**Métodos**

Teniendo en cuenta que el objetivo del trabajo es predecir la variable numérica **G3**, es decir, la nota final de cada alumno; se optó por realizar una regresión lineal con las variables provenientes de los *datasets*. Además, al ver que el número de variables que poseían los *dataset*s eran numerosas, se optó llevar a cabo una selección de variables por medio de diferentes métodos. El objetivo de reducir la cantidad de variables a utilizar es determinar las variables más significativas de cada *dataset* para llevar a cabo la predicción de la nota final. Los métodos utilizados para la selección de variables fueron la selección Stepwise y los métodos de regularización: Regresión Lasso, Ridge, y Elastic Net.

**Selección *Stepwise***

El método de selección *Stepwise* se basa en realizar diversos modelos de regresión con las variables explicativas que se le presente. Para cada modelo creado por el método, se considera una variable a ser adicionada o eliminada según el criterio de significación que esta posea. Mientras que para determinar los mejores modelos de regresión creados se utilizaron diversos criterios: ajustado, criterio de información bayesiano (BIC) y el Cp de Mallow. (explicar que son cada uno de ellos). En nuestro caso, el modelo seleccionado fue el indicado por el Cp de Mallow, ya que los criterios ajustado y BIC determinaban un rango en el cual el modelo no sufría de *underfitting* ni *overfitting* y el estadístico Cp de Mallows siempre se encontraba dentro de este rango establecido.

En la selección de variables por método ***forward*** se empieza con un modelo sencillo, es decir, que no contenga ninguna variable y, en cada paso, se van incorporando aquellas variables que otorguen una mejora estadísticamente significativa al modelo. Mientras que en el método ***backward*** se parte de un modelo complejo, que posee todas las variables explicativas, y en cada iteración se elimina la variable que produzca menor impacto al ajuste del modelo, es decir, que sea considerado estadísticamente insignificante. Luego, se tiene el método ***hybrid*** o también conocido como ***stepwise*** que consiste en una combinación de los métodos anteriormente mencionados. Este procede de la misma forma que el método *forward* pero se diferencian en el hecho de que cada variable adicionada también puede ser eliminada, esto es posible ya que la inclusión de una nueva variable puede provocar que otra dentro del modelo resulte redundante.

**Métodos de Penalización**

Los métodos mencionados anteriormente muchas veces son considerados inestables, una alternativa que se puede implementar son los métodos de regresión penalizada o, también conocidos como métodos de regresión contraída. La idea principal es evitar el *overfitting* debido al gran número de variables predictoras imponiendo una penalización o término de penalización (*tuning parameter)* **𝜆**, que de cierta forma permitirá “contraer” a los coeficientes estimados hacia el valor 0. Esto puede resultar en la disminución de manera significativa de la varianza, atenuar el efecto de correlación y minimizar la influencia en el modelo de los predictores menos relevantes. Entre las técnicas de regresión penalizada se encuentran: Regresión Ridge, Lasso y Elastic Net.

**Regresión *Ridge***

El método de regresión Ridge es similar al de mínimos cuadrados, a excepción de que los coeficiente se estiman minimizando una expresión diferente. Los coeficiente estimados por Ridge son aquellos que minimizan,

donde 𝜆 ≥ 0 es el parámetro de penalización o *tuning parameter.* Este método tiende a disminuir los coeficientes de regresión en función al término de penalización, es decir, cuanto mayor sea 𝜆, mayor es la penalización. En la expresión se puede ver que si , el término de penalización no tendrá ningún efecto y el resultado obtenido será idéntico al de mínimos cuadrados. Por lo que, la selección de dicho parámetro es esencial para la regresión. Hay que destacar que el método *Ridge* tiene un efecto de reducción de forma proporcional de todos los coeficiente pero sin que estos lleguen a cero, por lo que, los modelos obtenidos por dicho método van a contener todas la variables explicativas. Pero se diferencia respecto a **OLS** en el trade-off sesgo-varianza.

