FACULTAD DE CIENCIAS

SOCIALES Y JURÍDICAS DE ELCHE

TRABAJO FIN DE GRADO

FACTORIZACIÓN Y SELECCIÓN DE VARIBLES PARA MODELOS PREDICTIVOS EN EDUCACIÓN

***Alumna**: Inmaculada Meca Sáez*

***Tutor**: Alejandro Rabasa Dolado*

4º ESTADÍSTICA EMPRESARIAL

INDICE

1. RESUMEN

El análisis del rendimiento académico de los estudiantes constituye un tema de gran interés que es objeto de numerosos trabajos de investigación. Las técnicas de Minería de Datos constituyen una herramienta de gran eficacia para predecir el rendimiento académico de los estudiantes y permiten identificar y evaluar los factores que más influyen en el proceso de aprendizaje de los alumnos.

Esta investigación tiene como objetivo aplicar tres técnicas de Minería de Datos (Factorización, Selección de Atributos y Árboles de Decisión) a la información obtenida en dos Institutos de Educación Secundaria de Portugal durante el Curso 2005-2006, con la finalidad de analizar cómo afectan las diferentes formas de factorizar las variables numéricas a la importancia relativa y al nivel de precisión de los modelos de clasificación, permitiendo comparar la influencia de estas variables en ciencias y en letras. Por otra parte, se pretende deducir la forma óptima de discretización al objeto de predecir la variable objetivo calificación final (G3) en el caso propuesto, para encontrar un clasificador del rendimiento académico y detectar los patrones determinantes en éste.

Palabras clave: Rendimiento académico, Técnicas de Minería de Datos, Factorización, Selección de Atributos, Árboles de Decisión, Detección de Patrones.

2. INTRODUCCIÓN Y OBJETIVOS

2.1. INTRODUCCIÓN AL TFG

En la actualidad las instituciones educativas se enfrentan al reto de mejorar la calidad en la enseñanza. Uno de los factores que más influyen en esta calidad es el rendimiento académico de los estudiantes. El análisis de este rendimiento ha constituido desde hace décadas un tema de gran interés en todos los niveles educativos y en la actualidad son numerosos los trabajos de investigación a nivel internacional que centran sus estudios en este ámbito.

El rendimiento escolar es un índice que revela el grado en el que los alumnos demuestran su progreso en la adquisición del conocimiento y habilidades correspondientes a su etapa evolutiva e indirectamente el grado de eficacia de los sistemas instruccionales (Muelas y Beltrán, 2011). Cuando este índice se sitúa por debajo de la media en cada nivel educativo, las instituciones educativas consideran necesario evaluar los factores que influyen en el nivel de este índice con el objetivo de emprender acciones encaminadas a mejorarlo.

Las técnicas de Minería de Datos, aplicadas a la información obtenida a través de las instituciones educativas, permiten establecer unos modelos predictivos que constituyen una herramienta de gran eficacia para predecir el rendimiento académico de los estudiantes y para poder identificar y evaluar los factores que más influyen en el proceso de enseñanza y aprendizaje de los alumnos, proporcionando, de esta forma, una información sólida y fiable que permita mejorar este proceso educativo mediante acciones encaminadas a prevenir el bajo rendimiento escolar.

En la actualidad se ha generado una nueva comunidad de investigación en educación denominada Minería de Datos Educativa, que aplica las técnicas de Minería de Datos para analizar y evaluar los datos obtenidos en los entornos educativos y transformarlos en información útil, con la finalidad de comprender mejor los factores que influyen en el rendimiento académico y poder servir de apoyo a la toma de decisiones por parte de las instituciones con el fin de mejorar la calidad en el proceso de enseñanza y aprendizaje de los alumnos.

Esta investigación tiene como objetivo aplicar las tres técnicas de Minería de Datos (Factorización, Selección de Atributos y Árboles de Decisión) utilizando los datos del curso 2005-2006, obtenidos en dos centros de enseñanza públicos de Educación Secundaria de la Región del Alentejo en Portugal (Gabriel Pereira y Mousinho da Silveria), con la finalidad de analizar cómo afectan las diferentes formas de factorizar las variables numéricas a la importancia relativa y al nivel de precisión de los modelos de clasificación, permitiendo de esta forma comparar esta importancia relativa de las

variables en ciencias y en letras. Por otra parte, se pretende deducir la forma óptima de discretización al objeto de predecir la variable objetivo (G3) en el caso propuesto para encontrar un clasificador del rendimiento académico y detectar los patrones determinantes en este rendimiento estudiantil.

En este trabajo se ha utilizado la matriz de confusión para comparar y evaluar la precisión de los clasificadores. Los resultados muestran que la forma más adecuada, en el caso de ciencias, de clasificar a los estudiantes, es discretizando la variable objetivo (G3) en cuatro tramos, aunque la precisión no sea la más alta (66%). Sin embargo, en letras los resultados indican que la mejor tasa de clasificación se obtiene en el caso de la discretización en dos tramos de la variable objetivo "Calificación final" con una precisión de 71%.

2.2. OBJETIVOS DEL TFG

Los objetivos principales de este trabajo son dos: por una parte, analizar cómo afectan las diferentes formas de factorizar las variables numéricas a la importancia relativa y al nivel de precisión de los modelos de clasificación, permitiendo de esta forma comparar esta importancia relativa de las variables en ciencias y en letras. Por otra parte, elegir la forma óptima de discretización al objeto de predecir la variable objetivo (G3) en el caso propuesto.

2.3. OBJETIVOS PERSONALES

Los objetivos personales a conseguir con este trabajo son los siguientes:

- ✓ Aplicar los conocimientos obtenidos durante la formación en el Grado de Estadística Empresarial a un caso práctico en relación con el ámbito educativo.

- ✓ Iniciar actividades de investigación, llevando a cabo tareas de preprocesamiento, en el marco del Programa para la Realización de Prácticas en Actividades de Fomento de la Investigación en el Centro de Investigación Operativa de la Universidad Miguel Hernández de Elche.

3. ESTADO DE LA CUESTIÓN Y MARCO TEÓRICO

3.1. MÉTODOS DE FACTORIZACIÓN

La factorización de una variable (en algunos contextos también llamada discretización) es el proceso de convertir una variable numérica en categórica de una manera óptima. Dependiendo de la situación, esta conversión puede llevar a una mejor interpretación de la variable numérica, a la segmentación rápida del usuario o simplemente a una característica adicional para construir un modelo predictivo.

Uno de los métodos más utilizados es el de la **regla de Sturges** que es un método empírico muy utilizado en la estadística descriptiva para determinar el número de clases que deben existir en un histograma de frecuencias, y así poder clasificar un conjunto de datos que representan una muestra o población. Esta regla fue enunciada en 1926 por el matemático alemán Herbert Sturges.

Sturges propuso un método sencillo basado en el número de muestras x que permitiesen encontrar el número de clases y su amplitud de rango. La regla de Sturges se utiliza sobre todo en el área de la estadística, particularmente para construir histogramas de frecuencia.

Para establecer su regla Herbert Sturges consideró un diagrama de frecuencias ideal, que consta de K intervalos y estableció que el número óptimo de intervalos k es dado por la expresión:

$$k = 1 + \log_2 (N)$$

También expresada como:

$$k = 1 + 3{,}322 * \log_{10} (N)$$

Siendo ***k*** el número de clases y ***N*** el número total de observaciones de la muestra.

Por ejemplo, para la realización de un histograma de frecuencia que exprese una muestra aleatoria de la estatura de 142 niños, el número de intervalos o clases que tendrá la distribución en este caso es:

$$k = 1 + 3{,}322 * log10\ (N)$$

$$k = 1+3{,}322* log\ (142)$$

$$k = 1+3{,}322* 2{,}1523$$

$$k = 8{,}14 \approx 8$$

Finalmente, la distribución será en 8 intervalos.

Además, existen otros métodos de discretización como por ejemplo:

- La descomposición a paso constante entre los valores mínimo y máximo de la columna de valores seleccionada.
- El cálculo de clases óptimas frente a la minimización de la inercia intraclase con efectivos iguales en el caso de los datos no ponderados o, a peso constante, cuando los datos son ponderados.
- La importación de una lista de intervalos de clases.
- La modificación manual de los intervalos.

3.2. MÉTODOS DE CLASIFICACIÓN

En el ámbito del Data Science, se considera la clasificación como un proceso que asocia cada registro (o tupla) de un dataset con una clase predefinida. Se trata de un proceso de aprendizaje supervisado, ya que los métodos de clasificación infieren un determinado valor (discreto) de la variable de clase de cada tupla, a partir de su similitud con tuplas anteriores (previamente asignadas a una clase).

Existen varios algoritmos (llamados clasificadores) capaces de llevar a cabo esta tarea. Los más extendidos en la literatura, son los árboles de clasificación como el ID3 (Quinlan, 1986) y el C4.5 (Quinlan, 1993) que es una evolución del anterior, capaz de manejar antecedentes de naturaleza numérica. Los algoritmos de clasificación han sido ampliamente aplicados en contextos tan distintos como las telecomunicaciones, la medicina o las finanzas. Como salida, los métodos de clasificación, pueden proporcionar los típicos árboles de decisión en sí, o sistemas de reglas con el formato “si (antecedente)

entonces (consecuente)", donde el antecedente es una combinación de valores de variables explicativas y el consecuente es un determinado valor de la variable objetivo.

Tanto ID3 o C4.5 como otros clasificadores generan particiones recursivas de la muestra de entrenamiento (un subconjunto del dataset original) en función de los valores de sus variables explicativas y acaba asignando las subdivisiones finales generadas, a un valor de la variable objetivo.

En general, la precisión de los clasificadores es medida a través de la "accuracy" que es considerada como una métrica de la calidad global del modelo y que refleja la proporción entre el número de instancias bien clasificadas frente al total de instancias del dataset global. "Accuracy" toma valores entre 0 y 1 (1 cuando el total de la muestra es correctamente clasificada). Sin embargo, en la medida en que no todos los valores de la variable objetivo son predichos con diferente precisión, es necesario medir cómo se distribuye la precisión de la predicción de cada posible valor de la variable objetivo.

Esto se lleva a cabo con las matrices de confusión (Hernández Orallo, J. et al., 2004) que en el caso de variables objetivo binarias ("+" y "-" en el ejemplo siguiente), tienen un aspecto similar al que se muestra en la Tabla 1.

Valor real ↓ clasificado como →	+	-
+	300	20
-	40	200

Tabla 1: Ejemplo de una matriz de confusión
Fuente: elaboración propia

En este ejemplo, se dispone de 500 instancias bien clasificadas (suma de los valores de la diagonal: 300+200), y 60 instancias mal clasificadas (suma de valores fuera de la diagonal: 40+20), de un total de 560 instancias (suma de toda la matriz). Así, por ejemplo, 20 es el total de falsos negativos, es decir, instancias cuyo consecuente era "+" y que el modelo acabó clasificando como "-". El modelo del ejemplo, tendría una accuracy de: 500 / 560 = 0,89 (89%).
De manera general, las matrices de confusión también son aplicables a variables objetivo discretas de tamaño n (con $n>2$). En este caso, las matrices son de dimensión n x n.

3.3. MÉTODOS DE SELECCIÓN DE ATRIBUTOS

Sin abandonar el ámbito de los problemas de clasificación, el modelado de una variable objetivo discreta en función de un conjunto de variables explicativas, es absolutamente dependiente de cuáles de éstas sean elegidas. En problemas con un número manejable de variables explicativas, se suelen tener en cuenta todas ellas, delegando en el propio algoritmo clasificador, los mecanismos de dar a unas más importancias que a otras, gracias a los métodos de boosting y prunning que incorporan (Martínez Muñoz G. y Suárez A., 2007).

Sin embargo, es una práctica habitual recurrir a mecanismos de selección de atributos, que los ordenan por importancia (relación con la variable objetivo) para permitirnos elegir cuáles de éstos debemos utilizar para la construcción de los modelos de clasificación.

La selección de características es uno de los conceptos centrales en el aprendizaje automático que afecta enormemente al rendimiento del modelo. La selección de atributos y la limpieza de datos deben ser el primer paso y el más importante para el diseño del modelo, lo que hace que este sea más fácil de interpretar. Tener características irrelevantes en los datos puede disminuir la precisión de los modelos.

Los beneficios de realizar una selección de atributos antes de modelar los datos son los siguientes:

- ***Reduce el sobreajuste***: menos datos redundantes significa menos oportunidad de tomar decisiones basadas en el ruido.
- ***Mejora la precisión***: menos datos engañosos significa que la precisión del modelado mejora.
- ***Reduce el tiempo de entrenamiento***: menos puntos de datos reducen la complejidad del algoritmo y los algoritmos se entrenan más rápido.

En general, para reducir la dimensionalidad de los datos existen métodos como el Análisis de componentes principales, la descomposición del valor singular, etc. Por lo tanto, es natural plantearse el motivo por el que se necesitan otros métodos de selección de características. Lo que caracteriza a estas técnicas es que son formas no supervisadas de selección de características: por ejemplo, PCA utiliza la variación en los datos para encontrar los componentes. Estas técnicas no tienen en cuenta la información entre los valores de entidad y la clase o valores de destino. Además, existen ciertos supuestos, como la normalidad, asociados con tales métodos que requieren algún tipo de transformación antes de comenzar a aplicarlos. Estas restricciones no se aplican a todo tipo de datos.

Se pueden emplear alguna de estas tres técnicas de selección de características que son fáciles de usar y dan buenos resultados:

1. Selección univariable

Las pruebas estadísticas se pueden usar para seleccionar aquellas características que tienen la relación más fuerte con la variable de salida. La biblioteca scikit-learn proporciona la clase "SelectKBest" que se puede usar con un conjunto de diferentes pruebas estadísticas para seleccionar un número específico de funciones.

2. Importancia de la característica

Se puede obtener la importancia de cada característica de su conjunto de datos, utilizando la propiedad de importancia de la característica del modelo. La importancia de la característica le otorga una puntuación por cada característica de sus datos, mientras más alta o importante es la calificación hacia la variable de salida.

La importancia de la característica es una clase incorporada que viene con los clasificadores basados en árboles.

3. Correlación de matriz con mapa de calor

La correlación indica cómo se relacionan las características entre sí o con la variable de destino. La correlación puede ser positiva (el aumento en un valor de la característica aumenta el valor de la variable objetivo) o negativa (el aumento en un valor de la característica disminuye el valor de la variable objetivo).

Existen además, tres tipos de métodos de selección de atributos de forma genérica:

- ***Métodos de filtro***: los métodos de filtro se utilizan generalmente como un paso de preprocesamiento. La selección de características es independiente de cualquier algoritmo de aprendizaje automático. En cambio, las características se seleccionan en función de sus puntuaciones en diversas pruebas estadísticas para su correlación con la variable resultado. Algunos métodos de filtro comunes son métricas de correlación (Pearson, Spearman, distancia), prueba de Chi-cuadrado, Anova, puntaje de Fisher, etc.

- ***Métodos de envoltura***: en los métodos de envoltura, intenta utilizar un subconjunto de características y entrenar a un modelo usándolas. Sobre la base de las inferencias que saca del modelo anterior, decide agregar o eliminar características del subconjunto. Selección hacia adelante, eliminación hacia atrás son algunos de los ejemplos de métodos de envoltura.

- ***Métodos incrustados***: estos son los algoritmos que tienen sus propios métodos de selección de características incorporados. La regresión LASSO es un ejemplo de ello.

En este trabajo se utiliza uno de los métodos de envoltura que está disponible en R a través de un paquete llamado Boruta.

Selección de características con el Algoritmo de Boruta.

Boruta es un algoritmo envolvente de selección de características que tiene un carácter relevante, capaz de trabajar con cualquier método de clasificación que genere una medida de importancia variable (VIM); De forma predeterminada, Boruta utiliza Random Forest. El método realiza una búsqueda descendente de las características más importantes al comparar la importancia de los atributos originales con la importancia que se puede lograr al azar, utilizando sus copias permutadas y eliminando progresivamente las características irrelevantes para estabilizar esa prueba.

En la siguiente tabla, se muestra el resultado de aplicar la librería *Boruta*, que simplifica en una sola línea la llamada a un clasificador *(random forest)* junto con un algoritmo de búsqueda. Se muestra la importancia que Boruta asigna a cada atributo. El ranking se desarrolla denominando G3 (calificación final) como variable objetivo sin discretizar e incluyendo las calificaciones del primer (G1) y segundo período (G2). El dataset empleado se describe en detalle en la sección 5 de esta memoria.

RANKING DE ATRIBUTOS

(Variable objetivo G3)

	attribute	meanImp	medianImp	minImp	maxImp	normHits	decision
1	G2	35,46748503	36,03605771	29,54792591	38,88682021	1	Confirmed
2	G1	26,21167695	26,20970473	23,32670828	28,54655382	1	Confirmed
3	ausencias	23,24874689	23,8530205	15,97493161	28,09119091	1	Confirmed
4	faltas	10,9880068	11,01610585	9,080591163	12,66950205	1	Confirmed
5	schoolsup	5,720839044	5,902945131	2,716366986	7,644917233	1	Confirmed
6	edad	4,21399292	4,155988601	1,610298124	7,349643713	0,929292929	Confirmed
7	mas_alto	3,830149012	3,860769878	1,619885738	5,312454861	0,868686869	Confirmed
8	Medu	2,779109971	2,854885644	0,546740798	5,33505433	0,636363636	Confirmed
9	tutor	2,478328376	2,215671072	0,721975394	5,703387068	0,484848485	Confirmed
10	Mjob	2,270191662	2,152463685	-0,222311464	5,093438448	0,434343434	Rejected
11	pagadas	2,095819233	2,048336717	-0,541337454	5,027787982	0,464646465	Confirmed
12	romántico	1,627414175	1,64002593	-0,640170774	4,346330251	0,141414141	Rejected
13	Walc	1,377024022	1,438250563	-0,737894244	3,619943843	0,121212121	Rejected
14	salidas	1,242347541	1,178832776	-0,02387221	2,839883403	0,01010101	Rejected
15	Dalc	1,108433189	1,445905958	-1,024101281	2,087163984	0	Rejected
16	actividades	1,001163384	0,922407165	-0,914718424	2,744858458	0,01010101	Rejected
17	razón	0,910723293	0,660002055	-0,914020591	2,764036815	0,03030303	Rejected
18	sexo	0,808535632	0,743065145	-0,750048067	1,993951176	0,01010101	Rejected
19	dirección	0,740748272	1,229510061	-1,550054385	1,805885731	0	Rejected
20	tiempo_de_viaje	0,554574717	0,587767754	-1,180441934	1,768989814	0	Rejected
21	tiempo_de_estudio	0,548752195	0,487548607	-1,930351742	2,950049573	0,01010101	Rejected
22	Fedu	0,490870555	0,391497903	-0,838823325	1,427434733	0	Rejected
23	Psatatus	0,396604862	0,763801683	-1,69286229	1,343330929	0	Rejected
24	escuela	0,305056292	0,185170652	-1,151485349	1,630433955	0	Rejected
25	salud	0,293187863	0,360715433	-0,90000041	1,277870472	0	Rejected
26	tiempo_libre	0,237558346	0,042516237	-0,847092459	1,748728886	0	Rejected
27	internet	0,053347186	0,247565996	-1,046275566	1,559033722	0	Rejected
28	famrel	-0,04133632	-0,524066352	-1,659522887	2,155215386	0	Rejected
29	guardería	-0,238209761	-0,544778478	-2,252389733	1,625778362	0	Rejected
30	Fjob	-0,295538752	-0,326773638	-1,935616176	1,095257038	0	Rejected
31	famsize	-0,424776989	-0,416733753	-2,031082205	1,355313635	0	Rejected
32	famsup	-0,487363145	-0,744999126	-1,650761136	1,382153478	0	Rejected

Tabla 2. Importancia de los atributos según Boruta
Fuente: elaboración propia

Además de esta técnica existen algunas otras, todas ellas disponibles en R.

La diferencia entre las funciones proporcionadas por R y utilizadas en este trabajo para la selección de características consiste en que en el caso del algoritmo de *Boruta* el ranking de variables obtenido no sigue una distribución normal, sin embargo, utilizando el algoritmo de rankeo de la librería MachineLearning (*VariableRanker)* nos permite extraer un ranking en el que se asigna una importancia a las variables dentro de un rango normalizado entre 0 y 1.

3.4. ÁMBITO EDUCATIVO: SITUACIÓN ACTUAL Y MODELOS PREDICTIVOS

3.4.1.- Situación actual en el ámbito educativo en España y Portugal

A continuación, se refleja la estructura del sistema educativo en España y Portugal que ofrecen bastante similitud.

Estructura del sistema educativo en España

- De 3 a 6 años: Educación Infantil (No obligatoria)
- De 6 a 12 años: Educación Primaria (Obligatoria)
- De 12 a 16 años: Educación Secundaria (Obligatoria)
- De 16 a 18 años: Educación Secundaria Postobligatoria (No obligatoria)
- 18 años o más: Enseñanza Superior Universitaria (No obligatoria)

Estructura del sistema educativo en Portugal

- De 3 a 6 años: Educación Preescolar (No obligatoria)
- De 6 a 10 años: Primer Ciclo de Enseñanza Básica (Obligatoria)
- De 10 a 12 años: Segundo Ciclo de Enseñanza Básica (Obligatoria)
- De 12 a 15 años: Tercer Ciclo de Enseñanza Básica (Obligatoria)
- De 15 a 18 años: Enseñanza Secundaria (No obligatoria)
- 18 años o más: Enseñanza Superior Universitaria (No obligatoria)

3.4.2.- Modelos predictivos

Durante las distintas etapas del proceso llevado a cabo se han utilizado los datos que abordan el logro estudiantil en la educación secundaria de dos escuelas portuguesas (Gabriel Pereira y Mousinho da Silveira). Las herramientas de software utilizadas en esta investigación y estudiadas durante el Grado de Estadística Empresarial son las siguientes:

- ✓ Rstudio: entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráfico, para realizar la recopilación, integración y almacenamiento de los datos y seguidamente la depuración, selección y transformación de los datos. Este mismo programa ha sido útil para obtener los clasificadores aplicando técnicas de Minería de Datos.

- ✓ El programa informático Excel desarrollado y distribuido por Microsoft Corp ha sido un complemento para el desarrollo de las actividades llevadas a cabo con el software Rstudio.

Estos programas son una herramienta básica para llevar a cabo la realización de técnicas estadísticas y métodos para el aprendizaje de modelos *comprensibles* y *proposicionales* como Árboles de Decisión, Sistemas de Reglas, Regresión Logística etc. en este caso aplicadas al ámbito educativo.

El término "*modelo*" indica que estas técnicas construyen una hipótesis o representación de la regularidad existente en los datos. El término "*comprensible*" indica que estos modelos se pueden expresar de una manera simbólica, en forma de conjunto de condiciones (a diferencia de otros métodos, como las redes neuronales o las máquinas de vectores de soporte) y, por tanto, pueden tener como resultado formatos inteligibles para los seres humanos y también para sistemas semi- automáticos que procesen reglas. El término "*proposicional*" hace referencia que estos métodos se restringen a algoritmos que aprenden modelos sobre una única tabla de datos y que no establecen relaciones entre más de una fila de la tabla a la vez ni sobre más de un atributo a la vez.

En resumen, estos métodos utilizan la lógica de primer orden o de orden superior para representar los modelos y permiten establecer reglas que relacionan varios atributos y/o tablas.

De todos los métodos de aprendizaje, los sistemas de aprendizaje basados en árboles de decisión son quizás el método más fácil de utilizar y de entender.

Árboles de decisión (AD)

En términos generales, un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Los árboles de decisión se utilizan desde hace siglos, y son especialmente apropiados para expresar procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, etc.

Un árbol de decisión es un modelo jerárquico para el aprendizaje supervisado, que se puede aplicar a un problema de regresión o de clasificación. Un árbol de decisión es un modelo no paramétrico, que puede usarse para crear modelos predictivos automatizados que se puedan emplearse para el aprendizaje automático, la minería de datos y las estadísticas. El "aprendizaje basado en árboles de decisión" consiste en considerar las observaciones sobre un elemento para predecir su valor.

El árbol de decisión está constituido por un nodo raíz que se encuentra en la parte superior, un conjunto de nodos internos asociados cada uno a una variable predictora, y cuyas ramas representan validaciones o decisiones de los valores de la variable y un conjunto de nodos hojas o nodos terminales, que están etiquetadas con algún valor de la clase de la variable dependiente.

El modelo para un *árbol de clasificación* se presenta como una estructura jerárquica que establece las relaciones entre la variable dependiente y el conjunto de variables predictoras. Cada ramificación contiene un conjunto de atributos o reglas de clasificación asociadas a una etiqueta de clase específica que se encuentra al final de la ramificación.

En los casos en el los que los resultados son infinitos, posibles y continuos (la variable predicha es un número real) los árboles de decisión se llaman *"árboles de regresión"*. En estos casos se utiliza la regresión, normalmente lineal. Si llamamos A_{ij} al valor del atributo j para el ejemplo i si es continuo o la posición que ocupa entre los valores del atributo si el atributo es discreto, se genera un modelo "a priori" definido por la ecuación:

$$c_i = \alpha_0 + \sum_{j=1}^{a} \alpha_j \times A_{ij}$$

donde c_i es la clase para el ejemplo i, αj son los coeficientes del modelo, y a es el número de atributos del dominio. El objetivo de estos métodos es encontrar los valores de los αj de forma que se minimice el error cuadrático medio de las clasificaciones de los ejemplos de entrenamiento. Breiman y otros plantearon en los años ochenta el sistema CART que es parecido al ID3 y que generaba en los nodos hoja valores numéricos en lugar de clases discretas, a lo que se le denominó *árboles de regresión*.

Cuando un árbol representa la mayor cantidad de datos con el menor número de niveles o preguntas, se considera un árbol de decisión ideal. Los algoritmos diseñados para crear árboles de decisión optimizados incluyen CART, ASSISTANT, CLS y ID3/4/5. Cada método establece cuál es la mejor forma de dividir los datos en cada nivel.

Emplear árboles de decisión tiene ventajas, pero también desventajas. Una de las grandes ventajas es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes, lo que permite analizar una situación y, siguiendo el árbol de decisión apropiadamente, llegar a una sola acción o decisión a tomar. Existen otras ventajas: este tipo de árbol funciona para los datos numéricos o categóricos, puede modelar problemas con múltiples resultados, usa un modelo de caja blanca (lo que hace que los resultados sean fáciles de explicar) y la fiabilidad de un árbol se puede cuantificar y poner a prueba. Por el contrario, algunos de los inconvenientes son que cuando se presentan datos categóricos con múltiples niveles, la información obtenida se inclina a favor de los atributos con mayoría de niveles y además los cálculos pueden resultar complejos al lidiar con una cantidad importante de resultados relacionados.

Por otro lado, los sistemas de reglas son una generalización de los árboles de decisión, de hecho, un árbol de decisión se puede expresar como un conjunto de reglas en el que no se exige exclusión ni exhaustividad en las condiciones de las reglas, es decir, podría aplicarse más de una regla o ninguna.

En general, la diferencia más importante entre los sistemas de aprendizaje de árboles de decisión y los sistemas de inducción de reglas proposicionales es la filosofía del algoritmo que utilizan.

