17
7	1	9.00	NA	9.0	9	9
8	2	9.50	2.12	9.5	8	11

El cual se puede observar que todos los puntos donde **G3 = 0** se distribuyen en un único cluster, además es el grupo donde hay mayor cantidad de observaciones. Clusterizemos éste grupo; repitamos el proceso de clusterizar sin tener en cuenta a  $G3$  y luego una vez formado los grupos, volvamos a hacer el mismo análisis descriptivo de  $G3$  para los diferentes grupos formados:

Cuadro 17: Descripción de G3 (del grupo 2) para dataset de matemáticas, por grupo

Grupo	n	Media	SD	Mediana	Mínimo	Máximo
1	22	10.55	3.17	10.0	6	19
2	22	12.32	2.87	12.0	5	17
3	37	7.76	5.88	10.0	0	16
4	27	11.59	5.09	12.0	0	19
5	49	11.67	5.21	13.0	0	18
6	8	10.25	1.91	10.0	8	13
7	14	9.79	5.07	10.0	0	16
8	4	9.25	1.71	9.5	7	11
9	13	9.08	6.03	10.0	0	19
10	23	6.96	6.05	9.0	0	18
11	3	4.33	7.51	0.0	0	13
12	3	5.67	5.13	7.0	0	10

A partir de esta ultima tabla podemos concluir que los puntos **G3 = 0** son puntos outliers. Se puede observar como se distribuyen estos puntos en los diferentes clusters; esto se puede tener en cuenta al ver el valor **Mínimo = 0** que adopta G3 para los diferentes grupos. No hay una concentración de puntos en ningun cluster donde **G3 = 0**, pues no se observa ningún valor donde la **Media** adopte un valor cercano o igual a 0.

## A.2. Portugués

Repitamos el análisis para el dataset portugués:

Cuadro 18: Descripción de variable G3 para dataset de portugués, por grupo

Grupo	n	Media	SD	Mediana	Mínimo	Máximo
1	405	11.97	3.53	12.0	0	19
2	85	11.59	2.68	11.0	7	18
3	114	12.47	2.54	12.0	8	19
4	27	10.81	2.30	11.0	7	17
5	6	9.33	2.34	9.5	6	13
6	2	15.00	1.41	15.0	14	16
7	10	9.40	2.01	10.0	5	11

Observamos lo mismo que antes, hay una distribución de puntos donde **G3 = 0** en el grupo 1, y además es el grupo de mayor cantidad observaciones. Clusterizemos este grupo:

Cuadro 19: Descripción de variable G3 para dataset de portugués, por grupo

Grupo	n	Media	SD	Mediana	Mínimo	Máximo
1	107	12.17	2.90	12.0	0	18
2	46	10.65	4.33	11.0	0	17
3	73	13.03	2.67	13.0	0	19
4	40	10.35	3.38	10.0	0	17
5	48	13.25	4.38	14.0	0	18
6	33	12.48	3.59	13.0	0	18
7	16	12.06	2.05	12.0	7	16
8	16	9.75	3.91	9.5	0	18
9	18	12.44	2.59	13.0	8	18
10	8	8.75	3.62	10.0	0	11

Y aquí se puede observar otra vez las mismas conclusiones que en matemáticas. Para el dataset de portugués los puntos donde **G3** = 0 son también puntos outliers. Se observa como hay una distribución de puntos **G3** = 0 en diferentes grupos y además en ningún grupo hay una concentración de dichos puntos.