Regresión logística binaria (RL)

La Regresión Logística es un procedimiento cuantitativo muy útil para problemas donde la variable dependiente toma valores en un conjunto finito. Desde la década de los 80 este procedimiento se emplea cada vez con más frecuencia debido a las facilidades computacionales con que se cuenta desde entonces.

La Regresión Logística consiste en estudiar la dependencia funcional entre la variable dependiente categórica Y (dicotómica) que representa la ocurrencia o no de un suceso, y un conjunto de k variables independientes $X = (X_{1i}, X_{2i,\ldots,} X_{pi})$ que pueden ser cuantitativas o categóricas. El modelo de una regresión logística binaria permite predecir en términos de la probabilidad la ocurrencia del evento de interés, tomando así la variable dependiente Y valor 1 si ocurre el suceso (suspenso), y valor 0 si no ocurre el suceso (aprobado). De esta forma se tiene la ecuación:

$$P(Y=1/X) = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_p X_{pi}}}{1+e^{\beta_0+\beta_1 X1_i+\beta_2 X_{2i}+\ldots+\beta_p X_{pi}}}$$

El proceso de inferencia con la Regresión Logística, consiste en aplicar la ecuación estimada al vector de datos X^k_0 para predecir la clasificación del rendimiento académico de un estudiante (Suspenso o Aprobado). La ecuación estimada será:

$$P(Y=1/X^K_0) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{10}+\hat{\beta}_2 X_{20}+\ldots+\hat{\beta}_p X_{P0}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{10}+\hat{\beta}_2 X_{20}+\ldots+\hat{\beta}_p X_{P0}}}$$

4. HIPÓTESIS DE PARTIDA

4.1. ÁMBITO DE CLASIFICACIÓN Y LA NATURALEZA DE LAS VARIABLES

El presente TFG se plantea en el ámbito de la clasificación, técnica que exige que la variable objetivo sea de naturaleza discreta (categórica). Esta restricción, no es aplicable a las variables explicativas que pueden ser tanto continuas (numéricas) como discretas. No obstante, hay que ser conscientes de que hay clasificadores capaces de manejar variables explicativas de un tipo u otro, tal y como se explica en la sub-sección 3.2.

En el caso que nos ocupa, se han factorizado (discretizado) todas las variables que intervienen en el problema, y ello permite utilizar cualquiera de los algoritmos de clasificación existentes en la literatura.

4.2. LA FACTORIZACIÓN COMO PROCESO CRÍTICO

El tratamiento como variable discreta de una variable que originalmente no lo era, implica la pérdida parcial de la información intrínseca a su valor numérico. Sin embargo, en contextos decisionales, en los que se va a aplicar un mismo tratamiento a todos los sujetos que se muevan en una determinada horquilla (segmento) de valores de tal variable, poderla tratar como una variable discreta aporta la gran ventaja de poder aplicar algoritmos computacionalmente menos costosos y que a su vez proporcionan modelos más fácilmente interpretables e implementables en sistemas de apoyo a la decisión.

4.3. MODELOS PREDICTIVOS PRECISOS Y FÁCILES DE INTERPRETAR

De los diferentes modelos predictivos disponibles, se ha elegido el algoritmo CART (función RPART de la librería MachineLearning en R) por varios motivos. En primer lugar, por los altos niveles de precisión que lo caracterizan; en segundo lugar, por la sencillez a la hora de interpretar los árboles que proporciona como resultado; y, por último, por su versatilidad a la hora de ser parametrizado desde R-Studio.

4.4. ¿CÓMO AFECTA LA FACTORIZACIÓN A LA PRECISIÓN?

A la hora de factorizar una variable numérica, se debe decidir el número de segmentos en que se desea dividirla y también hay que elegir el mecanismo a aplicar para encontrar los límites de tales segmentos.

Por otro lado, “parece” evidente, que la forma en que se factoricen las diferentes variables de un problema (tanto las explicativas como la objetivo) dará lugar a data sets de entrada diferentes entre sí, a partir de los cuales se obtendrán predicciones más o menos precisas. Ello es debido a que fruto de diferentes discretizaciones, el conjunto de variables realmente relevantes (y sus pesos relativos) que debe considerar el modelo a su entrada, será diferente en cada caso.

En este TFG se pretende comprobar esta hipótesis de partida y medir hasta qué punto la precisión del modelo final se ve afectada por la factorización realizada previamente, a partir de un sencillo caso de uso en el ámbito educativo.

5. METODOLOGÍA

5.1. LOS DATOS

Estos datos abordan el logro estudiantil en la educación secundaria de dos escuelas portuguesas. Se proporcionan dos conjuntos de datos sobre el rendimiento en dos asignaturas distintas: **student-mat.csv** (curso de matemáticas) y **student-por.csv** (curso de idioma portugués) con 395 y 649 individuos respectivamente. En (Cortez y Silva, 2008), los dos conjuntos de datos se modelaron bajo tareas de clasificación y regresión binarias / de cinco niveles.

La calidad de los patrones que se obtienen con la minería de datos es directamente proporcional a la calidad de los datos utilizados. Esta fase es la responsable de obtener datos de alta calidad. Para lograr este objetivo se ha buscado detectar valores anómalos (outliers) y datos faltantes. Para la detección de los valores anómalos y de los datos faltantes se realizó un análisis exploratorio. Del resultado de este análisis se deduce la no existencia tanto de valores anómalos como de datos faltantes.

Los atributos que se recopilaron mediante el uso de informes y cuestionarios escolares de los conjuntos de datos de student-mat.csv (curso de matemáticas) y student-por.csv (curso de idioma portugués) para cada estudiante incluyen calificaciones de los estudiantes, características demográficas, sociales y relacionadas con la escuela. A continuación, se muestra la naturaleza estos atributos:

1. **escuela** - escuela de estudiantes (binario: 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira)
2. **sexo** - sexo del estudiante (binario: 'F' - femenino o 'M' - masculino)
3. **edad** - edad del estudiante (numérico: de 15 a 22)
4. **dirección** - tipo de domicilio del estudiante (binario: 'U' - urbano o 'R '- rural)
5. **famsize** - tamaño de la familia (binario:' LE3 '- menor o igual a 3 o' GT3 '- mayor que 3)
6. **Pstatus** - estado de convivencia de los padres (binario:' T '- viviendo juntos o' A '- aparte)
7. **Medu** - educación de la madre (numérico: 0 - ninguno, 1 - educación primaria (4º grado), 2 - 5º a 9º grado, 3 - educación secundaria o 4 - educación superior)
8. **Fedu** - educación del padre (numérico: 0 - ninguno, 1 - educación primaria (4to grado), 2 - 5to a 9no grado, 3 - educación secundaria o 4 - educación superior)
9. **Mjob** - trabajo de la madre (nominal: 'maestro', 'salud' relacionado con el cuidado, civil 'servicios' (por ejemplo, administrativo o policial), 'at_home' u 'otro')
10. **Fjob** - trabajo del padre (nominal: 'maestro', 'salud' relacionado con el cuidado, civil 'servicios' (por ejemplo, administración o policía), 'at_home' u 'otro')
11. **razón** - razón para elegir esta escuela (nominal: cerca de 'hogar', 'reputación' de la escuela, 'preferencia' del curso u 'otra')
12. **tutor** - tutor del estudiante (nominal: 'madre', 'padre' u 'otro')
13. **tiempo de viaje**: tiempo de viaje de la casa a la escuela (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. A 1 hora, o 4 -> 1 hora)
14. **tiempo de estudio** - tiempo de estudio semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, o 4 -> 10 horas)

15. **faltas** - número de faltas de clases pasadas (nominal: 0 - 0 faltas, 1 - 1 falta, 2 - 2 faltas, 3 - 3 o más faltas)
16. **schoolsup** - apoyo educativo extra (binario: sí o no)
17. **famsup** - apoyo educativo familiar (binario: sí o no)
18. **pagadas** - clases extra pagadas dentro de la materia del curso (matemáticas o portugués) (binario: sí o no)
19. **actividades** - actividades extracurriculares (binario: sí o no)
20. **guardería** - escuela infantil asistida (binario: sí o no)
21. **más alto** - desea obtener educación superior (binario: sí o no)
22. **Internet**: acceso a Internet en casa (binario: sí o no)
23. **romántico** - con una relación romántica (binario: sí o no)
24. **famrel** - calidad de las relaciones familiares (numérico: de 1 - muy malo a 5 - excelente)
25. **tiempo libre** - tiempo libre después de la escuela (numérico: de 1 - muy bajo a 5 - muy alto)
26. **salidas** - salir con amigos (numérico: de 1 - muy bajo a 5 - muy alto)
27. **Dalc** - consumo de alcohol en días hábiles (numérico: de 1 - muy bajo a 5 - muy alto)
28. **Walc** - consumo de alcohol los fines de semana (numérico: de 1 - muy bajo a 5 - muy alto)
29. **salud** - estado de salud actual (numérico: de 1 - muy malo a 5 - muy bueno)
30. **ausencias** - número de ausencias escolares (numérico: de 0 a 93)

Estas **calificaciones** están relacionadas con la **materia del curso,** matemáticas o portugués:

31. G1 – **calificación primer grado** (numérico: de 0 a 20)
32. G2: **calificación del segundo período** (numérico: de 0 a 20)
33. G3: **calificación final** (numérico: de 0 a 20, objetivo de salida)

Tabla 3. Variables procesadas relacionadas con el estudiante
Fuente: Cortez y Silva (2008)

El método de clasificación exige que la variable objetivo sea discreta por lo que se realizarán *distintas discretizaciones* de las variables **G1, G2 y G3** que son las **calificaciones que se utilizarán como variables a predecir** y a continuación, se llevarán a cabo una serie de métodos de selección de atributos. A partir de los cuales se elaborarán ***distintos rankings de variables*** de los que se obtendrán las conclusiones correspondientes.

El estudio va a consistir en analizar cómo afectan diferentes discretizaciones de las variables numéricas a la precisión de los clasificadores.

Al analizar los datos, lo primero que conviene hacer es formar una idea lo más exacta posible acerca de sus características. Para ello, es necesario realizar un análisis descriptivo de los datos. En él, se lleva a cabo un estudio de las relaciones existentes entre las diferentes variables y la forma en la que éstas afectan a la variable output *calificación final (G3).*

Las técnicas estadísticas utilizadas para este análisis son:

- ***Gráficos univariantes***: representan los datos disponibles en una variable de forma individual. Se utilizan los siguientes tipos: gráfico circular y gráficos de barras.
- ***Gráficos bivariantes***: reflejan las relaciones existentes entre dos variables. La mayoría de gráficos utilizados son gráficos de cajas y gráficos de barras.
- ***Gráficos multivariantes***: representan las relaciones existentes entre más de dos variables. Estas relaciones son representadas sobre gráficos de barras.
- ***Tablas***: muestran los resúmenes estadísticos de las diferentes variables.

A la vez que se realiza este análisis, se llevan a cabo diferentes contrastes de hipótesis entre las variables mediante el empleo de pruebas no paramétricas o de distribución libre basadas en los rangos de distribución de la variable.

El uso de estas pruebas se debe a que la distribución de la variable en estudio ***calificación final (G3)***, una vez realizado el test de normalidad de Shapiro-Wilk y obtenido un p-valor de 8.836×10^{-13}, se concluye que no cumple normalidad.

Particularmente, uno de los test utilizados es el test de Mann-Whitney-Wilcoxon, que es un test no paramétrico que contrasta si dos muestras proceden de poblaciones equidistribuidas. Una extensión de este es el test de Kruskall-Wallis que se utiliza para más de dos grupos (k poblaciones), también conocido como test H. Este test es la alternativa no paramétrica al test ANOVA y, a diferencia de éste, en el que se comparan medias, el test de Kruskall-Wallis contrasta si las diferentes muestras están equidistribuidas y que por lo tanto pertenecen a una misma distribución (población). Bajo ciertas simplificaciones, puede considerarse que ambos tests comparan las medianas.

El contraste de hipótesis es el siguiente:

H_0: todas las muestras provienen de la misma población (distribución).

H_1: al menos una muestra proviene de una población con una distibución distinta.

Para llevar a cabo este análisis, el software del que se dispone es RStudio. Este programa ofrece una amplia variedad de librerías, en este caso la más utilizada para la representación gráfica es ggplot2. Por otro lado, las funciones utilizadas para los contrastes de hipótesis son kruskal.test y wilcox.test.

RStudio está disponible para Windows, Mac y Linux o para navegadores conectados a RStudio Server o RStudio Server Pro (Debian / Ubuntu, RedHat / CentOS, y SUSE Linux) y tiene la misión de proporcionar el entorno informático estadístico R. Permite un análisis y desarrollo para poder analizar los datos con R.

En cuanto al hardware, el ordenador utilizado no es demasiado potente ya que los cálculos a realizar no lo requieren. Las especificaciones del dispositivo son las siguientes: el nombre es LAPTOP-A22B42J3, con un procesador Intel(R) Core (TM) i7-8850U CPU @ 1.80GHz 1.99GHz, memoria RAM de 16,0GB (15,9BG usable) y cuyo sistema operativo es de 64 bits, procesador basado en x64.

Se presentan a continuación las diferentes técnicas utilizadas para estudiar las relaciones existentes entre las distintas variables y la forma en la que estas afectan a la variable objetivo ***calificación final (G3)***. Esto se lleva a cabo para los dos conjuntos de datos de las dos asignaturas (matemáticas y lengua portuguesa), teniendo en cuenta que el número de individuos que cursan cada una de ellas es distinto: 395 y 649 respectivamente.

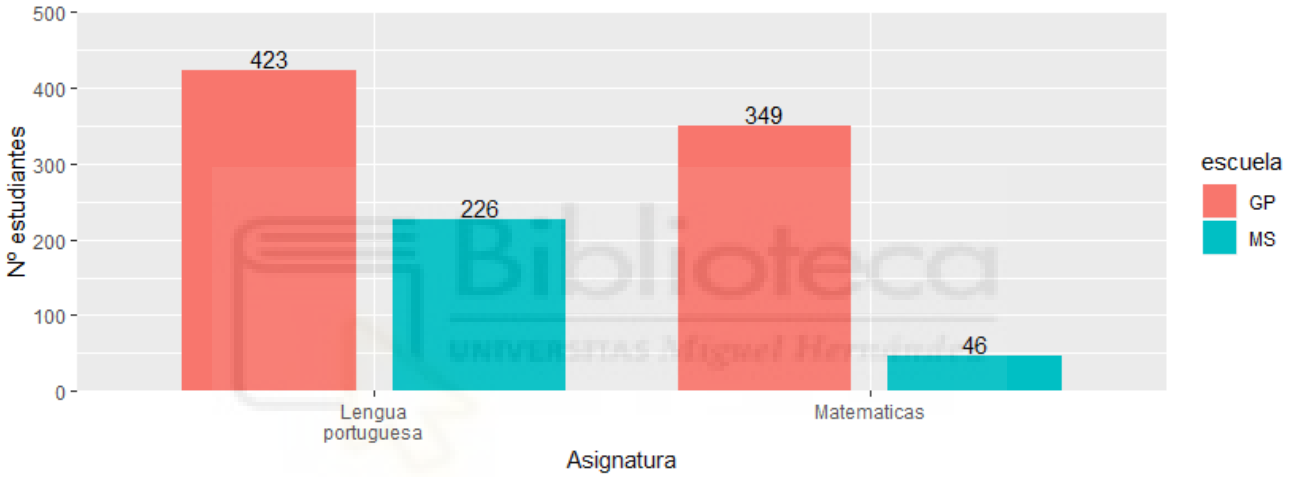

Fig. 1. Número de estudiantes por escuela en función de la asignatura cursada
Fuente: elaboración propia

En la figura 1 se observa que el número de estudiantes que cursan la asignatura de matemáticas en la escuela Gabriel Pereira es mucho mayor que el de la escuela Mousinho da Silveira. En el caso de Lengua portuguesa es también notable la diferencia, aunque en menor medida.

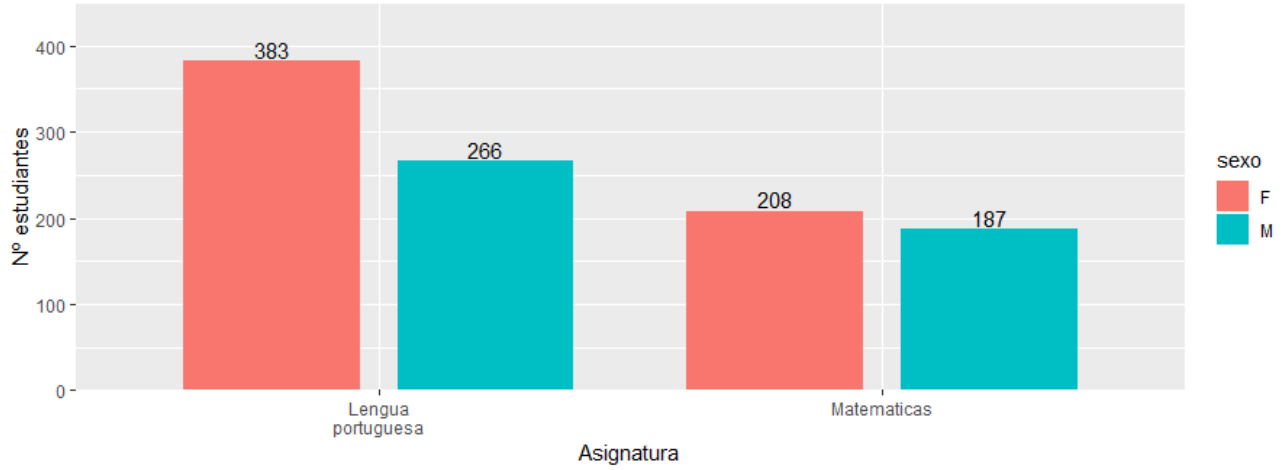

Fig. 2. Número de estudiantes por sexo en función de la asignatura cursada
Fuente: elaboración propia

De acuerdo con la figura 2 concluimos que la mayoría de los alumnos que estudian matemáticas pertenecen al sexo femenino, no apreciándose una diferencia notable con respecto al otro sexo. Lo mismo ocurre en la asignatura de lengua portuguesa, siendo mayor la diferencia entre ambos sexos.

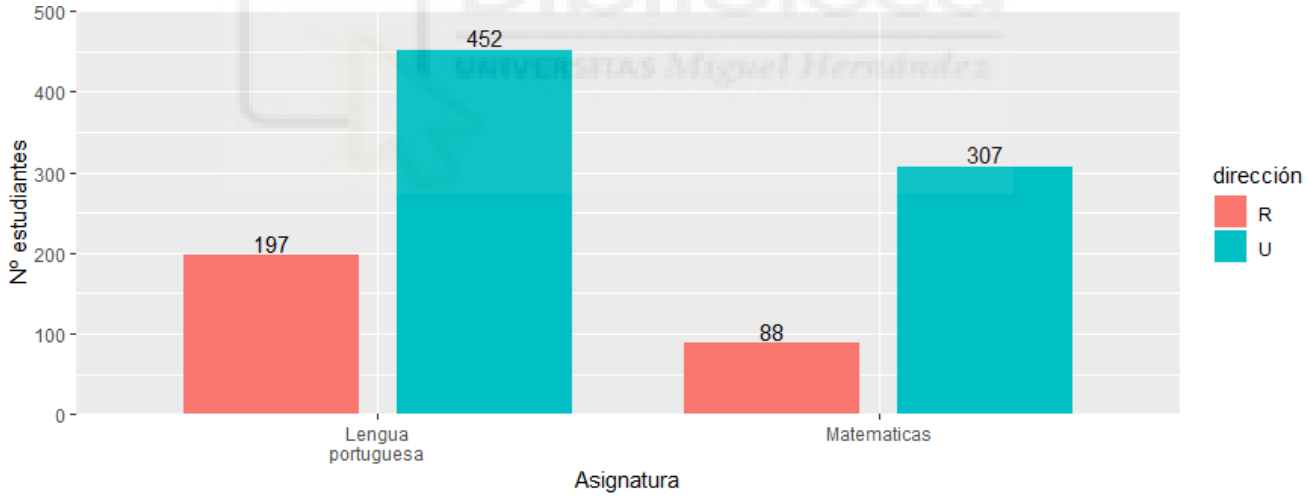

Fig. 3. Número de estudiantes en función de la dirección y la asignatura cursada
Fuente: elaboración propia

En cuanto al tipo de domicilio del estudiante (Fig. 3), destaca con un porcentaje elevado la zona urbana en los alumnos tanto de la asignatura de matemáticas como de lengua portuguesa.

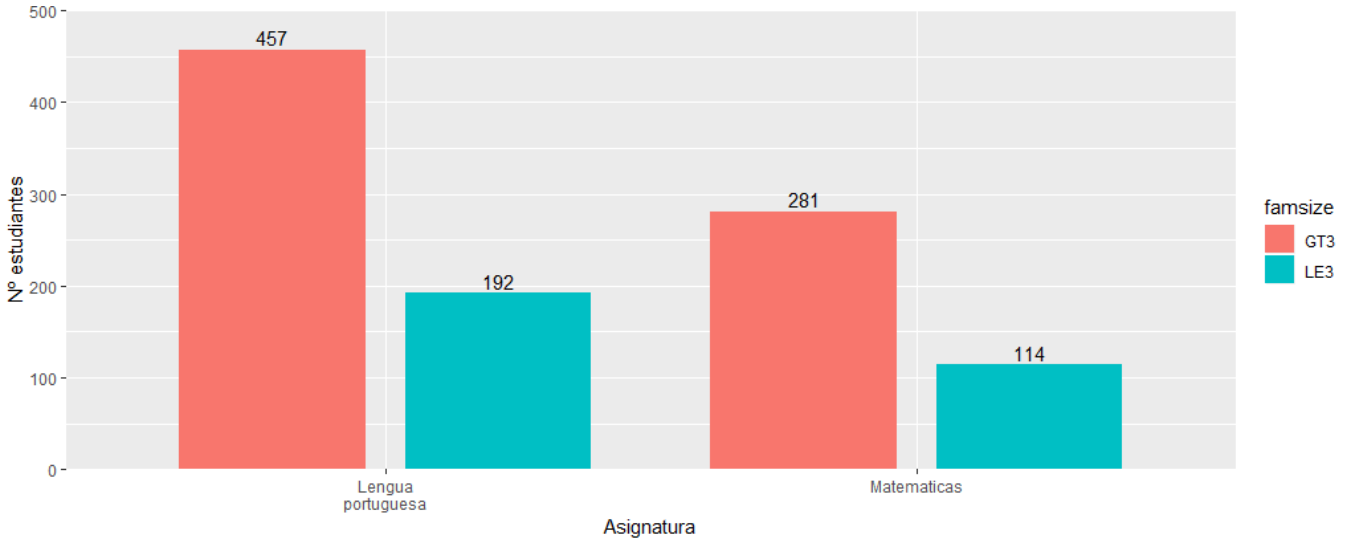

Fig. 4. Número de estudiantes en función del tamaño de la familia y la asignatura cursada Fuente: elaboración propia

En la figura 4 se observa que tanto en los estudiantes de la asignatura de matemáticas, como en los de lengua portuguesa, predomina el tamaño de la familia mayor que tres miembros con un porcentaje alto en ambas asignaturas.

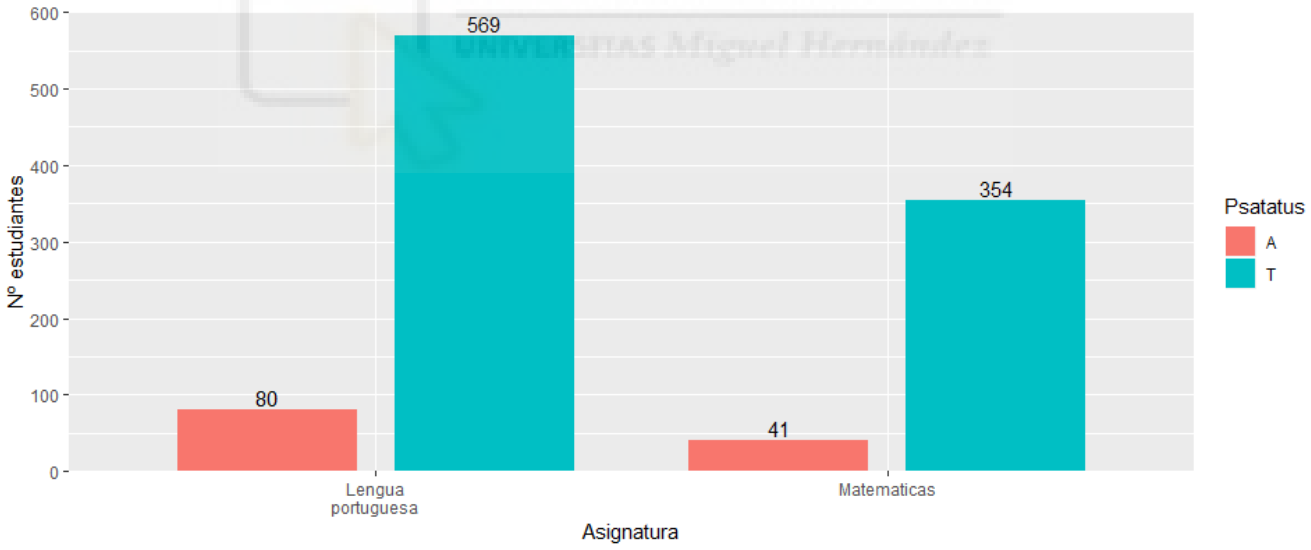

Fig. 5. Número de estudiantes en función del estado de convivencia de los padres y la asignatura cursada. Fuente: elaboración propia

Observado la figura 5, se puede concluir que los padres del 90% de los estudiantes, aproximadamente, que cursan la asignatura de matemáticas y la de lengua portuguesa viven en pareja estable.

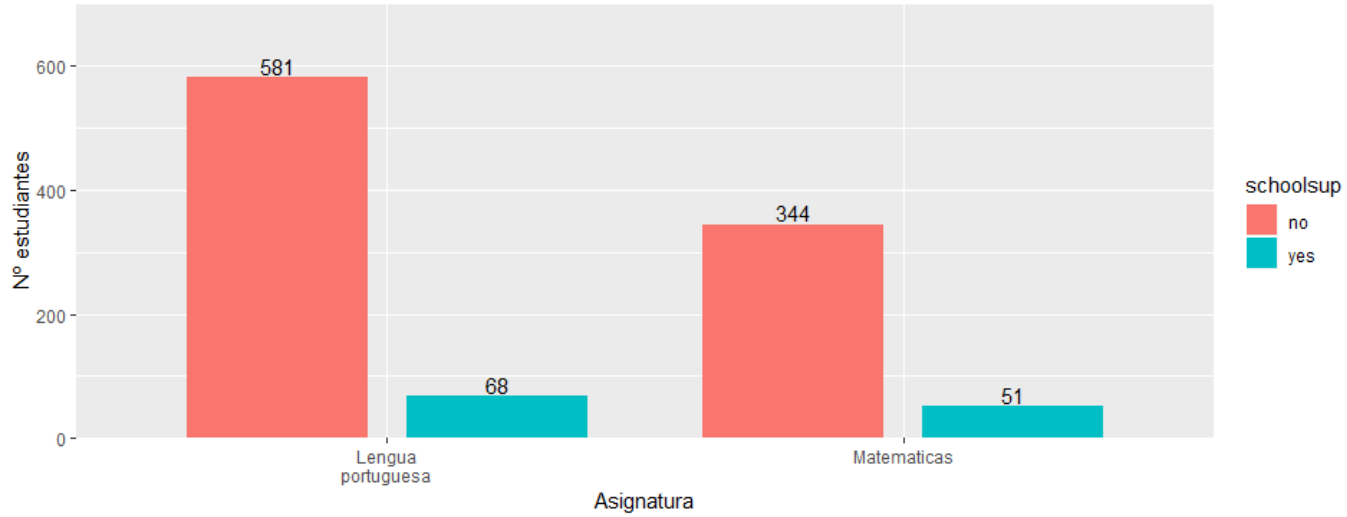

Fig. 6. Número de estudiantes en función del apoyo educativo extra y la asignatura cursada
Fuente: elaboración propia

Según el estudio realizado (Fig. 6) el 87% de los estudiantes que cursan la asignatura de matemáticas no recibe apoyo educativo extra, lo mismo ocurre con el 89.5% de los estudiantes que cursan la asignatura lengua portuguesa.

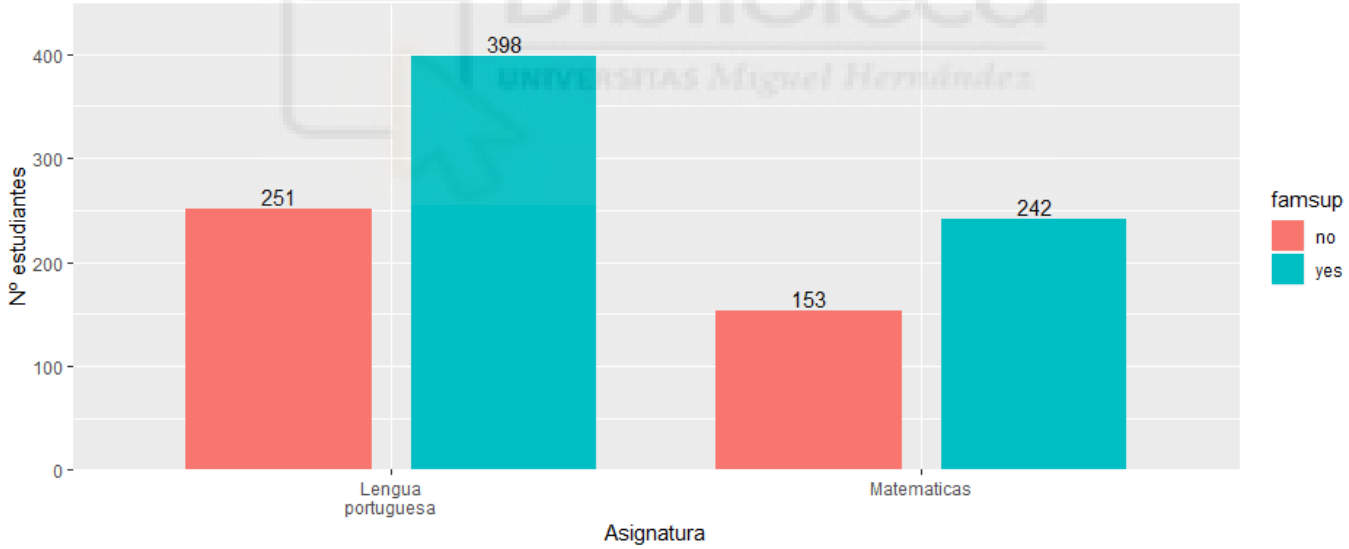

Fig. 7. Número de estudiantes en función del apoyo educativo familiar y la asignatura cursada
Fuente: elaboración propia

Aproximadamente (Fig. 7) un 61% de los estudiantes que cursan tanto la asignatura de matemáticas como la de lengua portuguesa reciben apoyo educativo por parte de familiares.

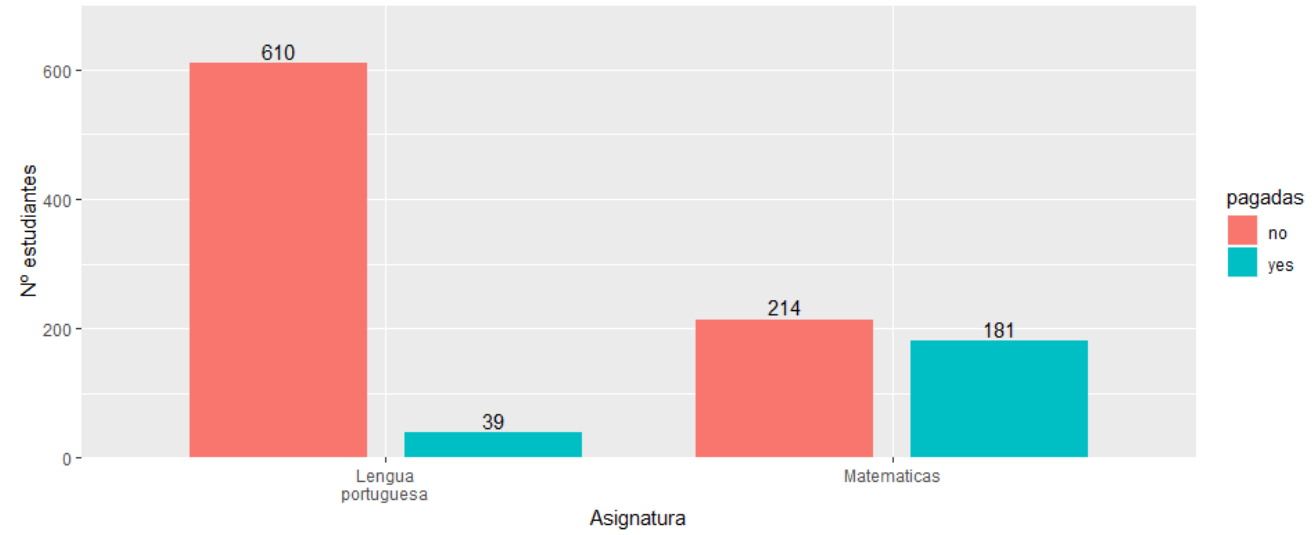

Fig. 8. Número de estudiantes en función de las clases extras pagadas y la asignatura cursada. Fuente: elaboración propia

En este caso, la figura 8 indica que para la asignatura de matemáticas las clases extras pagadas por los estudiantes son aproximadamente el 46%, sin embargo, para la asignatura lengua portuguesa este porcentaje es mucho menor (6%).

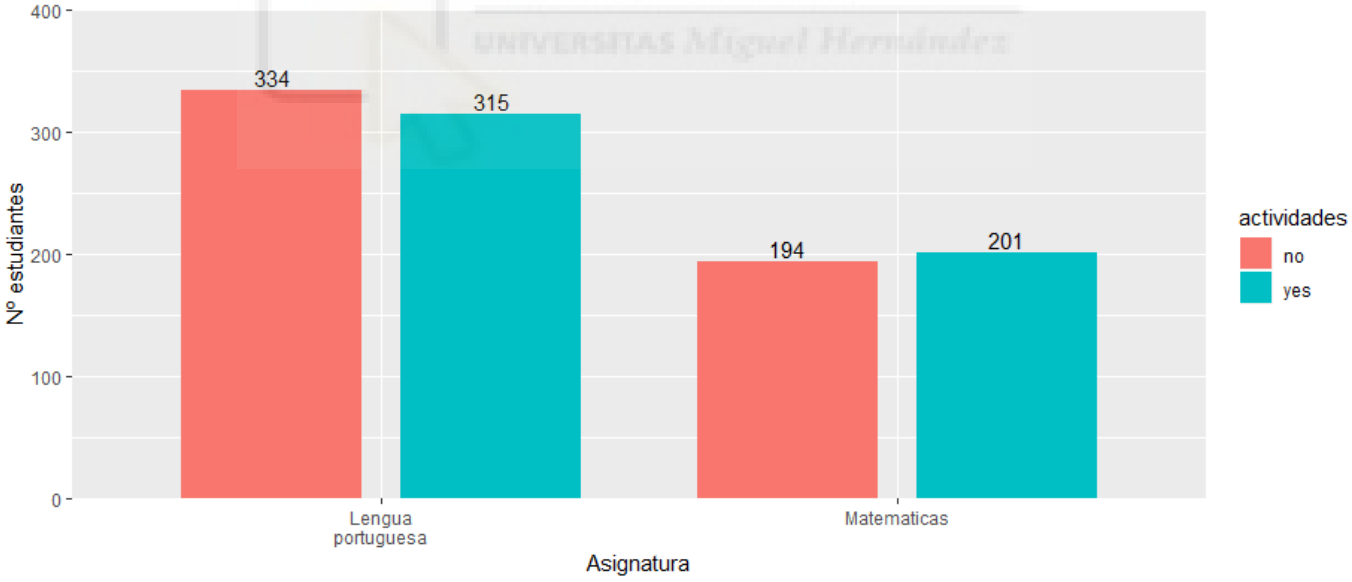

Fig. 9. Número de estudiantes en función de las actividades extracurriculares y la asignatura cursada. Fuente: elaboración propia

Según la figura anterior (Fig. 9), el número de estudiantes con actividades extracurriculares es similar al de estudiantes sin actividades extracurriculares tanto en matemáticas como en lengua portuguesa.

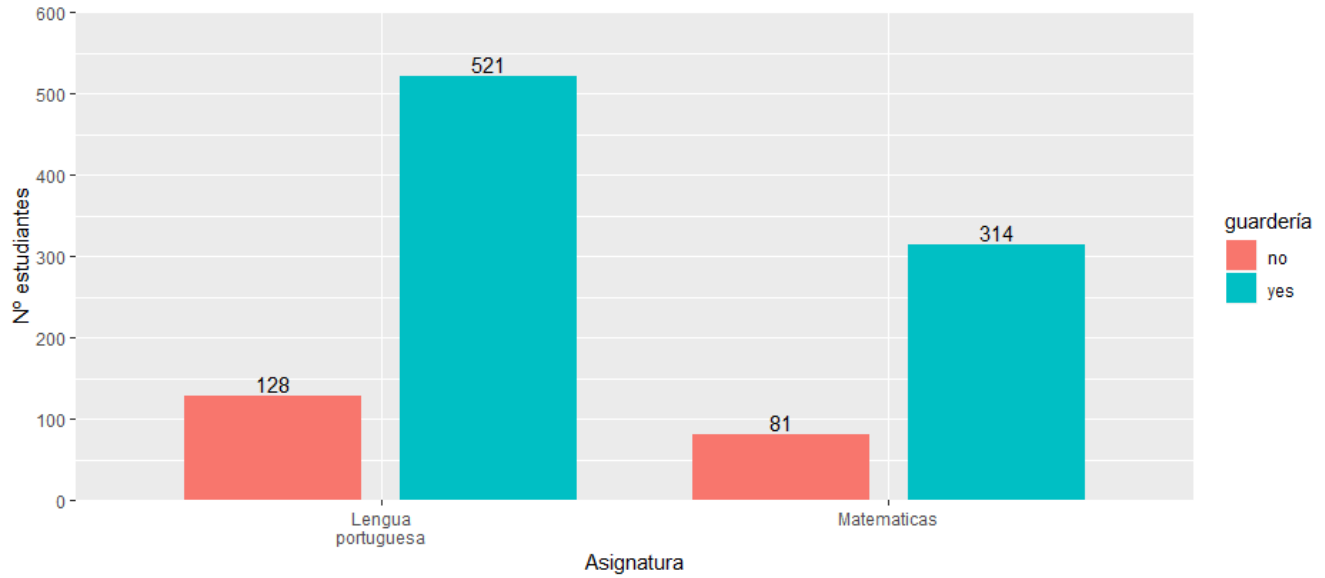

Fig. 10. Número de estudiantes en función de la escuela infantil asistida y la asignatura cursada. Fuente: elaboración propia

En la figura 10 se observa que el número de estudiantes que han asistido a una escuela infantil es mayor en ambas asignaturas (matemáticas y lengua portuguesa).

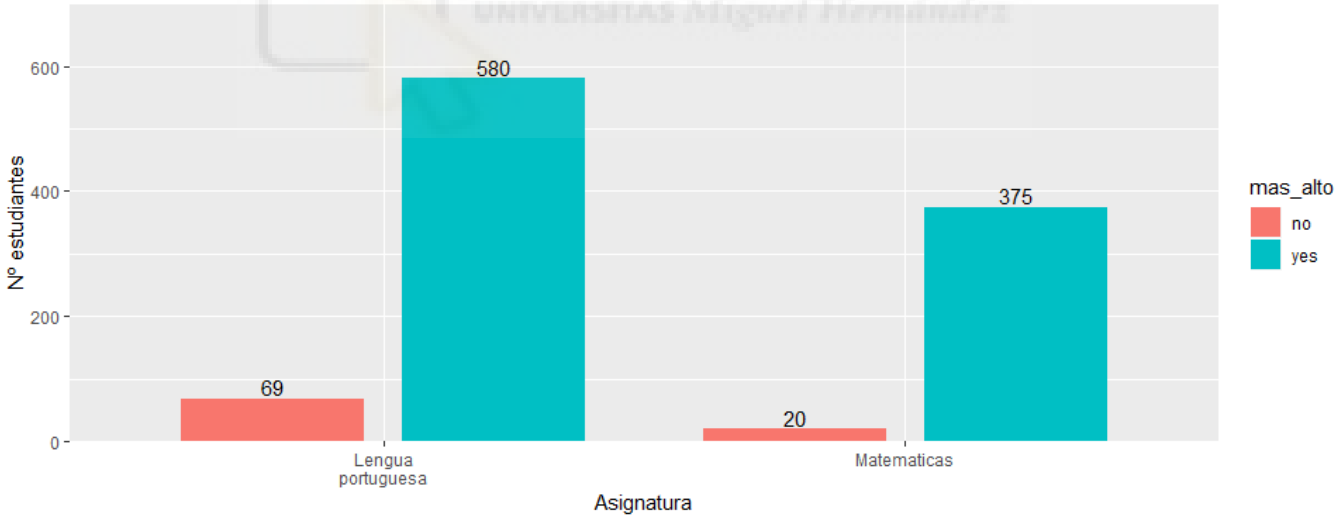

Fig. 11. Número de estudiantes en función del deseo de obtener educación superior y la asignatura cursada. Fuente: elaboración propia

El número de estudiantes, según se aprecia en la figura 11, que desea obtener educación superior es mucho mayor que el que no lo desea, tanto en la asignatura de matemáticas como en la de lengua portuguesa.

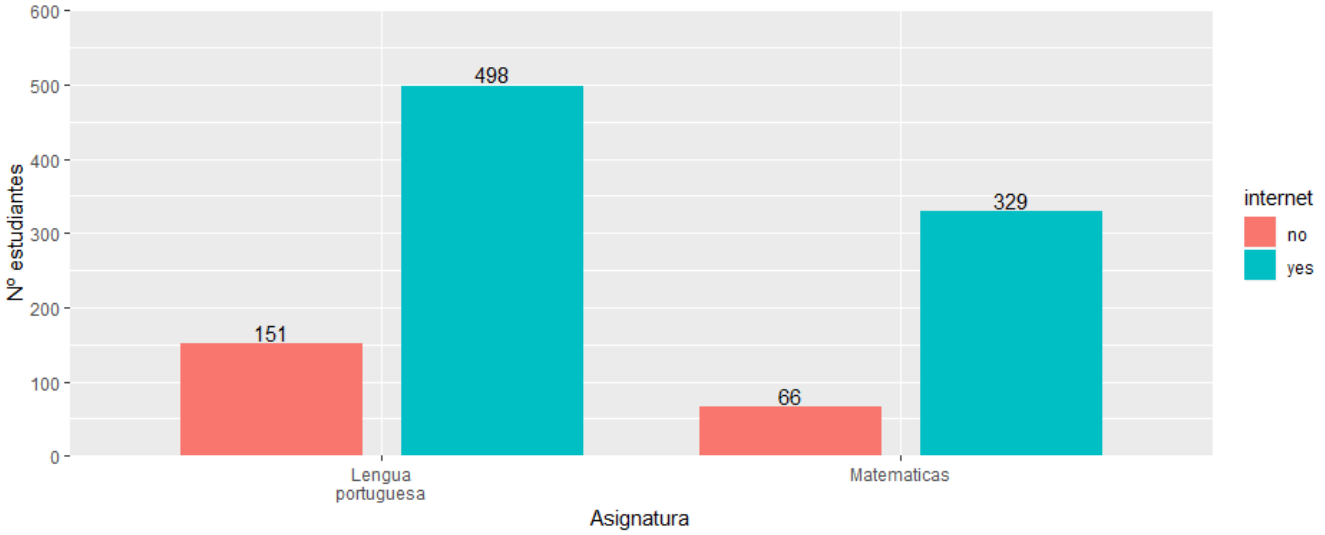

Fig. 12. Número de estudiantes en función del acceso a internet en casa y la asignatura cursada. Fuente: elaboración propia

En la figura 12 se puede observar que la mayoría de los alumnos que estudian la asignatura de matemáticas (83.3%) tienen acceso a internet en casa. Lo mismo ocurre con el 76.7% de los alumnos que cursan la asignatura lengua portuguesa.

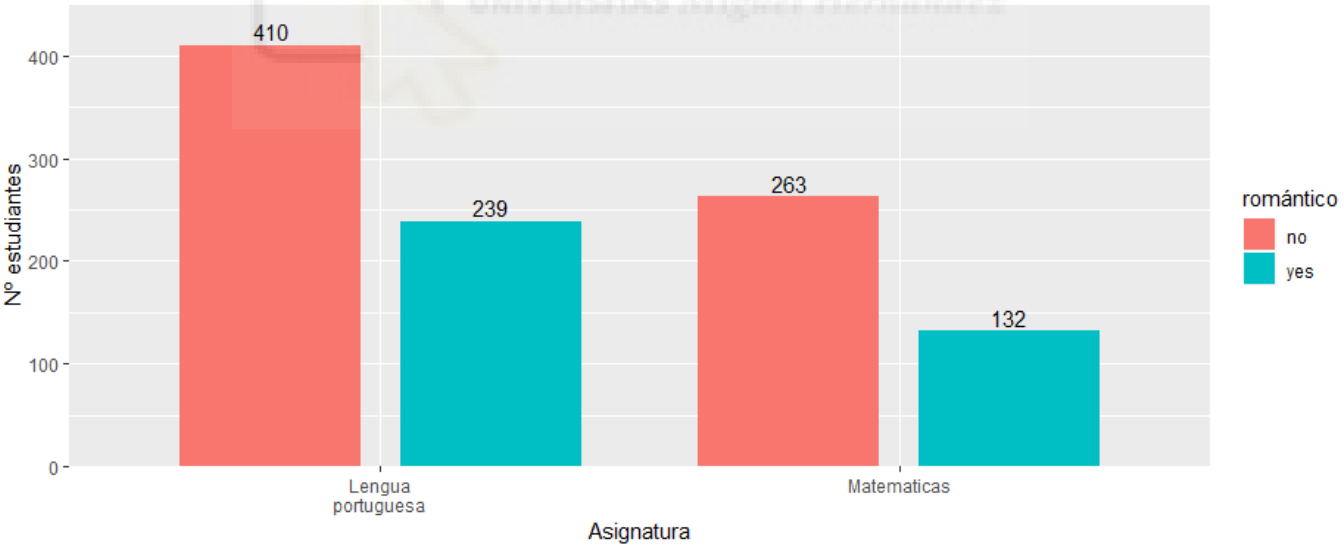

Fig. 13. Número de estudiantes en función de si tienen o no relación sentimental y la asignatura cursada. Fuente: elaboración propia

En cuanto al número de estudiantes con una relación sentimental (Fig. 13), es mayor el porcentaje de alumnos en ambas asignaturas que no tienen una relación sentimental.

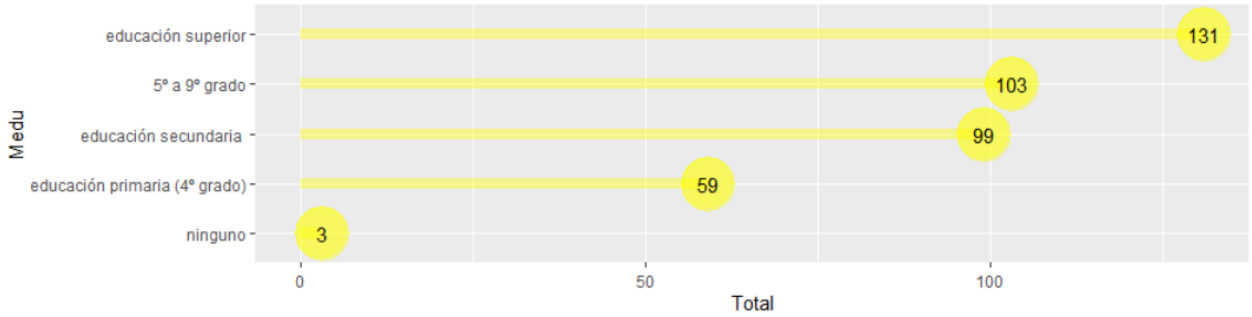

Fig. 14. Número de estudiantes en función de la educación de la madre (matemáticas)
Fuente: elaboración propia

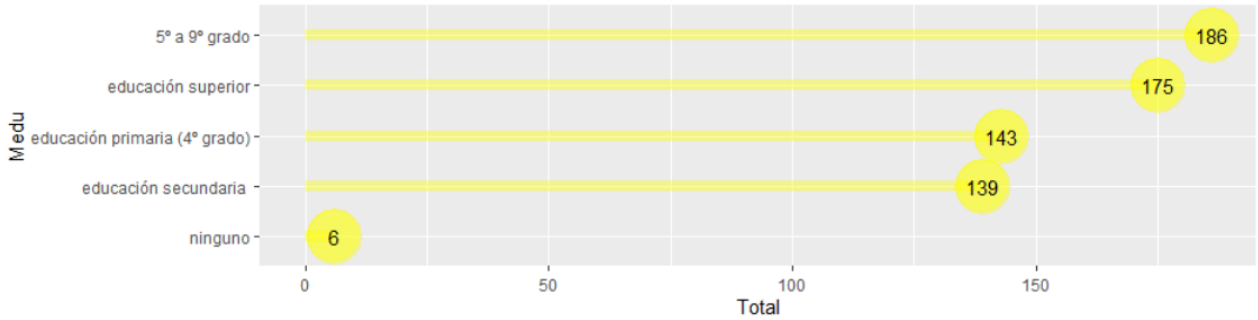

Fig. 15. Número de estudiantes en función de la educación de la madre (lengua portuguesa)
Fuente: elaboración propia

Observando las figuras 14 y 15, se concluye que las madres de los alumnos que cursan la asignaturas matemáticas y lengua portuguesa tienen un nivel educativo medio-alto.

Fig. 16. Número de estudiantes en función de la educación del padre (matemáticas)
Fuente: elaboración propia

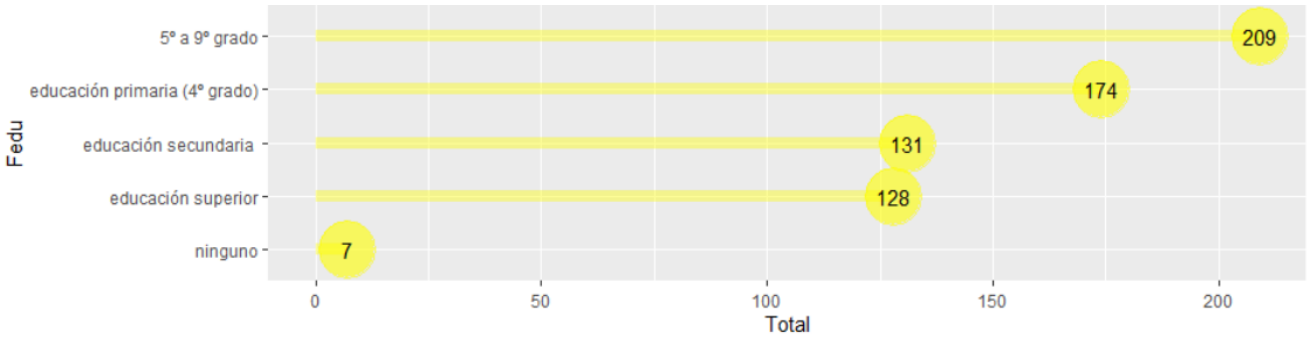

Fig. 17. Número de estudiantes en función de la educación del padre (lengua portuguesa)
Fuente: elaboración propia

Se concluye (Fig. 16 y Fig. 17) que los niveles educativos de los padres de los alumnos que cursan ambas asignaturas se reparten de forma homogénea en todos niveles educativos.

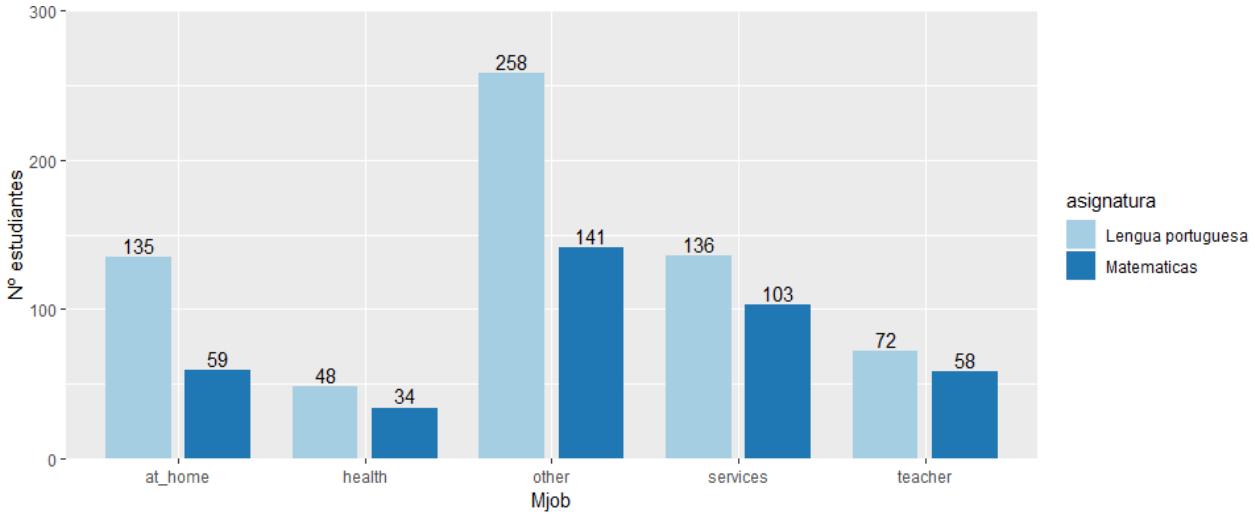

Fig. 18. Número de estudiantes en función del trabajo de la madre en las asignaturas de matemáticas y lengua portuguesa. Fuente: elaboración propia

Se observa en la figura 18 que los porcentajes más elevados en la actividad laboral de las madres de los alumnos que cursan ambas asignaturas se encuentran en el sector servicios y en otros sectores.

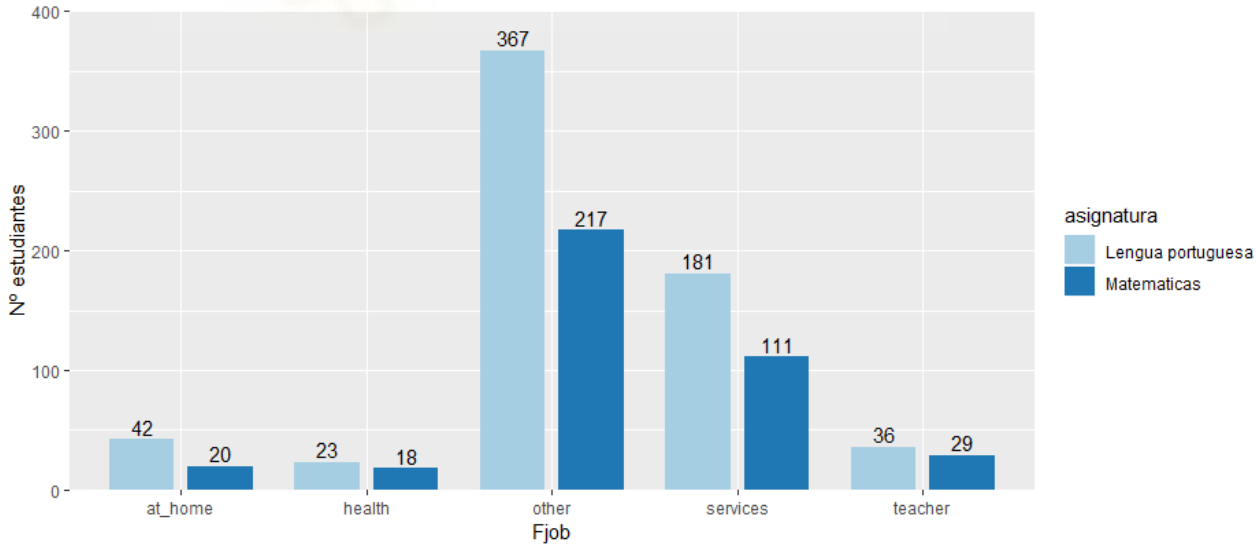

Fig. 19. Número de estudiantes en función del trabajo del padre en las asignaturas de matemáticas y lengua portuguesa. Fuente: elaboración propia

En la figura 19 se aprecia que los porcentajes más elevados en la actividad laboral de los padres de los alumnos que cursan ambas asignaturas se encuentran en el sector servicios y en otros sectores principalmente.

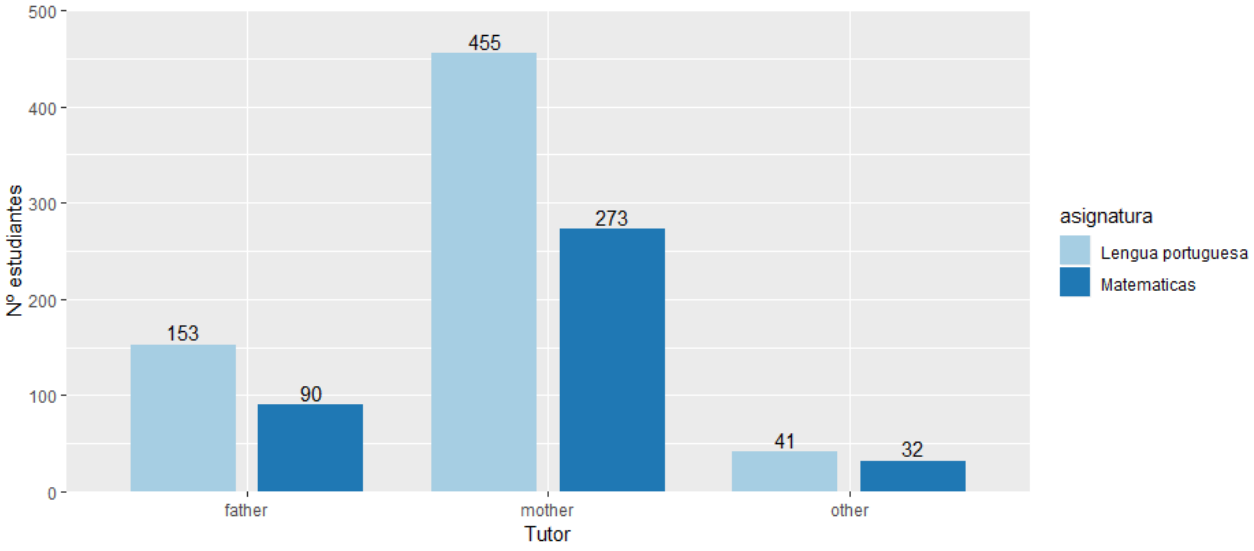

Fig. 20. Número de estudiantes en función del tutor y la asignatura
Fuente: elaboración propia

En la figura 20 se observa que la persona que ejerce como tutor de los alumnos que cursan ambas asignaturas corresponde en su mayoría a la madre.

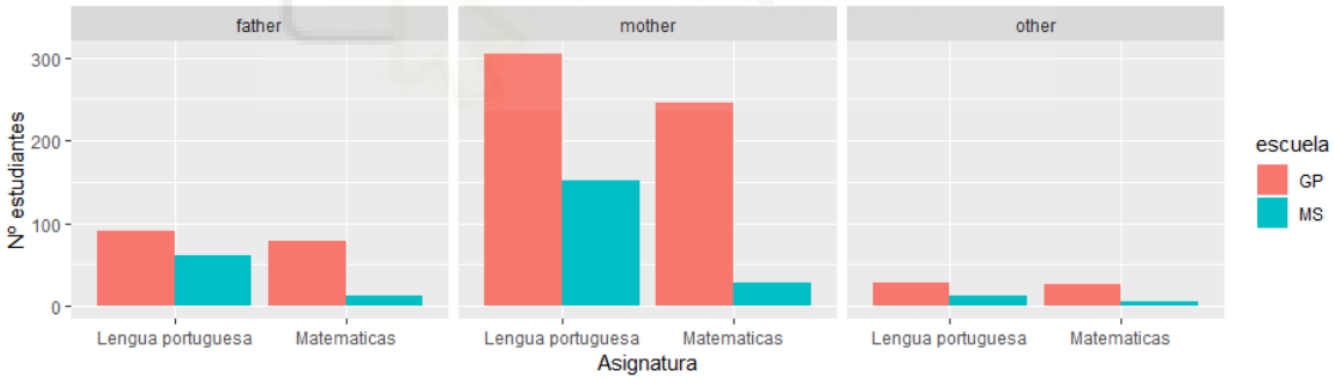

Fig. 21. Número de estudiantes en función de la escuela por asignatura y tutor
Fuente: elaboración propia

En la figura 21 se aprecia que la persona que ejerce como tutor de los alumnos que cursan ambas asignaturas, en este caso diferenciando ambas escuelas, corresponde en su mayoría a la madre.

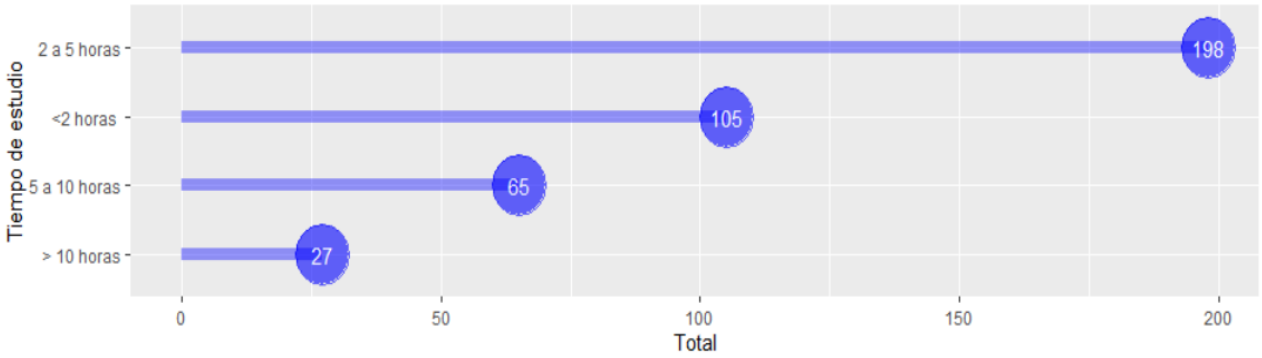

Fig. 22. Número de estudiantes en función del tiempo de estudio semanal (matemáticas) Fuente: elaboración propia

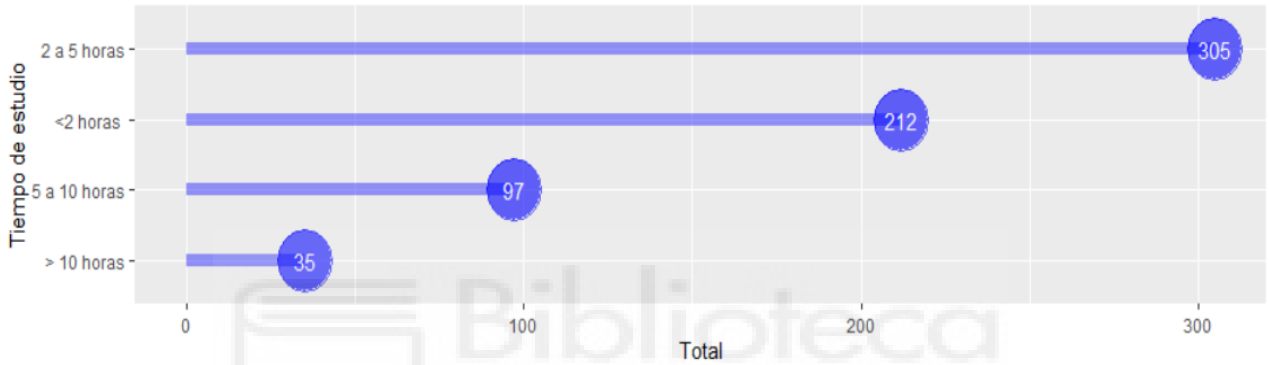

Fig. 23. Número de estudiantes en función del tiempo de estudio semanal (lengua portuguesa). Fuente: elaboración propia

En las figuras 22 y 23 se refleja que la mayoría de los alumnos que estudian ambas asignaturas dedican de dos a cinco horas semanales. Existe una minoría que dedica más de diez horas a la semana.

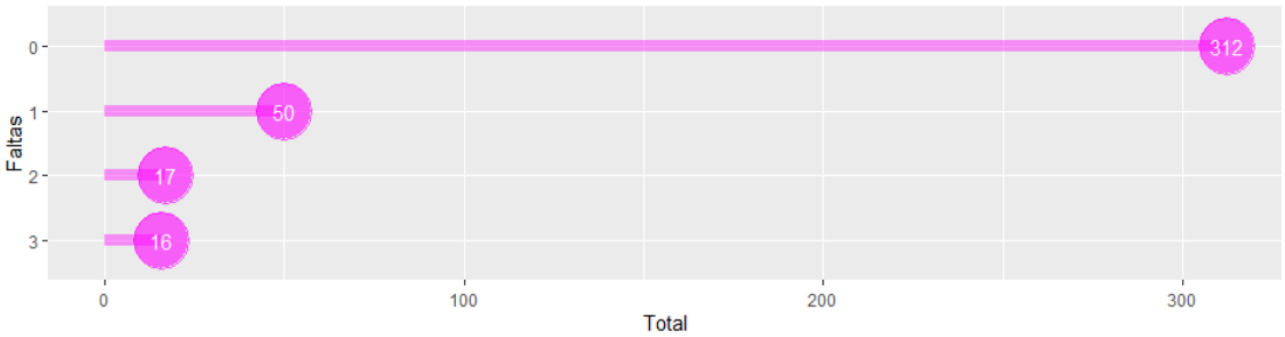

Fig. 24. Número de estudiantes en función del número de faltas a clase (matemáticas) Fuente: elaboración propia

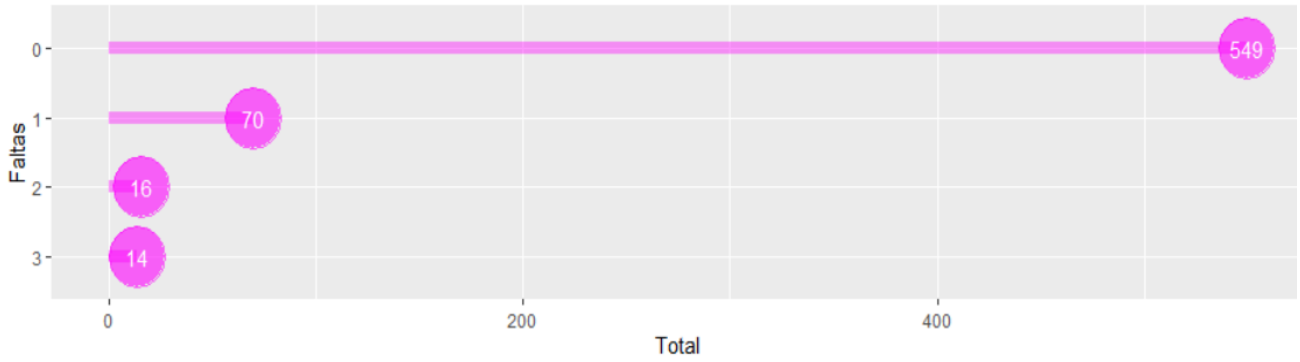

Fig. 25. Número de estudiantes en función del número de faltas a clase (lengua portuguesa)
Fuente: elaboración propia

En las figuras 24 y 25 se aprecia que la mayoría de los alumnos que estudian ambas asignaturas no tienen ninguna falta de asistencia a clase.

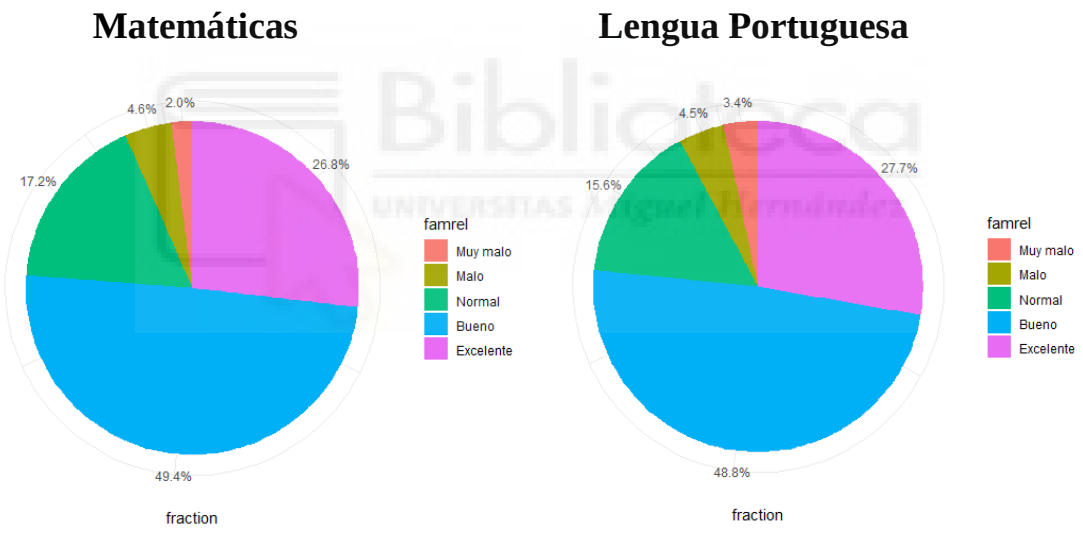

Fig. 26. Porcentaje de estudiantes en función de la calidad de las relaciones familiares
Fuente: elaboración propia

En la figura 26, podemos destacar que la calidad de las relaciones familiares en los estudiantes de enseñanza secundaria en ambas asignaturas es en su mayoría buena o excelente.

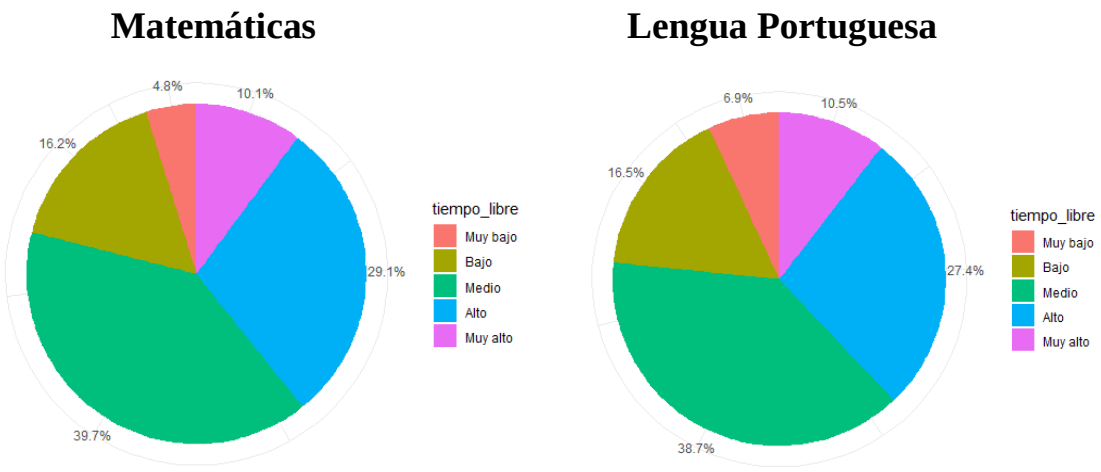

Fig. 27. Porcentaje de estudiantes en función del tiempo libre después de la escuela
Fuente: elaboración propia

En la figura 27, se aprecia que los estudiantes de enseñanza secundaria, en ambas asignaturas, disponen de tiempo libre después de la escuela en un rango medio/alto, incluyendo este más del 65% de los resultados.

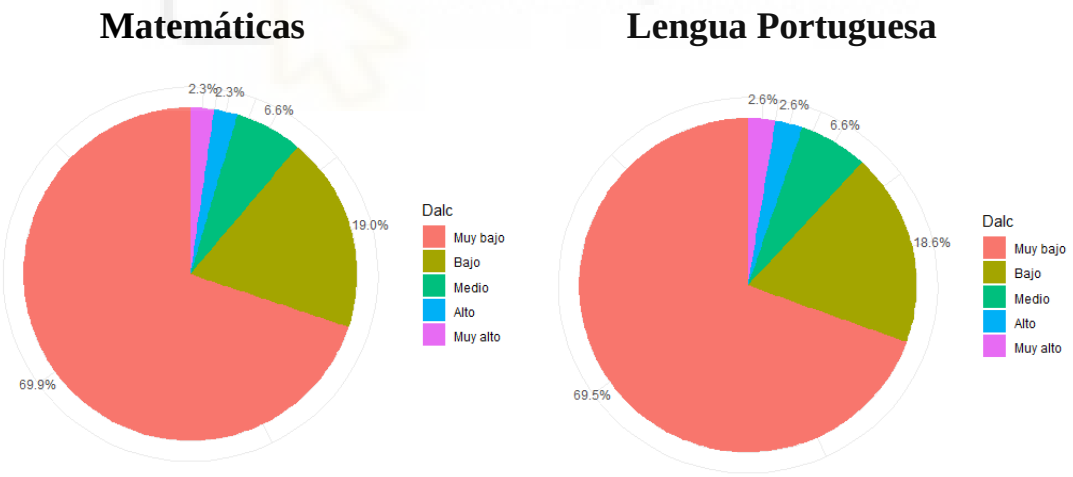

Fig. 28. Porcentaje de estudiantes en función del consumo de alcohol en días hábiles
Fuente: elaboración propia

En la figura 28, se refleja que el consumo de alcohol de los estudiantes de enseñanza secundaria de ambas asignaturas en días hábiles es notoriamente bajo.

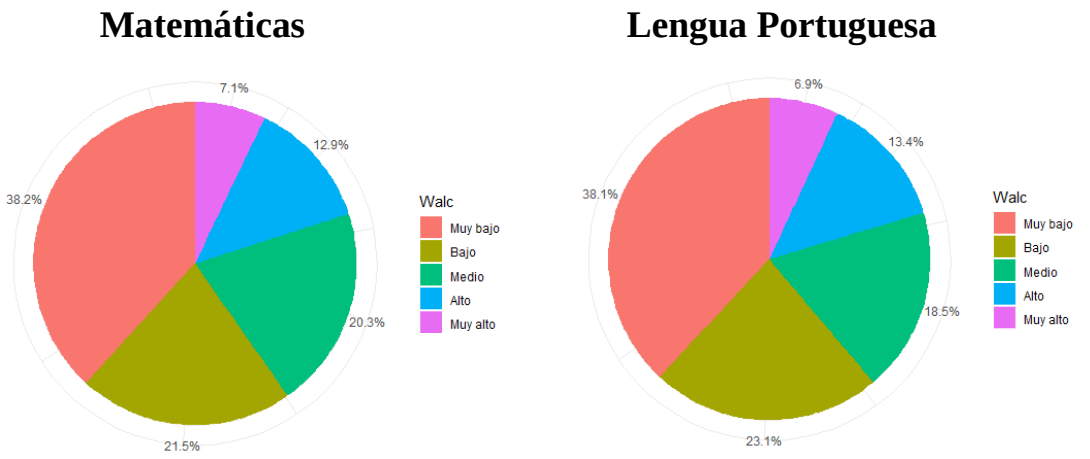

Fig. 29. Porcentaje de estudiantes en función del consumo de alcohol los fines de semana
Fuente: elaboración propia

En la figura 29, se observa que el consumo de alcohol durante los fines de semana de los estudiantes de enseñanza secundaria es en su mayoría bajo o muy bajo. Sumando de esta forma más de un 50% de los resultados obtenidos en ambas asignaturas.

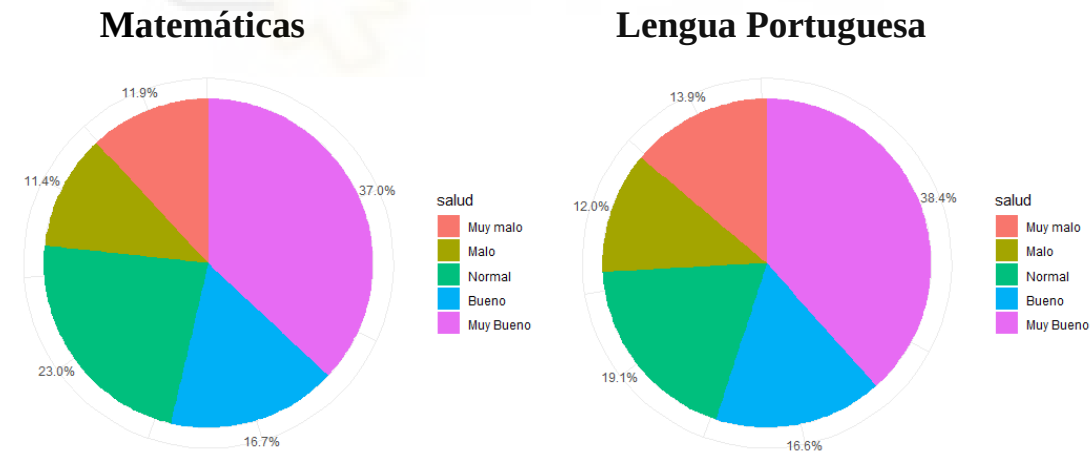

Fig. 30. Porcentaje de estudiantes en función del estado de salud actual
Fuente: elaboración propia

Como se puede apreciar en la figura 30, el estado de salud actual en los estudiantes de enseñanza secundaria en ambas asignaturas se encuentra con un mayor porcentaje en la franja entre normal y muy bueno.

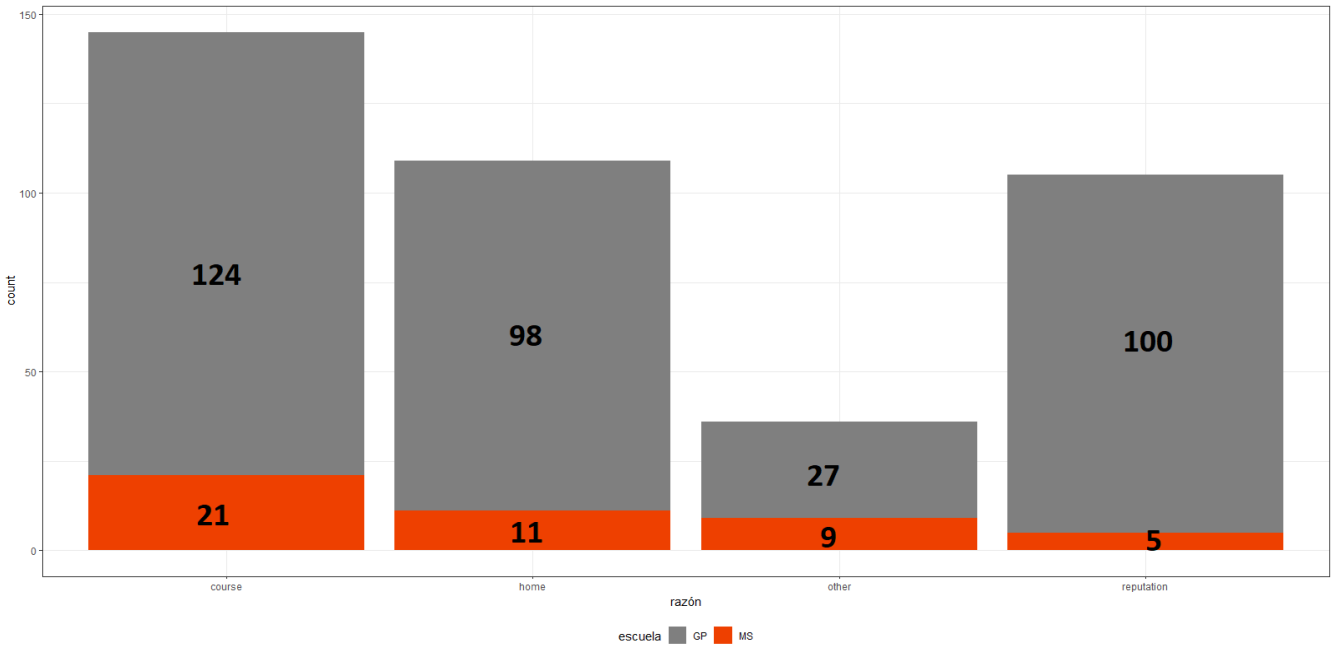

Fig. 31. Número de estudiantes en función de la razón de elección de la escuela
Fuente: elaboración propia

En esta figura 31 se observa la razón por la que se elige cada escuela. En ambos casos la razón principal de esta elección es la preferencia de los estudios posteriores a elegir.

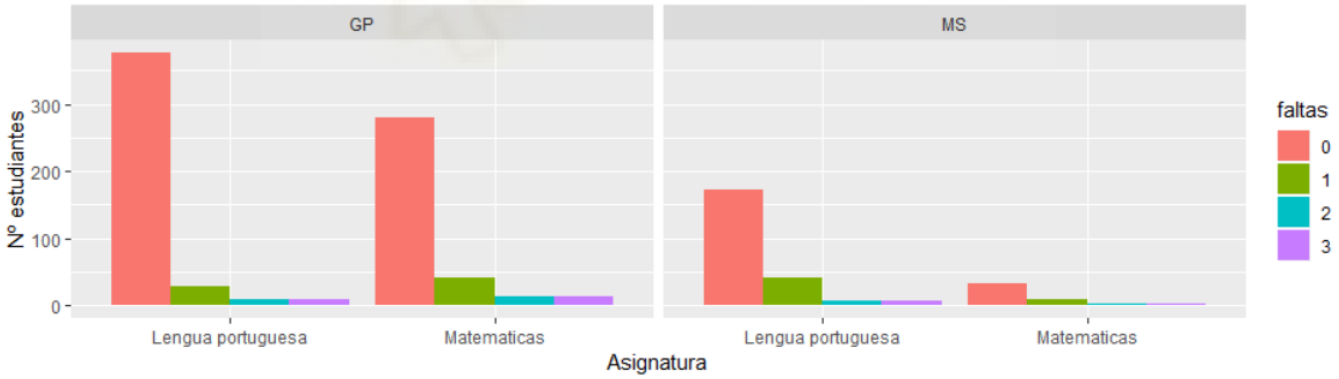

Fig. 32. Número de estudiantes en función de las faltas a clase por asignatura y escuela
Fuente: elaboración propia

En esta figura 32 se aprecia que el número de faltas de asistencia a clase que predomina entre los alumnos de cada una de las escuelas, tanto en lengua portuguesa como en matemáticas, es cero.

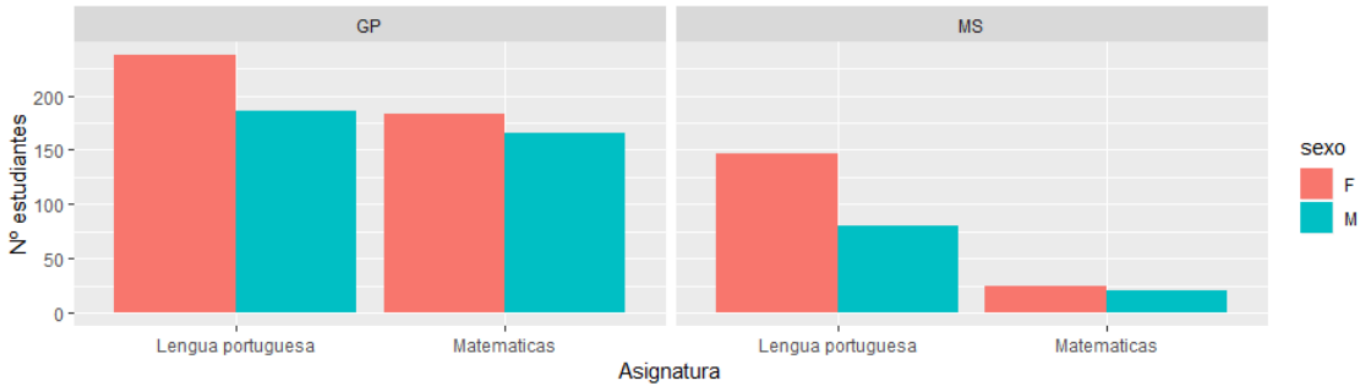

Fig. 33. Número de estudiantes en función del sexo por asignatura y escuela
Fuente: elaboración propia

Según el gráfico anterior (Fig. 33), se concluye que tanto en la escuela Gabriel Pereira (GP) como en la escuela Mousinho da Silveira (MS) para ambas asignaturas es mayor el número de estudiantes de sexo femenino.

En los siguientes gráficos, se representa la relación de las diferentes variables categóricas con la variable de estudio *calificación final* (*G3*). La comprobación del cumplimiento de la hipótesis nula se llevará a cabo para un intervalo de confianza a un nivel de significación del 95%.

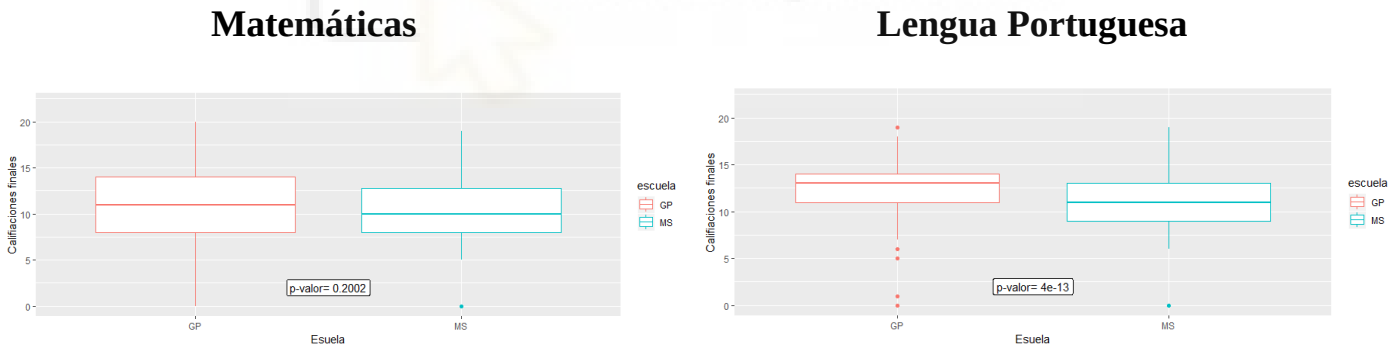

Fig. 34. Calificaciones finales en función de la escuela
Fuente: elaboración propia

En cuanto a la figura anterior (Fig. 34) se puede concluir que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales para ambas escuelas. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que las calificaciones finales en términos de medianas no son iguales para ambas escuelas.

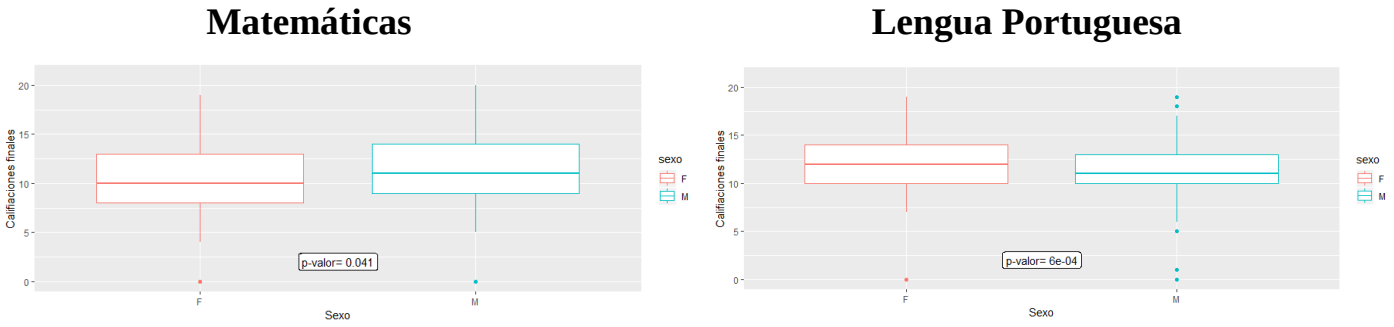

Fig. 35. Calificaciones finales en función del sexo del estudiante
Fuente: elaboración propia

Según la figura 35, observamos que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, se rechaza la hipótesis nula (p-valor < 0.05), por tanto, las calificaciones finales en términos de medianas no son iguales en ambos sexos.

Fig. 36. Calificaciones finales en función del tamaño de la familia
Fuente: elaboración propia

En la figura 36 se refleja que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del tamaño de la familia del estudiante.

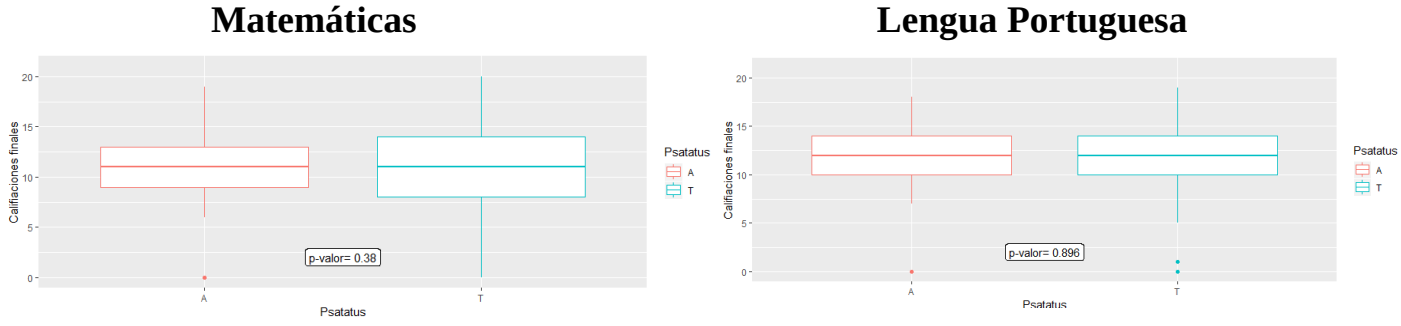

Fig. 37. Calificaciones finales en función del estado de convivencia de los padres
Fuente: elaboración propia

En la figura 37 se aprecia que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del estado de convivencia de los padres.

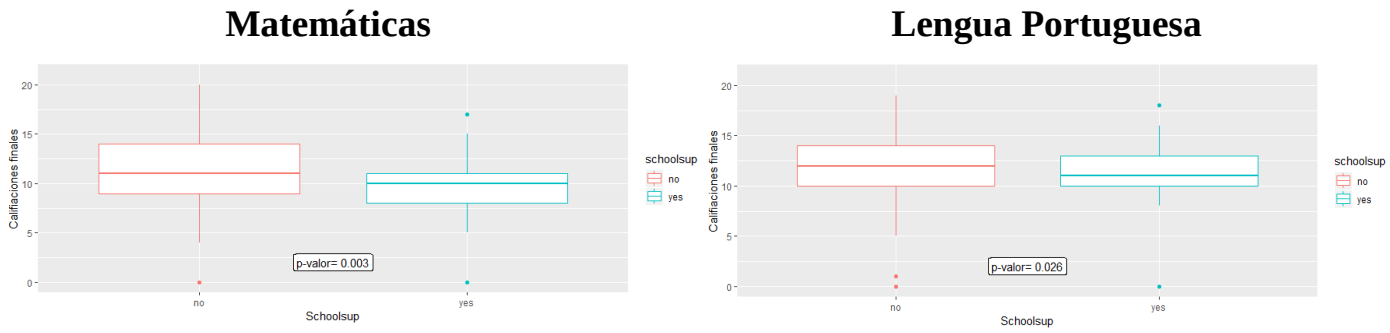

Fig. 38. Calificaciones finales en función del apoyo educativo extra
Fuente: elaboración propia

Según la figura 38, observamos que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, se rechaza la hipótesis nula (p-valor < 0.05), por tanto, las calificaciones finales en términos de medianas no son iguales si se tiene en cuenta el apoyo educativo extra.

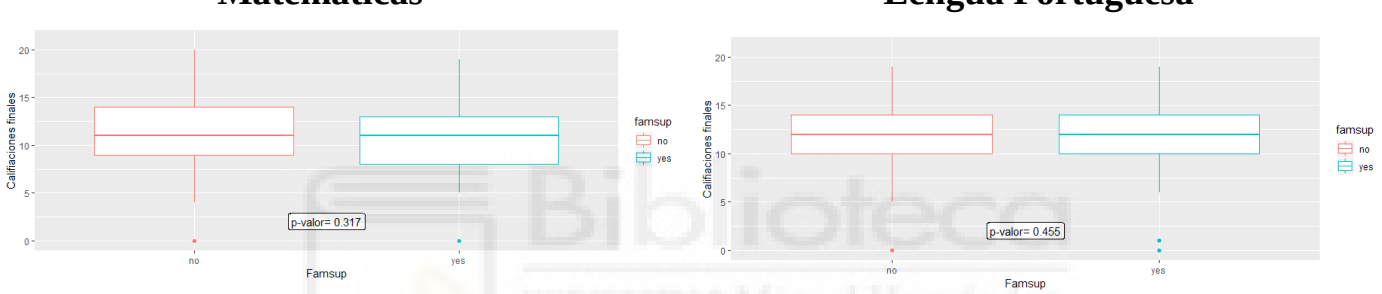

Fig. 39. Calificaciones finales en función del apoyo educativo familiar
Fuente: elaboración propia

En la figura 39 se refleja que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del apoyo educativo familiar.

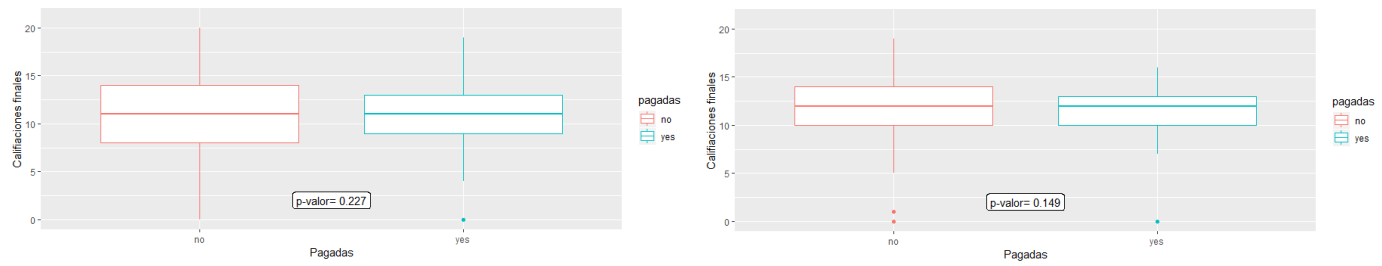

Fig. 40. Calificaciones finales en función de las clases extra
Fuente: elaboración propia

Según la figura 40, se concluye que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente de las clases extra.

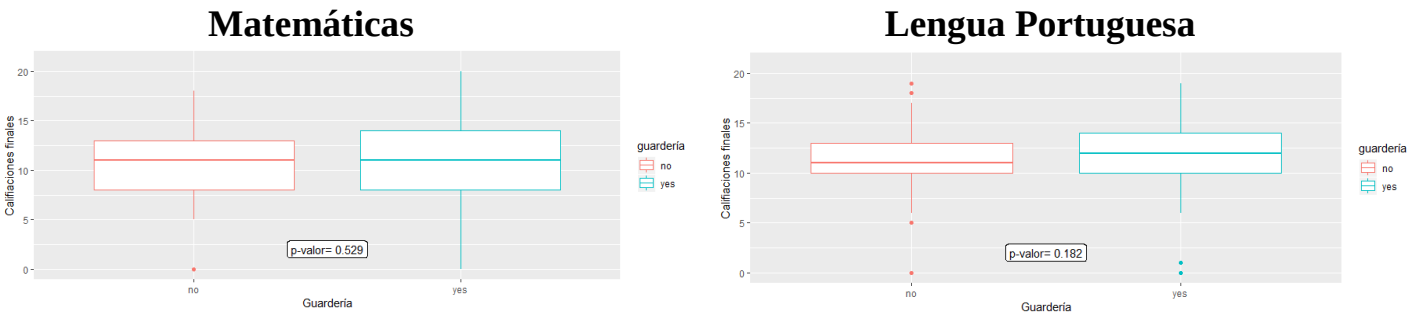

Fig. 41. Calificaciones finales en función de la asistencia a escuela infantil
Fuente: elaboración propia

En la figura 41 se observa que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente de si los estudiantes han asistido a escuelas infantiles o no.

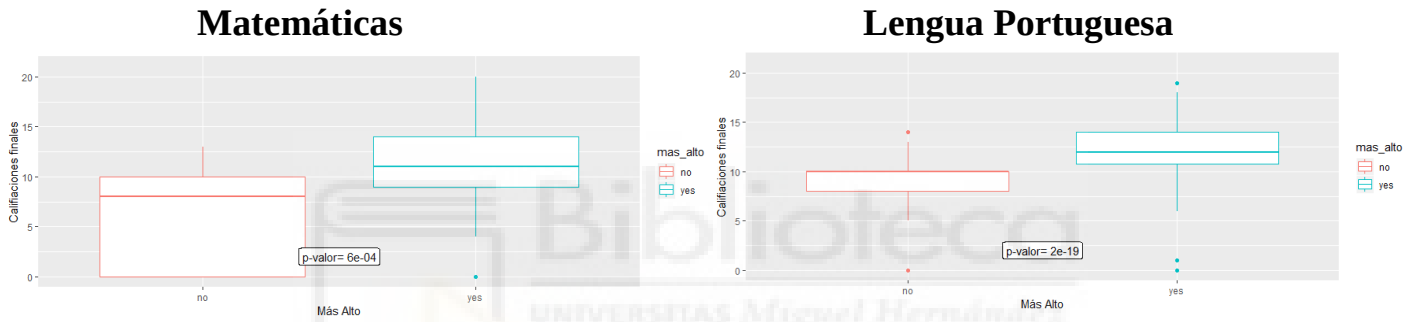

Fig. 42. Calificaciones finales en función del deseo de obtener educación superior
Fuente: elaboración propia

En la figura 42 se aprecia que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, se rechaza la hipótesis nula (p-valor < 0.05), por tanto, las calificaciones finales en términos de medianas no son iguales para los estudiantes que desean obtener educación superior y para los que no.

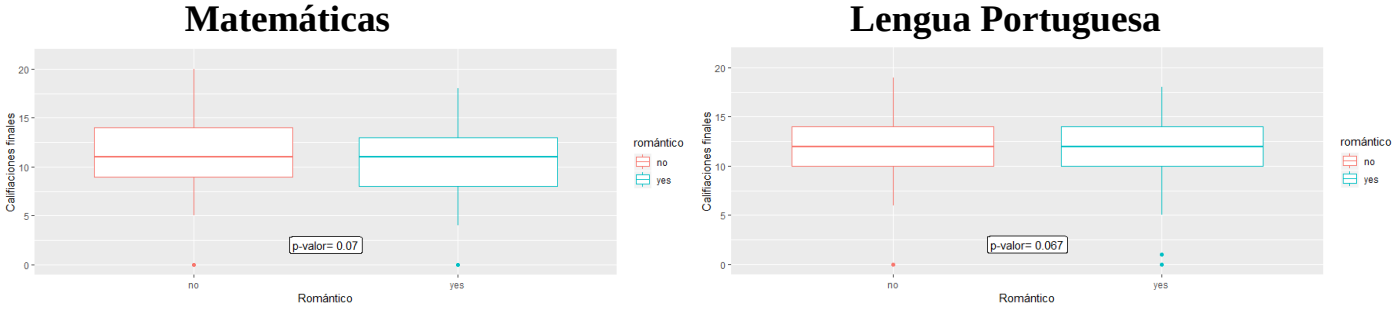

Fig. 43. Calificaciones finales en función de si tienen una relación sentimental
Fuente: elaboración propia

Según la figura 43, se concluye que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente de si los estudiantes tienen una relación sentimental o no.

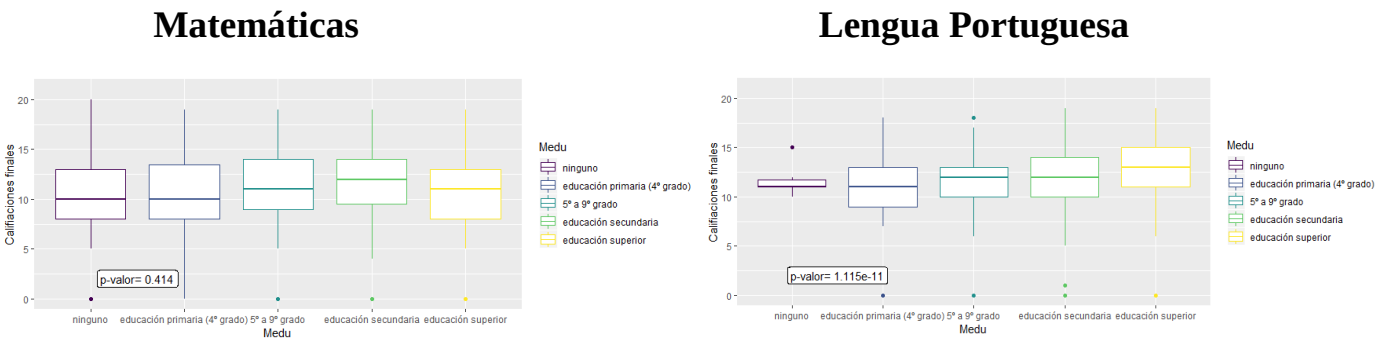

Fig. 44. Calificaciones finales en función de la educación de la madre
Fuente: elaboración propia

En cuanto a la figura anterior (Fig. 44) se puede concluir que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente de la educación de la madre. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de lo que se deduce que los estudiantes de lengua portuguesa obtienen mayores calificaciones cuando la madre ha cursado hasta educación superior.

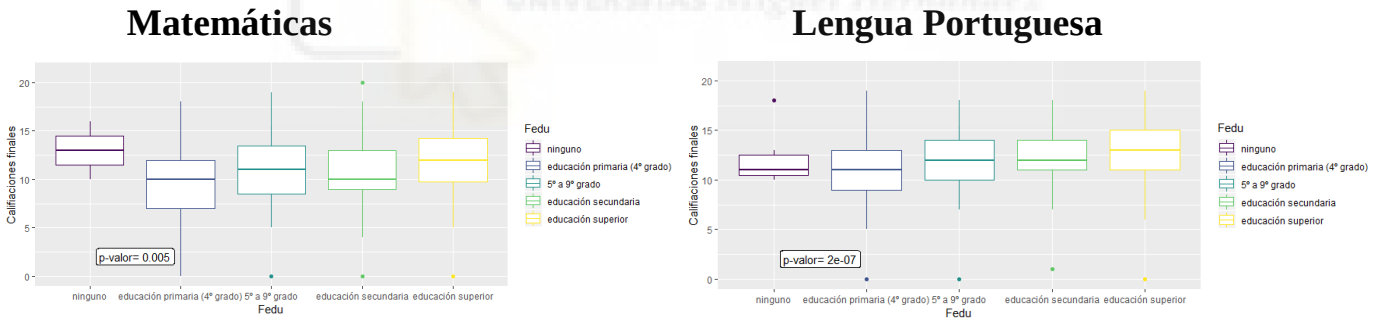

Fig. 45. Calificaciones finales en función de la educación del padre
Fuente: elaboración propia

Según la figura 45, observamos que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, se rechaza la hipótesis nula (p-valor < 0.05), por tanto, algunas de las calificaciones finales en términos de medianas son distintas, de forma que las calificaciones de los estudiantes son más altas cuando el padre ha cursado hasta educación superior.

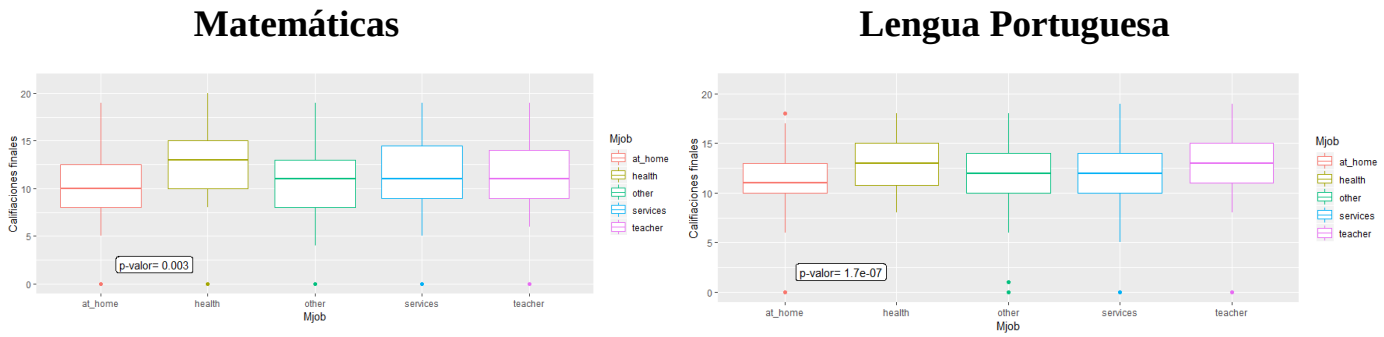

Fig. 46. Calificaciones finales en función del trabajo de la madre
Fuente: elaboración propia

En la figura 46 se aprecia que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, se rechaza la hipótesis nula (p-valor < 0.05), por tanto, algunas de las calificaciones finales en términos de medianas son distintas, deduciéndose de esta forma que las calificaciones de los estudiantes son más altas en la asignatura de matemáticas cuando el trabajo de la madre está relacionado con la salud y en el caso de lengua portuguesa cuando el trabajo de la madre está relacionado con la enseñanza. Siendo más bajas las calificaciones cuando la actividad de la madre es la de ama de casa.

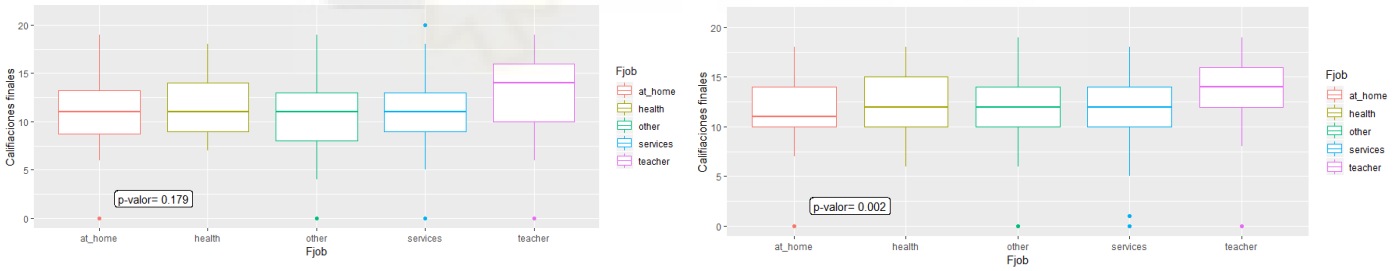

Fig. 47. Calificaciones finales en función del trabajo del padre
Fuente: elaboración propia

En cuanto a la figura anterior (Fig. 47) se puede concluir que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del trabajo del padre. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de lo que se deduce que los estudiantes de lengua portuguesa obtienen mayores calificaciones cuando la actividad laboral del padre está relacionada con la enseñanza.

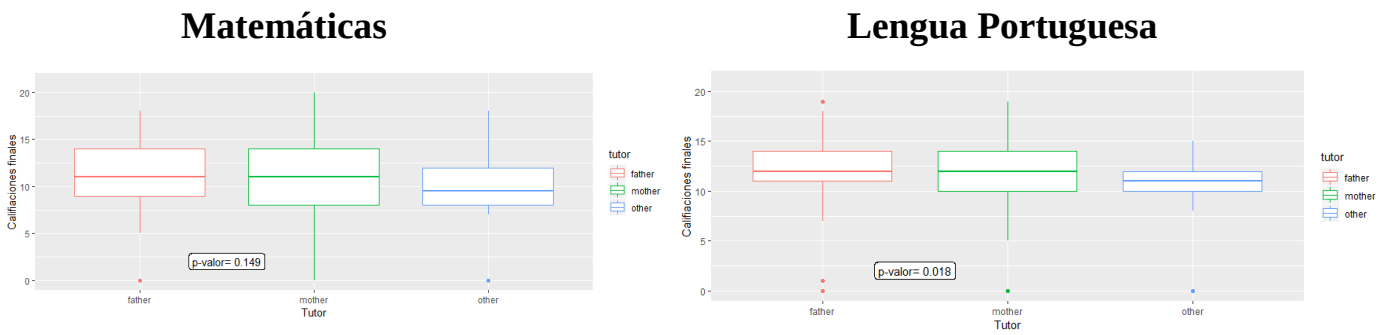

Fig. 48. Calificaciones finales en función del tutor
Fuente: elaboración propia

Según la figura 48, se puede concluir que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente de quién sea el tutor legal del estudiante. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de forma que los estudiantes de lengua portuguesa obtienen menores calificaciones cuando el tutor legal no es ni el padre ni la madre.

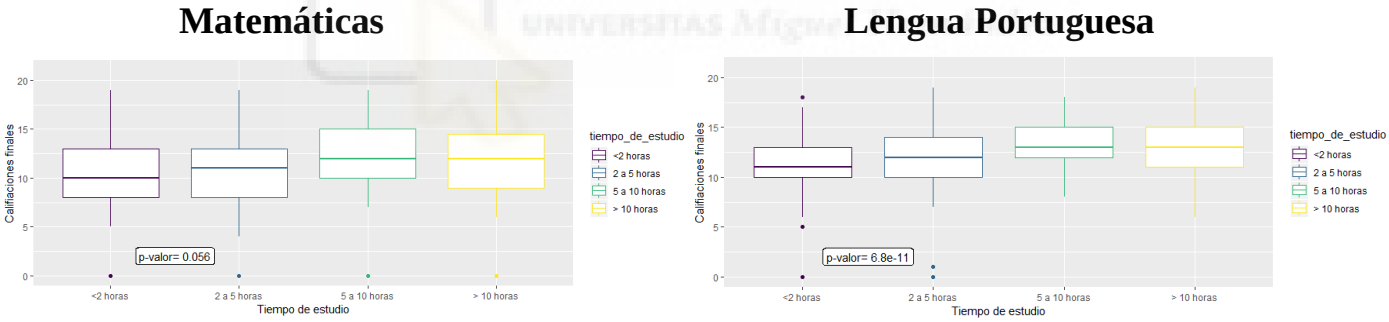

Fig. 49. Calificaciones finales en función del tiempo de estudio
Fuente: elaboración propia

En cuanto a la figura anterior (Fig. 49) se puede concluir que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del tiempo dedicado al estudio. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de lo que se deduce que los estudiantes de lengua portuguesa obtienen menores calificaciones cuando el tiempo dedicado al estudio es menos de dos horas.

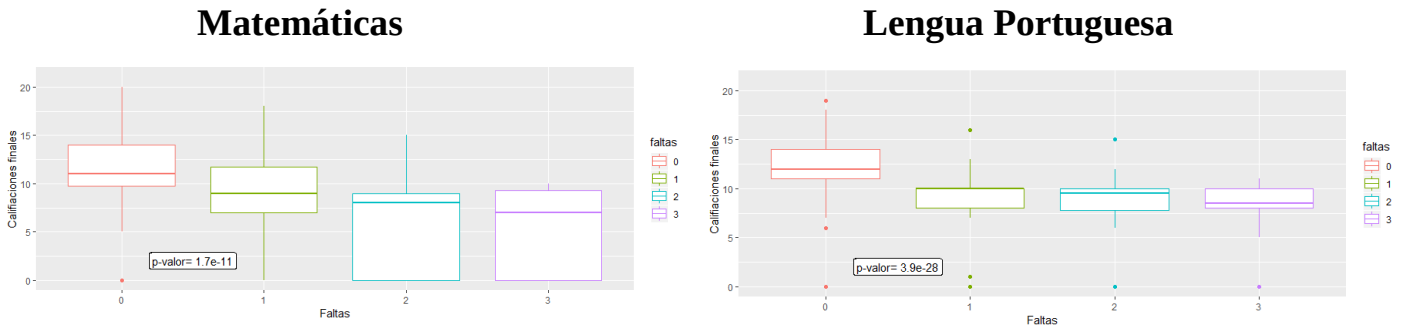

Fig. 50. Calificaciones finales en función del número de faltas a clase
Fuente: elaboración propia

En la figura 50 se aprecia que, tanto para la asignatura de matemáticas como para la de lengua portuguesa, se rechaza la hipótesis nula (p-valor < 0.05), por tanto, algunas de las calificaciones finales en términos de medianas son distintas, deduciéndose de esta forma que las calificaciones de los estudiantes en ambas asignaturas disminuyen conforme aumenta el número de faltas de asistencia a clase.

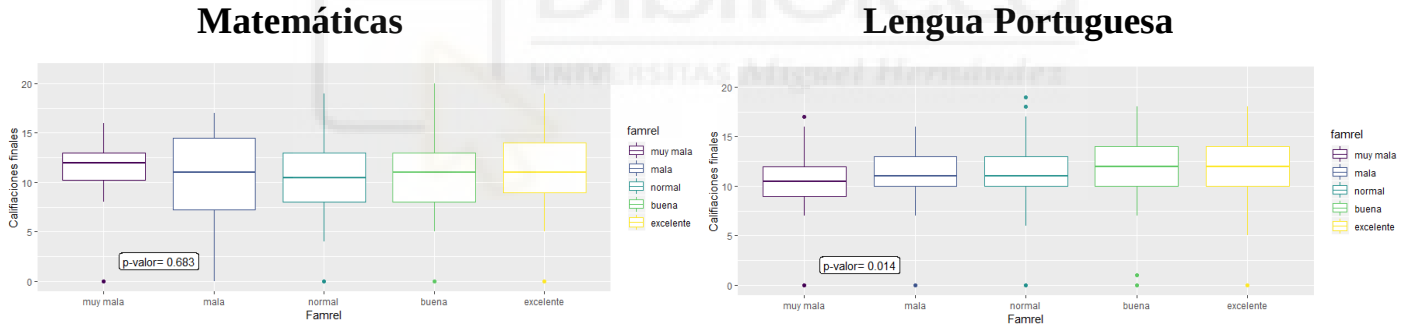

Fig. 51. Calificaciones finales en función de la calidad de las relaciones familiares
Fuente: elaboración propia

En cuanto a la figura anterior (Fig. 51) se puede concluir que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente de la calidad de las relaciones familiares. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de forma que los estudiantes de lengua portuguesa obtienen menores calificaciones cuando la calidad de las relaciones familiares es muy mala.

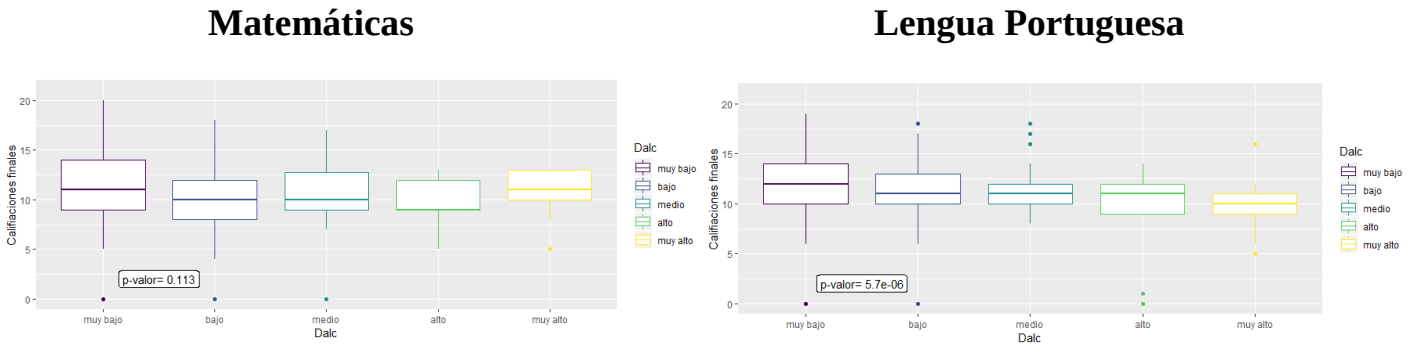

Fig. 52. Calificaciones finales en función del consumo de alcohol en días hábiles
Fuente: elaboración propia

En la figura 52 se observa que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del consumo de alcohol de los estudiantes en días hábiles. En el caso de lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de lo que se deduce que los estudiantes de lengua portuguesa obtienen menores calificaciones cuando el consumo del alcohol en días hábiles es muy alto.

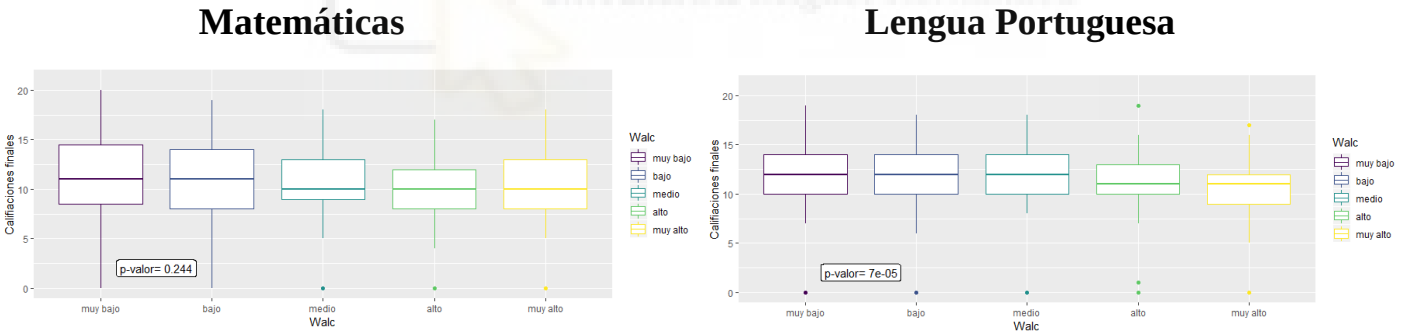

Fig. 53. Calificaciones finales en función del consumo de alcohol los fines de semana
Fuente: elaboración propia

En la figura anterior (Fig.53) se aprecia que, en la asignatura de matemáticas no se rechaza la hipótesis nula (p-valor > 0.05), por tanto, las calificaciones finales en términos de medianas son iguales independientemente del consumo de alcohol de los estudiantes los fines de semana. Sin embargo, en lengua portuguesa ocurre lo contrario (p-valor < 0.05), por lo que algunas de las calificaciones finales en términos de medianas son distintas, de forma que los estudiantes de lengua portuguesa obtienen menores calificaciones cuando el consumo del alcohol los fines de semana es muy alto.

A continuación, se realiza un análisis para observar la relación de las variables numéricas entre sí en ambas asignaturas:

d1 (Matemáticas)

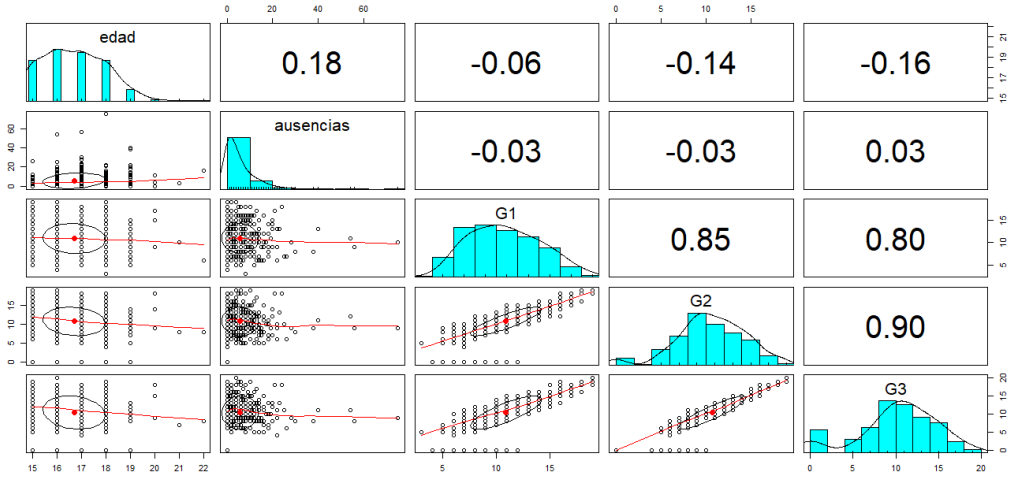

Fig. 54. Matriz de dispersión, histograma y correlación (matemáticas)
Fuente: elaboración propia

d2 (Lengua portuguesa)

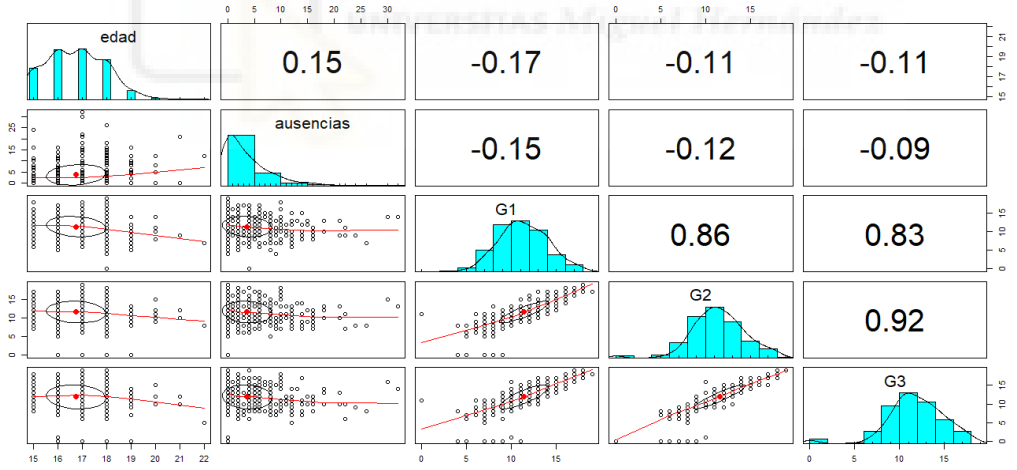

Fig. 55. Matriz de dispersión, histograma y correlación (lengua portuguesa)
Fuente: elaboración propia

Se observa (Fig. 54 y Fig. 55) que, tanto en ciencias como en letras, la correlación es alta entre las calificaciones (del primer, segundo y tercer período) motivo por el que se prescinde de G1 y G2 al realizar las discretizaciones. Sin embargo, la correlación entre la edad de los estudiantes de la asignatura de portugués y las calificaciones en los distintos periodos y la correlación entre el número de ausencias escolares y las calificaciones en los distintos períodos es cercana a cero, lo que indica que no existe apenas correlación entre estas variables.

Por tanto, se puede concluir que el atributo de destino G3 (*calificación final del año emitida en el tercer período*) tiene una fuerte correlación con los atributos G2 y G1 que corresponden a las calificaciones del primer y segundo período, tal y como se muestra en las figuras 54 y 55, lo que indica que el rendimiento de los estudiantes está muy influenciado por las evaluaciones anteriores. Siendo de esta forma más difícil predecir G3 sin G2 y G1, pero considerando dicha predicción mucho más útil.

5.2. DISCRETIZACIONES

DATASET 1: Matemáticas

Discretización del atributo *G3*

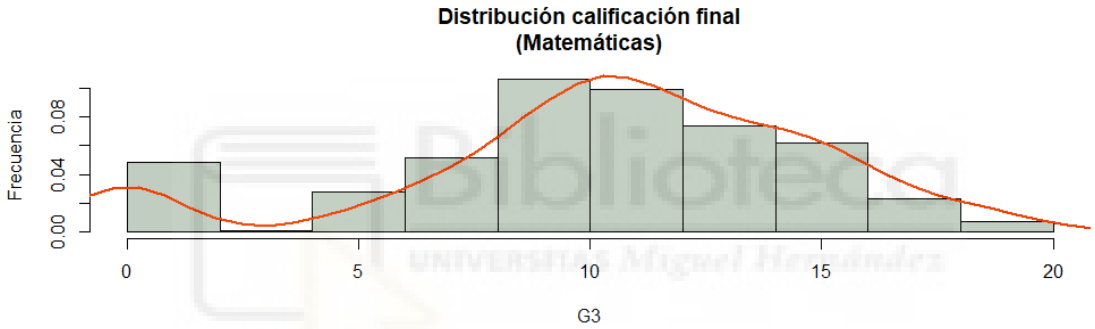

Fig. 56. Calificaciones matemáticas período final
Fuente: elaboración propia

En primer lugar (Tabla 4), se realiza una discretización de la variable en función del criterio que sigue el programa Rstudio. De ésta se obtienen tres categorías:

G3	Valor
[0,10)	*1- Necesidad de refuerzo intenso*
[10,12)	*2- Necesidad de mantenimiento*
[12,20]	*3- Mejora notable del rendimiento*

Tabla 4. Discretización en tres tramos del atributo calificación final (G3)
Fuente: elaboración propia

De la misma forma se realiza esta discretización en G2 y G1 (Tabla 5 y Tabla 6). Se obtienen discretizaciones parecidas por la correlación explicada anteriormente entre G1, G2 y G3 respectivamente.

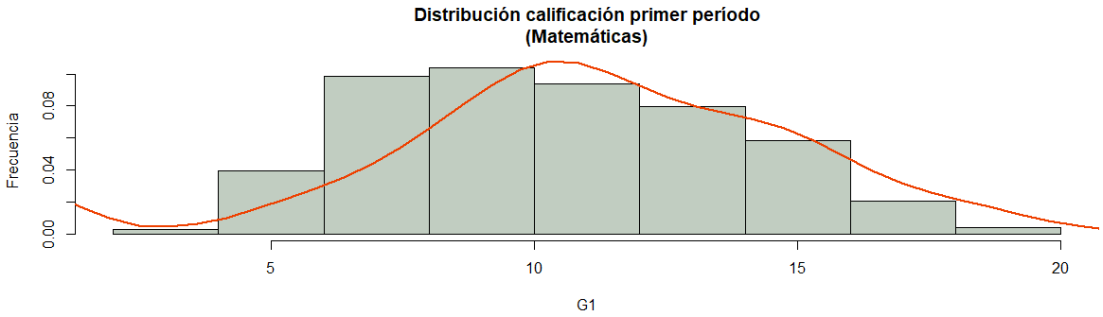

Fig. 57. Calificaciones matemáticas primer período
Fuente: elaboración propia

G1	Valor
[3,9)	*1- Necesidad de refuerzo intenso*
[9,12)	*2- Necesidad de mantenimiento*
[12,19]	*3- Mejora notable del rendimiento*

Tabla 5. Discretización en tres tramos del atributo calificación primer período (G1)
Fuente: elaboración propia

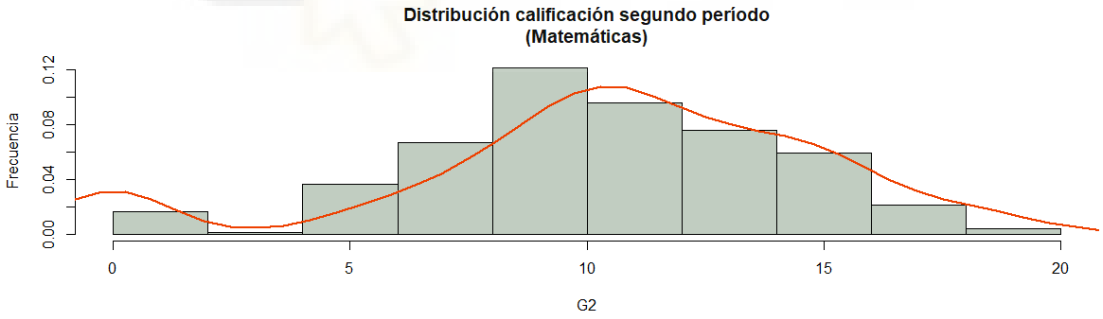

Fig. 58. Calificaciones matemáticas segundo período
Fuente: elaboración propia

G2	Valor
[0,9)	*1- Necesidad de refuerzo intenso*
[9,12)	*2- Necesidad de mantenimiento*
[12,19]	*3- Mejora notable del rendimiento*

Tabla 6. Discretización en tres tramos del atributo calificación segundo período (G2)
Fuente: elaboración propia

En segundo lugar, se realiza una discretización de la variable en dos categorías (Tabla 7) y se obtiene lo siguiente:

G3	Valor
[0,11)	*1- Suspenso*
[11,20]	*2- Aprobado*

Tabla 7. Discretización en dos tramos del atributo calificación final (G3)
Fuente: elaboración propia

Ocurre lo mismo que anteriormente. De la misma forma se realiza esta discretización en G2 y G1. Se obtienen discretizaciones parecidas por la correlación explicada con anterioridad entre G1, G2 y G3 respectivamente.

En tercer lugar, se realiza una discretización de la variable en cuatro categorías (Tabla 8) y se obtiene lo siguiente:

G3	Valor
[0,8)	*1- Menor rendimiento académico*
[8,11)	*2- Refuerzo para aprobar*
[11,14)	*3- Refuerzo para mantener*
[14,20]	*4- Posibilidad de mejora notable*

Tabla 8. Discretización en cuatro tramos del atributo calificación final (G3)
Fuente: elaboración propia

Ocurre lo mismo que anteriormente. De la misma forma se realiza esta discretización en G2 y G1. Se obtienen discretizaciones parecidas por la correlación explicada anteriormente entre G1, G2 y G3 respectivamente.

Discretización del atributo *ausencias*

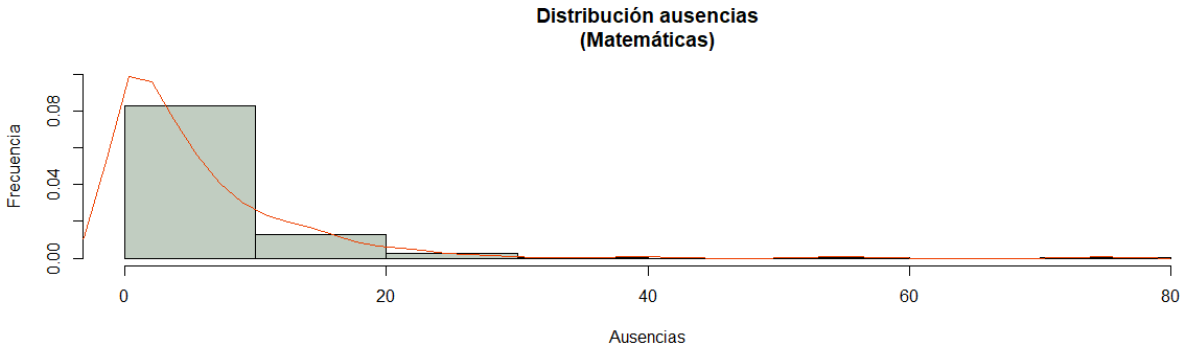

Fig. 59. Número de ausencias escolares en la asignatura de matemáticas
Fuente: elaboración propia

En primer lugar, se realiza una discretización de la variable en función del criterio que sigue el programa Rstudio (Tabla 9). De esta discretización se obtienen las tres categorías siguientes:

ausencias	Valor
[0,2)	*1- Puntuales*
[2,6)	*2- Moderadas*
[6,75]	*3- Frecuentes*

Tabla 9. Discretización en tres tramos del atributo ausencias
Fuente: elaboración propia

NOTA:

*llega hasta 75 porque, aunque la variable (número de ausencias escolares) esté definida de 0 a 93, el máximo de ausencias escolares es 75:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	4	5,71	8	75

En segundo lugar, se realiza una discretización de la variable en dos categorías (Tabla 10) y se obtiene lo siguiente:

ausencias	Valor
[0,4)	*1- Puntuales*
[4,75]	*2- Moderadas/Frecuentes*

Tabla 10. Discretización en dos tramos del atributo ausencias
Fuente: elaboración propia

Discretización del atributo *edad*

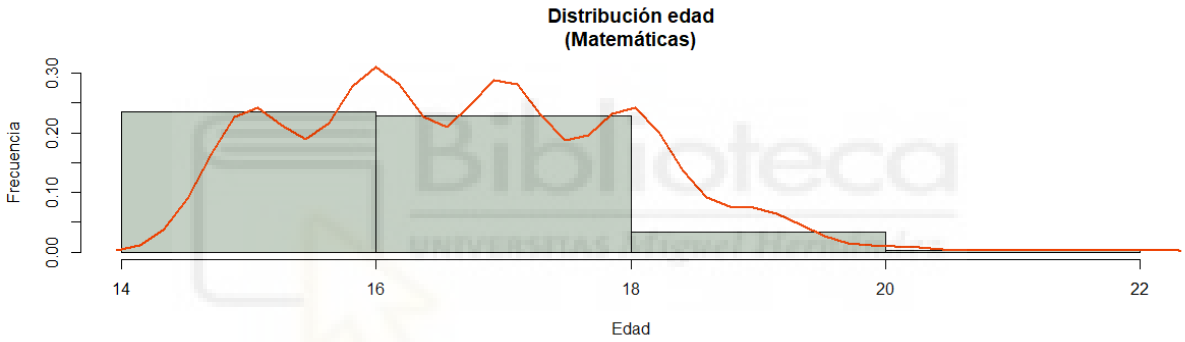

Fig. 60. Edades de los estudiantes de la asignatura de matemáticas
Fuente: elaboración propia

En primer lugar, se realiza una discretización de la variable edad (Tabla 11) en función del criterio que sigue el programa Rstudio. Se obtienen las tres categorías siguientes:

edad	Valor
[15,16)	*1- Preadolescentes*
[16,17)	*2- Adolescentes*
[17,22]	*3- Adolescentes adultos*

Tabla 11. Discretización en tres tramos del atributo edad
Fuente: elaboración propia

En segundo lugar, se realiza una discretización de la variable edad en dos categorías (Tabla 12) y se obtiene lo siguiente:

edad	Valor
[15,17)	*1- Adolescentes*
[17,22]	*2- Adolescentes adultos*

Tabla 12. Discretización en dos tramos del atributo edad
Fuente: elaboración propia

DATASET 2: Lengua portuguesa

Discretización del atributo *G3*

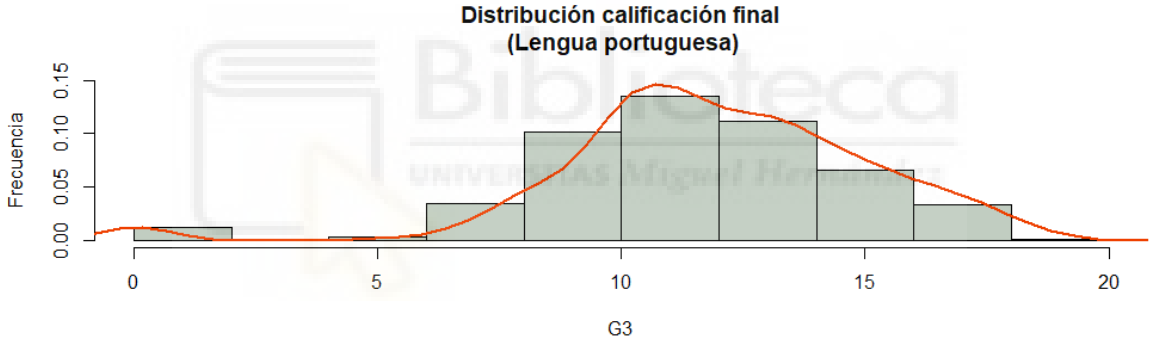

Fig. 61. Calificaciones lengua portuguesa período final
Fuente: elaboración propia

En primer lugar, se realiza una discretización de la variable en función del criterio que sigue el programa Rstudio (Tabla 13). Se obtienen las tres categorías siguientes:

G3	Valor
[0,11)	*1- Necesidad de refuerzo intenso*
[11,13)	*2- Necesidad de mantenimiento*
[13,19]	*3- Mejora notable del rendimiento*

Tabla 13. Discretización en tres tramos del atributo calificación final (G3)
Fuente: elaboración propia

De la misma forma se realiza esta discretización en G2 y G1. Se obtienen discretizaciones parecidas (Tabla 14 y Tabla 15) por la correlación explicada anteriormente entre G1, G2 y G3 respectivamente.

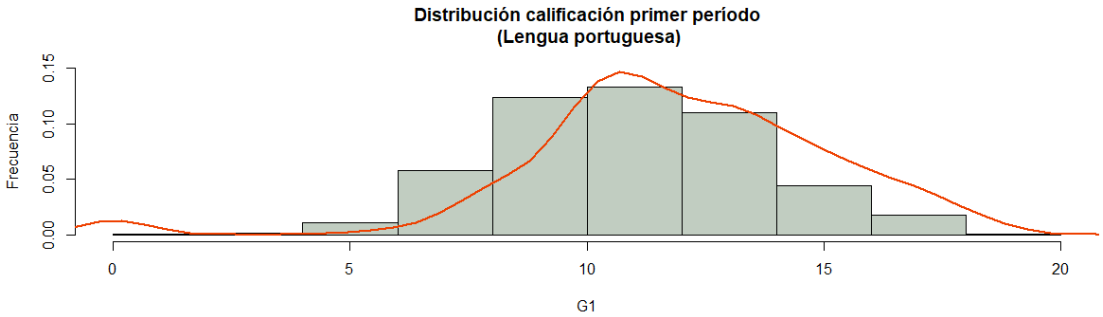

Fig. 62. Calificaciones lengua portuguesa primer período
Fuente: elaboración propia

G1	Valor
[0,10)	*1- Necesidad de refuerzo intenso*
[10,13)	*2- Necesidad de mantenimiento*
[13,19]	*3- Mejora notable del rendimiento*

Tabla 14. Discretización en tres tramos del atributo calificación primer período (G1)
Fuente: elaboración propia

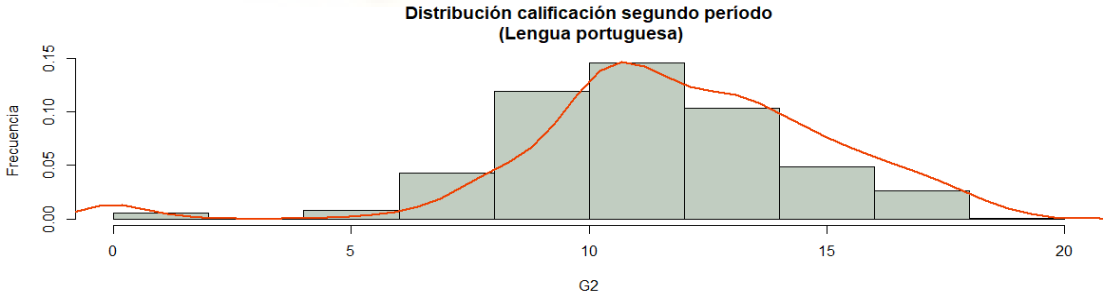

Fig. 63. Calificaciones lengua portuguesa segundo período
Fuente: elaboración propia

G2	Valor
[0,10)	*1- Necesidad de refuerzo intenso*
[10,13)	*2- Necesidad de mantenimiento*
[13,19]	*3- Mejora notable del rendimiento*

Tabla 15. Discretización en tres tramos del atributo calificación segundo período (G2)
Fuente: elaboración propia

En segundo lugar, se realiza una discretización de la variable en dos categorías (Tabla 16) y se obtiene lo siguiente:

G3	Valor
[0,12)	*1- Suspenso*
[12,19]	*2- Aprobado*

Tabla 16. Discretización en dos tramos del atributo calificación final (G3)
Fuente: elaboración propia

Ocurre lo mismo que anteriormente. De la misma forma se realiza esta discretización en G2 y G1. Se obtienen discretizaciones parecidas por la correlación explicada anteriormente entre G1, G2 y G3 respectivamente.

En tercer lugar, se realiza una discretización de la variable en cuatro categorías (Tabla 17) y se obtiene lo siguiente:

G3	Valor
[0,10)	*1- Menor rendimiento académico*
[10,12)	*2- Refuerzo para aprobar*
[12,14)	*3- Refuerzo para mantener*
[14,19]	*4- Posibilidad de mejora notable*

Tabla 17. Discretización en cuatro tramos del atributo calificación final (G3)
Fuente: elaboración propia

Ocurre lo mismo que anteriormente. De la misma forma se realiza esta discretización en G2 y G1. Se obtienen discretizaciones parecidas por la correlación explicada anteriormente entre G1, G2 y G3 respectivamente.

Discretización del atributo *ausencias*

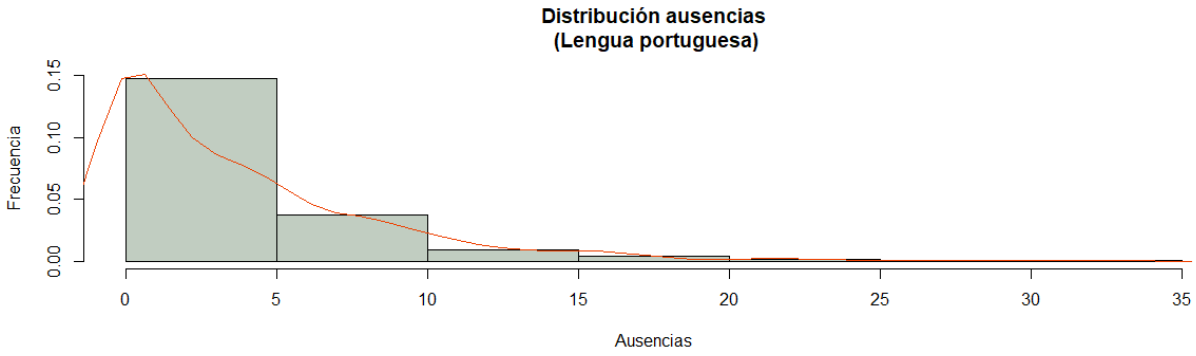

Fig. 64. Número de ausencias escolares en la asignatura de lengua portuguesa
Fuente: elaboración propia

En este caso, se realiza una discretización de la variable en dos categorías (Tabla 18) y se obtiene lo siguiente:

ausencias	Valor
[0,2)	*1- Puntuales*
[2,32]	*2- Moderadas/Frecuentes*

Tabla 18. Discretización en dos tramos del atributo ausencias
Fuente: elaboración propia

NOTA:

*llega hasta 32 porque, aunque la variable (número de ausencias escolares) esté definida de 0 a 93, el máximo de ausencias escolares es 32:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	2	3,66	6	32

Discretización del atributo *edad*

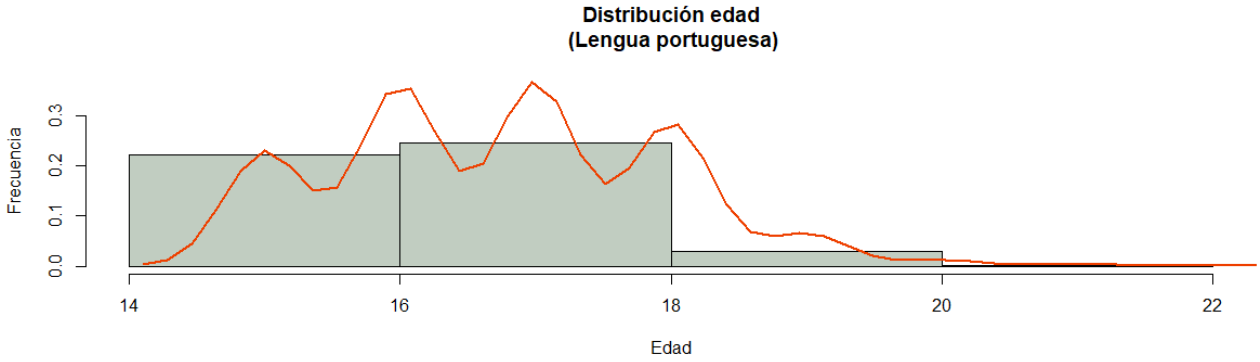

Fig. 65. Edades de los estudiantes de la asignatura de lengua portuguesa
Fuente: elaboración propia

En primer lugar, se realiza una discretización de la variable edad (Tabla 19) en función del criterio que sigue el programa Rstudio. De esta discretización se obtienen las siguientes tres categorías:

edad	Valor
[15,16)	*1- Preadolescentes*
[16,17)	*2- Adolescentes*
[17,22]	*3- Adolescentes adultos*

Tabla 19. Discretización en tres tramos del atributo edad
Fuente: elaboración propia

En segundo lugar, se realiza una discretización de la variable en dos categorías (Tabla 20) y se obtiene lo siguiente:

edad	Valor
[15,17)	*1- Adolescentes*
[17,22]	*2- Adolescentes adultos*

Tabla 20. Discretización en dos tramos del atributo edad
Fuente: elaboración propia

5.3. RANKINGS DE VARIABLES

El método de selección de atributos consiste en encontrar el grado de influencia de los antecedentes o combinaciones de éstos sobre la variable consecuente. Es una técnica descriptiva que se utiliza como paso previo en tareas de clasificación y de regresión.

En este caso, la selección de atributos para la realización del modelo predictivo se ha llevado a cabo siguiendo el criterio de visualización de los gráficos presentados a continuación.

	[0,10) [10,12) [12,20]	
	DISC.1	
	attributes	importance
1	faltas	0,097432467
2	mas_alto	0,057708888
3	schoolsup	0,047927652
4	Medu	0,017483291
5	edad_2	0,016394333
6	salidas	0,016016166
7	Fedu	0,012338411
8	Dalc	0,010795888
9	tiempo_libre	0,010225594
10	sexo	0,009837398
11	Mjob	0,009814902
12	pagadas	0,009758904
13	dirección	0,009699262
14	tiempo_de_viaje	0,008952199
15	romántico	0,008842128
16	tutor	0,008525819
17	Fjob	0,007691863
18	tiempo_de_estudio	0,00763327
19	famrel	0,006902759
20	razón	0,006815277
21	salud	0,006445354
22	Walc	0,006314469
23	internet	0,006197154
24	escuela	0,005518731
25	ausencias_2	0,004831916
26	famsize	0,004161795
27	Psatatus	0,004070101
28	famsup	0,003135783
29	guardería	0,000706667
30	actividades	0,000197137

	[0,11) [11,20]	
	DISC.2	
	attributes	importance
1	faltas	8,90E-02
2	mas_alto	6,09E-02
3	schoolsup	2,42E-02
4	Fedu	1,53E-02
5	edad_2	1,38E-02
6	Mjob	1,37E-02
7	Medu	1,34E-02
8	tiempo_de_viaje	1,13E-02
9	Dalc	1,11E-02
10	salidas	1,10E-02
11	dirección	1,03E-02
12	tiempo_libre	1,02E-02
13	escuela	9,91E-03
14	internet	9,81E-03
15	Walc	7,40E-03
16	tiempo_de_estudio	5,48E-03
17	Psatatus	4,58E-03
18	salud	4,11E-03
19	Fjob	3,93E-03
20	sexo	3,72E-03
21	famrel	3,63E-03
22	tutor	3,37E-03
23	razón	1,92E-03
24	famsize	1,42E-03
25	pagadas	1,35E-03
26	ausencias_2	1,15E-03
27	famsup	7,54E-04
28	guardería	7,15E-04
29	romántico	6,45E-05
30	actividades	3,12E-05

	[0,8) [8,11) [11,14) [14,20]	
	DISC.3	
	attributes	importance
1	faltas	0,121484723
2	mas_alto	0,095991129
3	schoolsup	0,052833951
4	Dalc	0,037894818
5	Medu	0,035341017
6	ausencias_2	0,028788041
7	pagadas	0,025676081
8	Walc	0,025625111
9	Fedu	0,025623837
10	Mjob	0,024041328
11	dirección	0,023460727
12	Fjob	0,020324738
13	internet	0,018074537
14	edad_2	0,017978783
15	tiempo_de_viaje	0,017974597
16	romántico	0,016563291
17	tiempo_de_estudio	0,015180418
18	escuela	0,014410164
19	salidas	0,013909806
20	salud	0,01241843
21	famrel	0,011965871
22	tiempo_libre	0,011919507
23	tutor	0,010305758
24	sexo	0,008682287
25	razón	0,008678878
26	Psatatus	0,00830089
27	guardería	0,006436647
28	famsize	0,005569771
29	actividades	0,003588154
30	famsup	0,001483338

Tabla 21. Rankings para las diferentes discretizaciones de la variable G3 en la asignatura de matemáticas. Fuente: elaboración propia

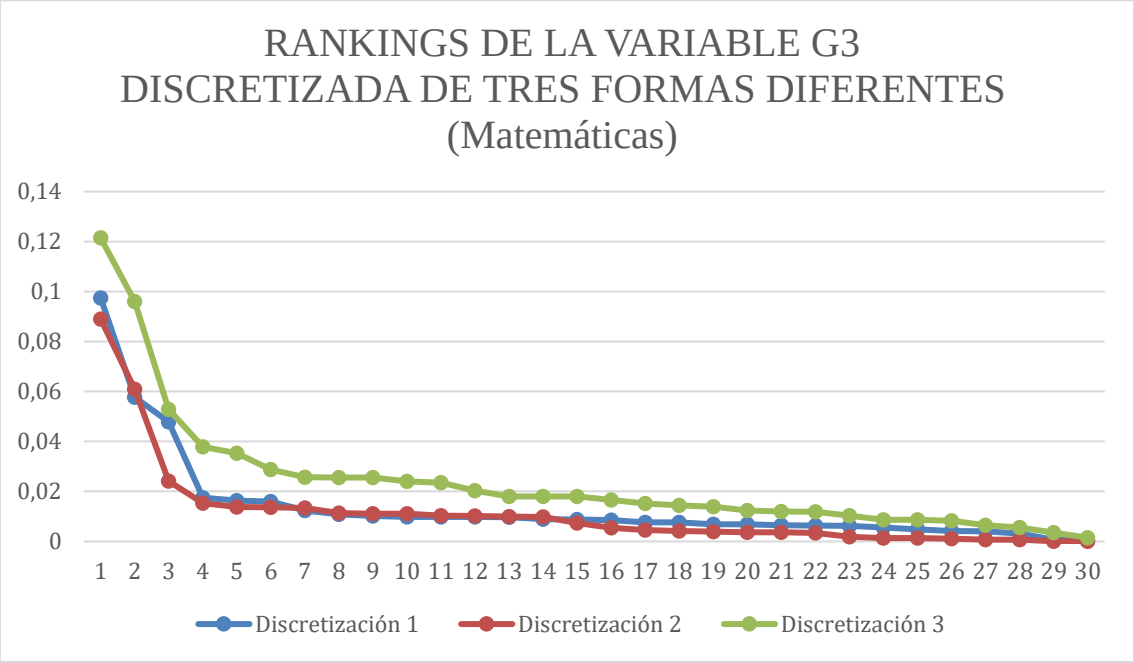

Fig. 66. Rankings según las diferentes discretizaciones de G3 en matemáticas
Fuente: elaboración propia

Según lo representado en la Fig. 66, lo más adecuado es escoger las tres primeras variables de las discretizaciones 1 y 2 y las cinco primeras variables de la discretización 3, siendo la discretización 1 la discretización que viene dada por el algoritmo utilizado.

En la Tabla 21 se puede observar el diferente posicionamiento de la variable edad discretizada en dos tramos ([15,17) [17,22]), en función de la discretización de la variable "Calificación final" (G3). Si factorizamos G3 en dos y tres tramos, la variable edad (discretizada en dos tramos) ocupa el puesto número 5 en los rankings. Sin embargo, al discretizarla en cuatro tramos, el puesto que ocupa esta variable es mucho más bajo, en este caso el 14.

Con la variable ausencias discretizada en dos tramos ([0,4) [4,75]) ocurre lo contrario. Discretizando la calificación final (variable objetivo) en dos y tres tramos los puestos que ocupa el atributo ausencias (discretizada en dos tramos) son muy bajos (25 y 26), lo cual no ocurre si discretizamos la variable G3 en cuatro tramos, quedando la factorización del atributo ausencias en un sexto puesto.

Además, se observa en la Tabla 21 que al discretizar la calificación final de la asignatura de Matemáticas (G3) en cualquiera de las tres discretizaciones realizadas, las variables que más relación tienen con esta variable objetivo son las siguientes:

- Número de faltas de clases pasadas (faltas).
- Deseo de la obtención de una educación superior (mas_alto).
- Apoyo educativo extra (schoolsup).

	[0,11) [11,13) [13,19]	
	DISC.1	
	attributes	importance
1	faltas	0,201996525
2	mas_alto	0,195684747
3	escuela	0,072904703
4	Medu	0,032765831
5	tiempo_de_estudio	0,031411923
6	Dalc	0,030032627
7	dirección	0,0264792
8	Fedu	0,022515826
9	internet	0,020945167
10	edad_2	0,020063606
11	Mjob	0,019917846
12	razón	0,019623025
13	ausencias_2	0,014819402
14	tutor	0,014432658
15	Fjob	0,01419309
16	schoolsup	0,013942338
17	Walc	0,013718358
18	sexo	0,012759999
19	tiempo_libre	0,011696522
20	tiempo_de_viaje	0,011649388
21	pagadas	0,010992392
22	salud	0,01064973
23	famrel	0,008424818
24	salidas	0,008052159
25	romántico	0,007638111
26	guardería	0,006355879
27	actividades	0,003359576
28	famsize	0,001726496
29	famsup	0,001193944
30	Psatatus	0,000593096

	[0,12) [12,19]	
	DISC.2	
	attributes	importance
1	mas_alto	1,73E-01
2	faltas	1,66E-01
3	escuela	5,57E-02
4	tiempo_de_estudio	3,02E-02
5	Medu	2,78E-02
6	Dalc	2,57E-02
7	dirección	2,42E-02
8	Mjob	2,18E-02
9	internet	2,17E-02
10	Fedu	1,97E-02
11	tiempo_de_viaje	1,77E-02
12	razón	1,46E-02
13	sexo	1,37E-02
14	Walc	1,36E-02
15	ausencias_2	1,29E-02
16	actividades	1,12E-02
17	Fjob	1,04E-02
18	tiempo_libre	7,96E-03
19	salidas	7,65E-03
20	tutor	7,02E-03
21	salud	5,09E-03
22	famrel	4,70E-03
23	guardería	4,52E-03
24	schoolsup	4,51E-03
25	romántico	1,59E-03
26	edad_2	1,39E-03
27	famsup	3,88E-04
28	pagadas	3,09E-04
29	Psatatus	1,26E-06
30	famsize	8,61E-08

	[0,10) [10,12) [12,14) [14,19]	
	DISC.3	
	attributes	importance
1	mas_alto	0,2055431
2	faltas	0,202360965
3	escuela	0,084177236
4	schoolsup	0,05430495
5	tiempo_de_estudio	0,03478018
6	Medu	0,034556737
7	pagadas	0,034270682
8	Dalc	0,030466164
9	Fedu	0,027726665
10	Mjob	0,026855745
11	dirección	0,026302556
12	internet	0,025419229
13	razón	0,022403878
14	tiempo_de_viaje	0,020488132
15	Walc	0,017350217
16	tiempo_libre	0,015424275
17	edad_2	0,015396075
18	sexo	0,01423922
19	salud	0,014108923
20	Fjob	0,013735097
21	famrel	0,013687449
22	ausencias_2	0,01324662
23	salidas	0,012072968
24	actividades	0,011267784
25	tutor	0,010733052
26	guardería	0,008758152
27	romántico	0,004985782
28	famsize	0,003363938
29	famsup	0,001999132
30	Psatatus	0,000261546

Tabla 22. Rankings para las diferentes discretizaciones de la variable G3 en la asignatura de lengua portuguesa. Fuente: elaboración propia

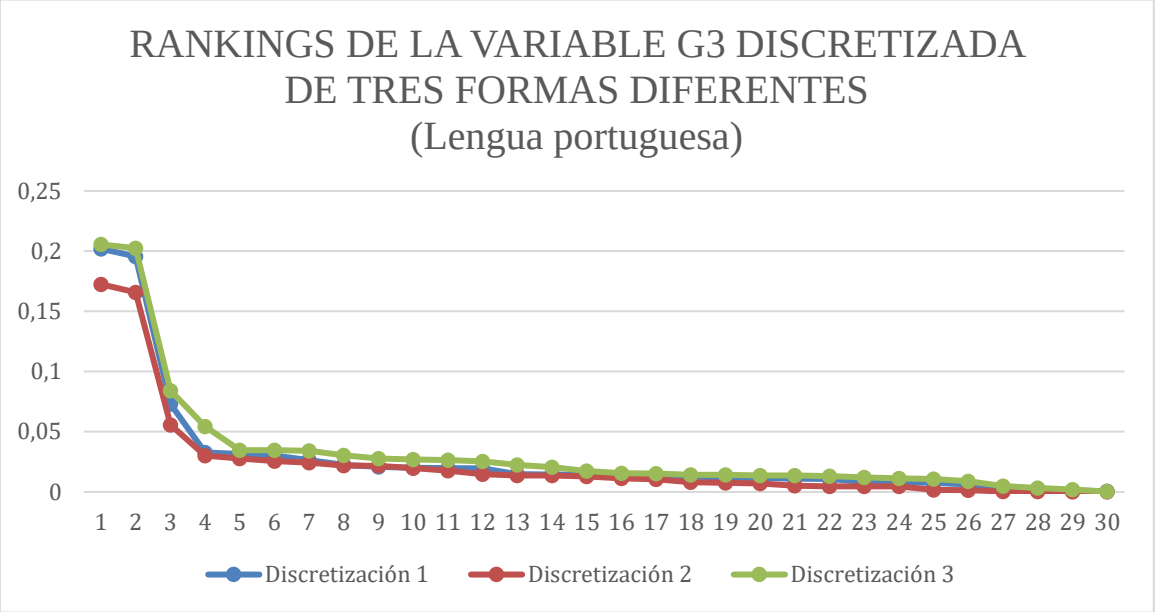

Fig. 67. Rankings según las diferentes discretizaciones de G3 en lengua portuguesa
Fuente: elaboración propia

Según lo expuesto en la Fig. 67, lo más conveniente es escoger las tres primeras variables de las discretizaciones 1 y 2 y las cuatro primeras variables de la discretización 3, siendo la discretización 1 la que viene dada por el algoritmo utilizado.

De los rankings de la Tabla 22 podemos destacar que la variable edad discretizada en dos tramos ([15,17) [17,22]) ocupa un lugar más alto si la variable "Calificación final" se discretiza en tres tramos.

Además, se conoce según la Tabla 22 que al discretizar la calificación final de la asignatura Lengua portuguesa (G3) en cualquiera de las tres discretizaciones realizadas, las variables que más relación tienen con esta variable objetivo son la siguientes:

- Número de faltas de clases pasadas (faltas).
- Deseo de la obtención de una educación superior (mas_alto).
- Escuela de estudiantes "Gabriel Pereira" o "Mousinho da Silveira" (escuela).

Podemos concluir que, tanto en la asignatura de Matemáticas (Ciencias) como en la asignatura de Lengua portuguesa (Letras), el número de faltas de clases pasadas (faltas) y el deseo de la obtención de una educación superior (mas_alto) son las variables que más afectan a las calificaciones finales en estas dos asignaturas.

5.4. MODELOS PREDICTIVOS

Los términos necesarios para abordar los modelos predictivos que se obtienen a continuación son los siguientes:

- ***Árbol de clasificación***: modelo predictivo que permite la clasificación de un conjunto de instancias siguiendo un conjunto de reglas situadas en los nodos de una estructura de datos de tipo árbol. Este modelo tiene una precisión en concreto.

- ***Algoritmo de clasificación (Rpart)***: "Recursive Partitioning and Regression Trees" es un algoritmo que divide de forma recursiva el conjunto de entrenamiento siguiendo una regla que involucra un atributo independiente, este proceso continúa con los conjuntos resultantes hasta que un criterio de terminación se cumple.

La regla de división elegida es aquella que sigue el criterio de reducir, en mayor medida, la heterogeneidad del atributo dependiente del conjunto original. Este criterio toma el nombre de impureza (impurity).

- ***Accuracy***: es el porcentaje del total de aciertos de nuestro modelo. Es importante señalar la diferencia entre "Accuracy" y "Precisión", dos factores importantes a considerar cuando se toman mediciones de datos. Ambos reflejan qué tan cerca está una medida de un valor real, pero la diferencia reside en que la exactitud (accuracy) refleja la cercanía de una medida a un valor conocido o aceptado, mientras que la precisión refleja qué tan reproducibles son las mediciones, incluso si están lejos del valor aceptado.

A continuación, se muestra un ejemplo donde se visualiza esta diferencia de conceptos:

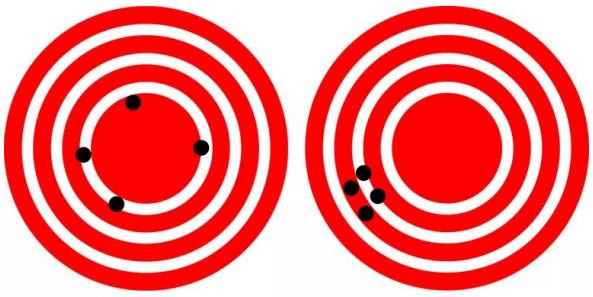

Fig. 68. Accuracy vs. Precisión
Fuente: recurso web número 17

En la figura 68, la imagen de la izquierda muestra un golpe de blanco con un alto grado de *exactitud (accuracy)*, pero de baja *precisión*. Mientras que la imagen de la derecha muestra el impacto del objetivo con alta *precisión*, pero baja *exactitud.*

- ***Matriz de confusión***: indica la forma en la que se distribuye el error. Siendo mejor siempre el modelo que distribuye su error de manera equitativa entre valores.

DATASET 1: Matemáticas

MODELO PARA LA DISCRETIZACIÓN EN DOS TRAMOS [0,11) [11,20] (DISC.2)

$$DISC.2 \sim faltas + mas\_alto + schoolsup$$

Fig. 69. Variables más influyentes usando DISC.2
Fuente: elaboración propia

Para la realización del modelo presentado a continuación, se tienen en cuenta las variables situadas en los primeros puestos del ranking obtenido al discretizar la variable "calificación final" en la asignatura de matemáticas en dos tramos. El árbol resultante se muestra en Fig. 70.

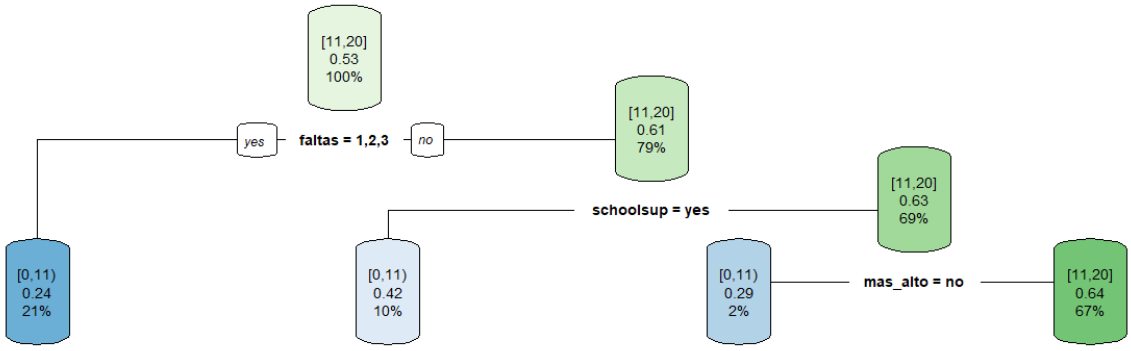

Fig. 70. Árbol de decisión obtenido con el método RPART
Fuente: elaboración propia

Una vez obtenido este modelo (Fig. 70), por la rama más a la izquierda, se puede concluir que el alumno que ha faltado alguna vez tiene un 76% de probabilidad de suspender (1-0.24). Sin embargo, por la rama más a la derecha, el alumno que no ha faltado nunca, que no recibe apoyo educativo extra (schoolsup = no) y que desea obtener educación superior (mas_alto = yes) tiene un 64% (0.64) de probabilidad de aprobar.

	[0,11)	[11,20]
[0,11)	91	95
[11,20]	39	170

Tabla 23. Matriz de confusión generada para el árbol de la figura XX
Fuente: elaboración propia

Accuracy: 0.6607595 (66%)

Esta matriz de confusión (Tabla 23), asociada a una discretización en dos tramos, tiene una precisión media aceptable. Parece que el modelo resulta bueno ya que, a pesar de clasificar los casos positivos con una probabilidad del 48,92%, los casos negativos los clasifica con una probabilidad del 81,34%. Esto puede derivar de lo observado en la Tabla 23., los falsos aprobados (errores de tipo II), es decir, predicciones que indican que los estudiantes van a aprobar [11,20] aunque realmente su calificación final será suspensa [0,11).

MODELO PARA LA DISCRETIZACIÓN EN TRES TRAMOS
[0,10) [10,12) [12,20]

DISC.1 ~ faltas + mas_alto + schoolsup

Fig. 71. Variables más influyentes usando DISC.1
Fuente: elaboración propia

En la obtención del modelo presentado a continuación, se tienen en cuenta las variables situadas en los primeros puestos del ranking una vez discretizada la variable "calificación final" en la asignatura de matemáticas en tres tramos. El árbol resultante se muestra en Fig. 72.

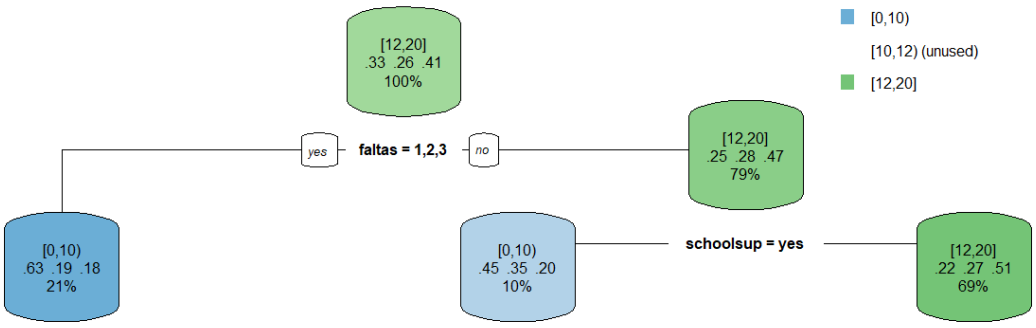

Fig. 72. Árbol de decisión obtenido con el método RPART
Fuente: elaboración propia

Como resultado de este modelo (Fig.72), por la rama más a la izquierda, se puede concluir que el alumno que ha faltado alguna vez tiene un 63% de probabilidad de suspender (0.63). A diferencia del alumno (por la rama más a la derecha) que no ha faltado nunca y que no recibe apoyo educativo extra (schoolsup = no) el cual tiene un 51% (0.51) de probabilidad de obtener una buena calificación final.

	[0,10)	[10,12)	[12,20]
[0,10)	70	0	60
[10,12)	30	0	73
[12,20]	23	0	139

Tabla 24. Matriz de confusión generada para el árbol de la figura XX
Fuente: elaboración propia

Accuracy: 0.5291139 (53%)

En la matriz de confusión (Tabla 24) se aprecia que, pese a discretizar previamente en tres tramos, el modelo no clasifica ninguna instancia en el tramo central [10,12). Así pues, la matriz de confusión pone de manifiesto que esta discretización no es conveniente pese a tener una precisión media aceptable.

MODELO PARA LA DISCRETIZACIÓN EN CUATRO TRAMOS
[0,8) [8,11) [11,14) [14,20]

$$DISC.3 \sim faltas + mas\_alto + schoolsup + Dalc + Medu$$

Fig. 73. Variables más influyentes usando DISC.3
Fuente: elaboración propia

Para la ejecución del modelo presentado a continuación, se tienen en cuenta las variables situadas en los primeros puestos del ranking obtenido al discretizar la variable "calificación final" en la asignatura de matemáticas en cuatro tramos. El árbol resultante se muestra en Fig. 73.

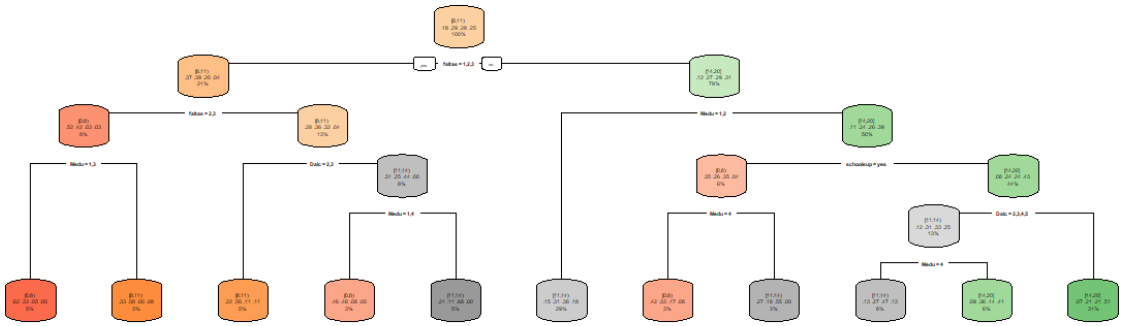

Fig. 73. Árbol de decisión obtenido con el método RPART
Fuente: elaboración propia

Al realizar este modelo (Fig.73), dónde rojo [0,8), naranja [8,11), gris [11,14) y verde [14,20], por la rama más a la izquierda, se concluye que el alumno que ha faltado alguna vez (2 veces o más) y que el nivel de formación de la madre es educación primaria o educación secundaria, tiene una probabilidad de suspender del 62% (0.62). En cambio, por la rama más a la derecha, el alumno cuya formación de la madre es nula, educación secundaria o educación superior, que además no recibe apoyo educativo extra (schoolsup = no) y que tiene un consumo de alcohol en días hábiles muy bajo, tiene un 51% (0.51) de probabilidad de obtener una muy buena calificación final.

	[0,8)	[8,11)	[11,14)	[14,20]
[0,8)	24	8	28	10
[8,11)	17	17	48	34
[11,14)	4	2	75	28
[14,20]	1	3	25	71

Tabla 25. Matriz de confusión generada para el árbol de la figura XX
Fuente: elaboración propia

Accuracy: 0.4734177 (47%)

Esta matriz de confusión (Tabla 25), asociada a una discretización en cuatro tramos, también tiene una precisión media aceptable (teniendo en cuenta que ahora hay 4 posibles segmentos de nota) y muestra un reparto más equitativo de los errores de clasificación.

DATASET 2: Lengua portuguesa

MODELO PARA LA DISCRETIZACIÓN EN DOS TRAMOS
[0,12) [12,19]

$$DISC.2 \sim mas\_alto + faltas + escuela$$

Fig. 74. Variables más influyentes usando DISC.2
Fuente: elaboración propia

Para la realización del modelo presentado a continuación, se tienen en cuenta las variables situadas en los primeros puestos del ranking obtenido al discretizar la variable "calificación final" en la asignatura de lengua portuguesa en dos tramos. El árbol resultante se muestra en Fig. 75.

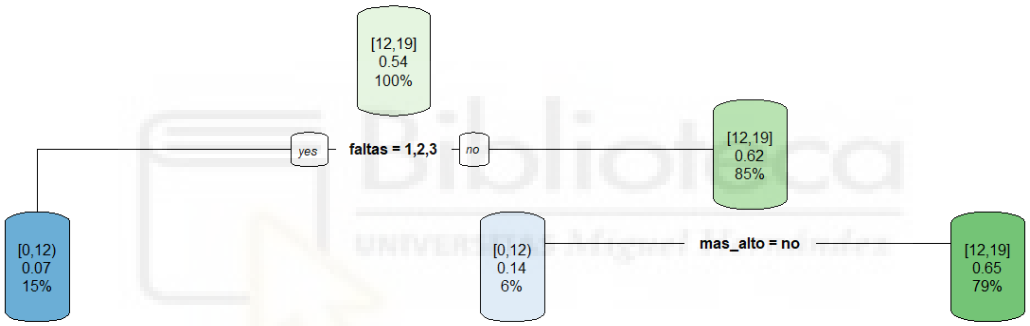

Fig. 75. Árbol de decisión obtenido con el método RPART
Fuente: elaboración propia

Con el modelo resultante (Fig. 75), por la rama más a la izquierda, se puede concluir que el alumno que ha faltado alguna vez tiene un 93% (1-0.07) de probabilidad de suspender. Sin embargo, por la rama más a la derecha, el alumno que no ha faltado nunca y que desea obtener educación superior (mas_alto = yes) tiene un 65% (0.65) de probabilidad de aprobar.

	[0,12)	[12,19]
[0,12)	124	177
[12,19]	12	336

Tabla 26. Matriz de confusión generada para el árbol de la figura XX
Fuente: elaboración propia

Accuracy: 0.7087827 (71%)

En este caso, la matriz de confusión obtenida (Tabla 26), asociada a una discretización en dos tramos, tiene una precisión media del 71%. Lo que indica que el modelo resulta bastante bueno ya que, aunque clasifica los casos positivos con una probabilidad del 41,20%, los casos negativos los clasifica con una probabilidad del 96,55%. Esto puede derivar de los falsos aprobados (errores de tipo II) al igual que ocurre en el caso de ciencias, es decir, predicciones que indican que los alumnos van a aprobar [12,19] aunque realmente la calificación final obtenida por estos será suspensa [0,12).

MODELO PARA LA DISCRETIZACIÓN EN TRES TRAMOS
[0,11) [11,13) [13,19]

$$DISC.1 \sim faltas + mas\_alto + escuela$$

Fig. 76. Variables más influyentes usando DISC.1
Fuente: elaboración propia

En la obtención del modelo presentado a continuación, se tienen en cuenta las variables situadas en los primeros puestos del ranking una vez discretizada la variable "calificación final" en la asignatura de matemáticas en tres tramos. El árbol resultante se muestra en Fig. 77.

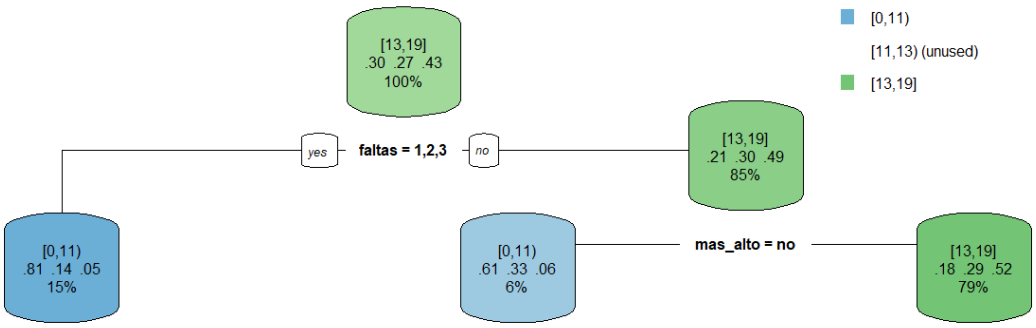

Fig. 77. Árbol de decisión obtenido con el método RPART
Fuente: elaboración propia

Obtenido este modelo (Fig. 77), por la rama más a la izquierda, se puede concluir que el alumno que ha faltado alguna vez tiene un 81% (0.81) de probabilidad de suspender. En cambio, por la rama más a la derecha, el alumno que no ha faltado nunca y que desea obtener educación superior (mas_alto = yes) tiene un 52% (0.52) de probabilidad de obtener una calificación alta.

	[0,11)	[11,13)	[13,19]
[0,11)	103	0	94
[11,13)	26	0	150
[13,19]	7	0	269

Tabla 27. Matriz de confusión generada para el árbol de la figura XX
Fuente: elaboración propia

Accuracy: 0.5731895 (57%)

De nuevo, se puede apreciar en la Tabla 27 que no se predice ninguna nota en su segmento central, por lo que, pese a una precisión media del 57%, esta discretización debe evitarse.

MODELO PARA LA DISCRETIZACIÓN EN CUATRO TRAMOS [0,10) [10,12) [12,14) [14,19]

DISC.3 ~ mas_alto + faltas + escuela + schoolsup

Fig. 78. Variables más influyentes usando DISC.3
Fuente: elaboración propia

Para la ejecución del modelo presentado a continuación, se tienen en cuenta las variables situadas en los primeros puestos del ranking obtenido al discretizar la variable "calificación final" en la asignatura de matemáticas en cuatro tramos. El árbol resultante se muestra en Fig. 79.

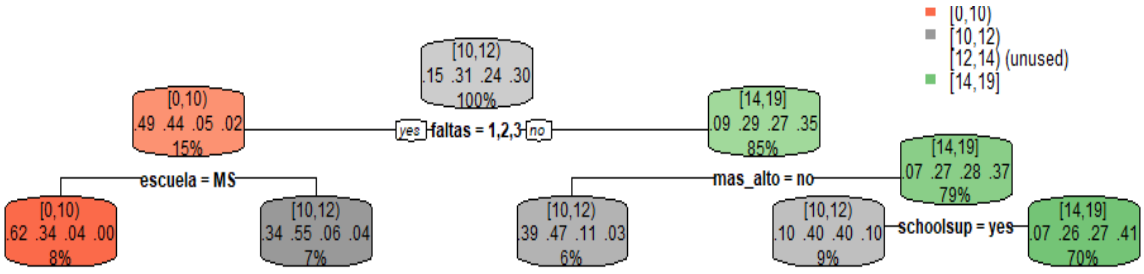

Fig. 79. Árbol de decisión obtenido con el método RPART
Fuente: elaboración propia

Al realizar este modelo (Fig. 79), por la rama más a la izquierda, concluimos que el alumno que ha faltado alguna vez y que cursa la asignatura de lengua portuguesa en la escuela Mousinho da Silveira (MS) tiene un 62% (0.62) de probabilidad de suspender. Sin embargo, por la rama más a la derecha, el alumno que no ha faltado nunca, que desea obtener educación superior (mas_alto = yes) y que no recibe apoyo educativo extra (schoolsup = no) tiene un 41% (0.41) de probabilidad de obtener una calificación muy alta.

	[0,10)	[10,12)	[12,14)	[14,19]
[0,10)	33	36	0	31
[10,12)	18	66	0	117
[12,14)	2	30	0	122
[14,19]	0	9	0	185

Tabla 28. Matriz de confusión generada para el árbol de la figura XX
Fuente: elaboración propia

Accuracy: 0.4375963 (44%)

Finalmente, se concluye según lo observado en la Tabla 28 que la discretización en 4 tramos no es adecuada, al no tener instancias clasificadas en el tercer segmento.

5.5 COMPARATIVA

	MATEMÁTICAS	***LENGUA PORTUGUESA***
DISC. EN DOS TRAMOS	*0.6607595 (66%)*	*0.7087827 (71%)*
DISC. EN TRES TRAMOS	*0.5291139 (53%)*	*0.5731895 (57%)*
DISC. EN CUATRO TRAMOS	*0.4734177 (47%)*	*0.4375963 (44%)*

Tabla 29. Accuracy de los modelos predictivos en ciencias y letras
Fuente: elaboración propia

Los resultados (Tabla 29) muestran que en el caso de ciencias (Matemáticas) con la discretización en dos tramos (***[0,11) [11,20]***) el método RPART logra los mejores resultados junto con la selección de características basada en el ranking de importancia de variables normalizado realizado con *VariableRanker* de la librería "MachineLearning". Aunque es la discretización en cuatro tramos la que permitiría intervenir tempranamente en el alumnado, a pesar de obtener una precisión final más baja en el modelo predictivo.

En cuanto a letras (Lengua portuguesa) las pruebas han indicado que el método RPART para la discretización en dos tramos (***[0,12) [12,19]***) logra los mejores resultados junto con la selección de características basada en el ranking de importancia de variables normalizado realizado con *VariableRanker* de la librería "MachineLearning".

Por tanto, podemos concluir que, en el caso de ciencias, es factible realizar una discretización de la variable calificación final (G3) en cuatro tramos para clasificar a los estudiantes utilizando aprendizaje automático con una precisión de (47%). En letras, también lo es, pero en este caso en dos tramos para poder clasificar a los estudiantes con alta precisión (71%).

6. INTERPRETACIÓN DE LOS RESULTADOS

Para efectuar la evaluación de los modelos se ha tomado como medida el porcentaje total de aciertos al clasificar una instancia en su respectiva clase. Esto se puede observar en la Tabla 28. donde viene representado la ***accuracy*** de los modelos predictivos en ciencias y letras.

Se puede concluir por tanto que, en el caso de ciencias, el modelo obtenido con la discretización en dos tramos de la variable ***calificación final*** (G3) tuvo un mejor desempeño que el que resulta con la discretización en cuatro tramos de esta variable objetivo. Pero al analizar las matrices de confusión (Tablas 23 y 25) se puede ver que el modelo extraído con la discretización en cuatro tramos permite una mejor injerencia en el alumnado, ya que tiene una precisión media aceptable (teniendo en cuenta los cuatro posibles segmentos de calificación) y muestra un reparto más equitativo de los errores de clasificación.

En cambio, en el caso de letras el modelo resultante de la discretización en dos tramos de la variable ***calificación final*** (G3) es el modelo que mejor clasifica al alumnado con un porcentaje total de aciertos del 71%.

7. CONCLUSIONES Y PROPUESTAS

Tras los experimentos realizados, ya estamos en disposición de responder a la pregunta planteada como hipótesis de partida: "¿Cómo afecta la factorización a la precisión de los modelos?".

La experiencia computacional llevada a cabo en las secciones anteriores pone de manifiesto que la factorización es crítica. Así, el número de segmentos generados sobre la variable objetivo (o los límites de cada segmento) conducen a modelos predictivos de diferentes precisiones medias. Y no sólo eso, sino que, además, en ocasiones las matrices de confusión asociadas a los modelos, ponen de manifiesto que la discretización propuesta fue arbitraria ya que no se clasifican instancias en algunos de sus segmentos.

Demostrada ya la importancia de la factorización sobre los modelos predictivos, se abre una línea de investigación sobre métodos para encontrar la factorización que conduce en cada caso a modelos predictivos óptimos (en caso de que los haya).

Desde el punto de vista meramente educativo, la conclusión que se desprende de este estudio es que la segmentación de las calificaciones a predecir (y los umbrales asociados a dichas segmentaciones) varía entre disciplinas de ciencias y de letras, así como probablemente lo haga en función de las respectivas escuelas. Es por ello, que el problema de la alerta temprana sobre el rendimiento de los estudiantes debe ser abordado desde un enfoque absolutamente dependiente del dato (data-driven).

A la vista de los resultados, otras líneas futuras de investigación que se abren son las siguientes:

- ✓ Aplicar modelos predictivos de regresión sobre calificaciones numéricas (no factorizadas) y comparar con precisiones obtenidas en los modelos de clasificación desarrollados en este trabajo.

- ✓ Diseñar nuevos procedimientos de selección automática de características que conduzcan a modelos precisos y sencillos en sistemas de alerta temprana en el ámbito educativo.

- ✓ Estudiar los mecanismos para implementar este tipo de árboles en DSS orientados a la planificación de la docencia

- ✓ Extender el estudio a otros ámbitos de aplicación distintos a la docencia, para poner de manifiesto qué factorizaciones son las más adecuadas, en cada caso, desde el punto de vista decisional.

8. BIBLIOGRAFÍA

CORTEZ, P. y SILVA, A. (2008): Uso de la Minería de Datos para predecir el rendimiento de los estudiantes de Educación Secundaria. En A. Brito y J. Teixeira Eds., Actas de la 5ta. Conferencia de TECnología de FUture BUsiness (FUBUTEC 2008) págs. 5-12, Oporto, Portugal, abril de 2008, EUROSIS.

FAJARDO BULLÓN, et al. (2017): Análisis del rendimiento académico de los alumnos de Educación Secundaria Obligatoria según las variables familiares. *Revista Educación XXI*, 20 (1), págs. 209-232.

GARCÍA, J. et al. (2018): Ciencia de datos. Técnicas analíticas y aprendizaje estadístico en un enfoque práctico. Editorial Altaria. Barcelona.

GONCALVEZ, T.C. et al. (2018): Técnicas de minería de datos: un estudio de caso de deserción en la educación superior del Instituto Federal de Maranhao. *Revista brasileña de computación aplicada*, Volumen 10, nº 3, noviembre 2018, págs. 11-20.

GOROSTIAGA, A. y ROJO-ÁLVAREZ, J.L. (2016): Sobre el uso de las técnicas convencionales y de aprendizaje para el análisis de los resultados PISA en España. *Revista Neurocomputing*, Volumen 171, 1 de enero de 2016, págs. 625-637.

GUTIÉRREZ, G. et al. (2018): Minería: Comentarios de los alumnos sobre la evaluación del desempeño docente utilizando algoritmos de aprendizaje automático. *Revista internacional de problemas de optimización combinatoria e informática*, Volumen 9, nº 3, septiembre - diciembre 2018, págs. 26-40.

HERNÁNDEZ ORALLO, J. et al. (2004): Introducción a la Minería de Datos. Editorial Pearson Educación. Madrid.

LIU, QT. et al. (2018): Teachers' online discussion text data shed light on their reflexive thinking. *Revista IEEE Transacciones en tecnologías de aprendizaje*, Volumen 11, nº 2, abril - junio 2018, págs. 243-254.

MARTÍNEZ MUÑOZ G. y SUÁREZ A. (2007): Using boosting to prune bagging ensembles. *Revista Pattern Recognition Letters*, Volumen 8, nº 1, enero 2007, págs. 156-165.

MENACHO CHIOK, C.H. (2017): Predicción del rendimiento académico aplicando técnicas de minería de datos. *Revista de anales científicos*, 78 (1), págs. 26-33.

MENES CAMEJO, I. et al. (2015): Desempeño de algoritmos de minería en indicadores académicos: Árbol de Decisión y Regresión Logística. *Revista cubana de ciencias informáticas*, Volumen 9, nº 4, octubre - diciembre 2015, págs. 104-117.

MUELAS, A. y BELTRÁN, J.A. (2011): Variables influyentes en el rendimiento académico de los estudiantes. *Revista de psicología y educación*, nº6, 2011, págs.173-196.

MURZINA, I.V. y KAZAKOVA, S.V. (2019): Direcciones de perspectiva de la educación patriótica. *Revista educación y ciencia*, Volumen 21, nº 2, febrero 2019, págs. 155-175.

NEELEMAN, A. (2019): El alcance de la autonomía escolar en la práctica: una clasificación empírica de las intervenciones escolares. *Revista de cambio educativo*, Volumen 20, nº 1, feb 2019, págs. 31-55.

QUINLAN, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers. Series in machine learning. USA. (1993). *Revista Machine Learning*, Volumen 16, nº3, septiembre 1994, págs 235-240.

QUINLAN, J.R. (1986): Inducción de árboles de decisión. *Revista Machine Learning*, Volumen 1, nº 1, marzo 1986, págs. 81-106.

SPOSITTO, O.M. et al. (2010): Aplicación de Técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. Memorias de la Novena Conferencia Iberoamericana de Sistemas. Cibernética e Informática (CISCI 2010), Orlando, Florida. EEUU.

TIMARAN, R. y JIMENEZ, J. (2014): Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM. Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación, noviembre 2014.

RECURSOS WEB

En este apartado se presentan diferentes enlaces web con materiales que pueden resultar de utilidad para la materia tratada en este trabajo.

1. Statistical Tools For High-Throughput Data Analysis: http://www.sthda.com

2. Modelización: modelos de clasificación: http://www.rpubs.com/JoanClaverol/414202

3. R para profesionales de los datos: una introducción https://datanalytics.com/libro_r/_main.pdf

4. Diagrama de Dispersión y Regresión en R https://rpubs.com/MSiguenas/97848

5. Rminer: Data Mining Classification and Regression Methods: https://cran.r-project.org/web/packages/rminer/index.html

6. Un análisis con R. Datos multivariantes: http://www.ub.edu/stat/docencia/EADB/Ejemplo.pdf

7. Feature selection techniques with R: http://dataaspirant.com/2018/01/15/feature-selection-techniques-r/

8. Importancia de variables: https://rpubs.com/gualberto/importancia_variables

9. Package 'randomForest': https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

10. Machine Learning: http://www.diegocalvo.es/

11. Análisis y Tratamiento de Datos: aplicaciones con R: http://www.dma.ulpgc.es/profesores/personal/asp/Docencia/doctoMed/RdoctoMed.pdf

12. Regla de Sturges: https://es.wikipedia.org/wiki/Regla_de_Sturges

13. Feature Selection Techniques in Machine Learning with Python: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

14. Estudiar en Portugal. Estructura del sistema educativo: http://www.universia.es/estudiar-extranjero/portugal/sistema-educativo/estructura-del-sistema-educativo/2315

15. Estructura del sistema educativo en Portugal: https://www.dgb.sep.gob.mx/tramites/revalidacion/Estruc_sist_edu/Estud-PORTUGAL.pdf

16. Regla de Sturges: Explicación, Aplicaciones y Ejemplos: https://www.lifeder.com/regla-sturges/

17. What is the Difference Between Accuracy and Precision?: https://www.thoughtco.com/difference-between-accuracy-and-precision-609328

18. Inducción de árboles de decisión: https://doi.org/10.1007/BF00116251