**Regresión *Lasso***

La regresión *Lasso* es similar a *Ridge* pero su gran diferencia radica en que este método tiene un efecto a forzar a que los coeficientes de los predictores tiendan a cero. Esto sucede ya que *Lasso* penaliza la suma del valor absoluto de los coeficientes de regresión, como se puede ver en la siguiente expresión,

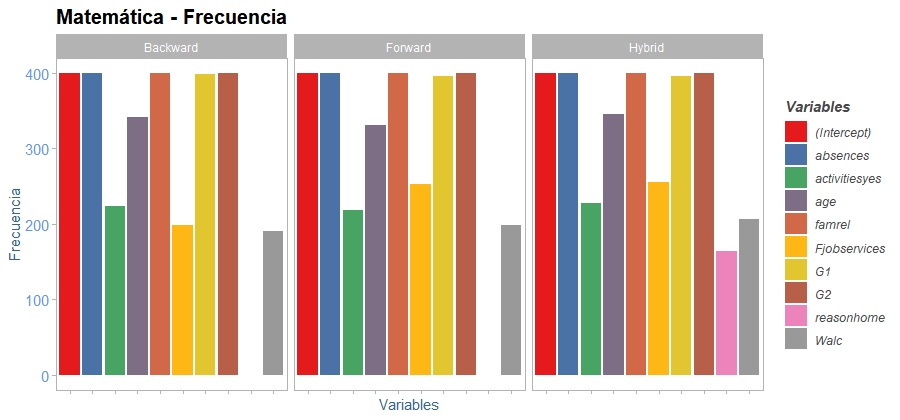
Esto permite obtener modelos con variables reducidas ya que *Lasso* permite eliminar los predictores menos relevantes.

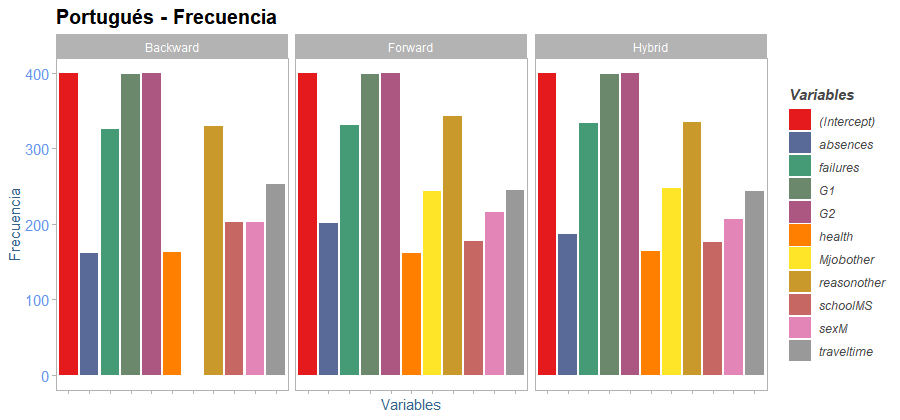
**Regresión *Elastic Net***

Por último, la regresión por el método *Elastic Net* es un método de penalización que combina la penalización por *Ridge* y por *Lasso.* Para determinar el grado de influencia de cada penalización se utiliza un parámetro α. Este es un parámetro que tiene un valor comprendido entre 0 y 1. Cuando el valor de α = 0, se aplica la penalización de *Ridge* y cuando α = 1, se aplica *Lasso*. El objetivo de utilizar *Elastic Net* es lograr seleccionar predictores y obtener una cierta estabilidad en el caso de que algunos de ellos estén correlacionados.

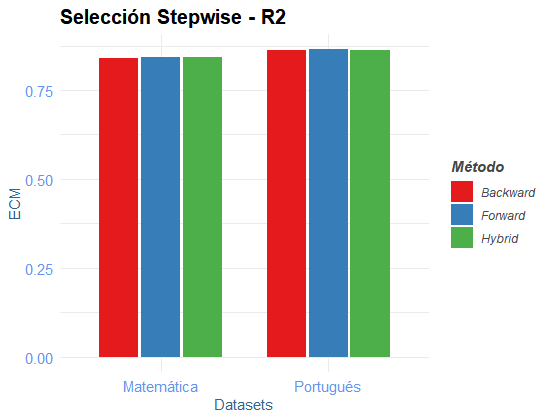
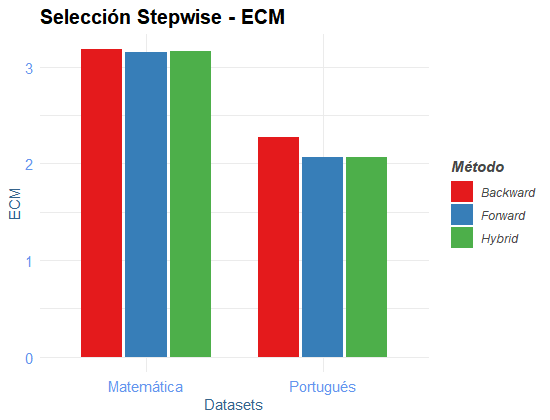
Cabe aclarar que para encontrar el valor óptimo de *tuning parameter* se utilizó *cross validation* en todos los casos anteriormente mencionados.

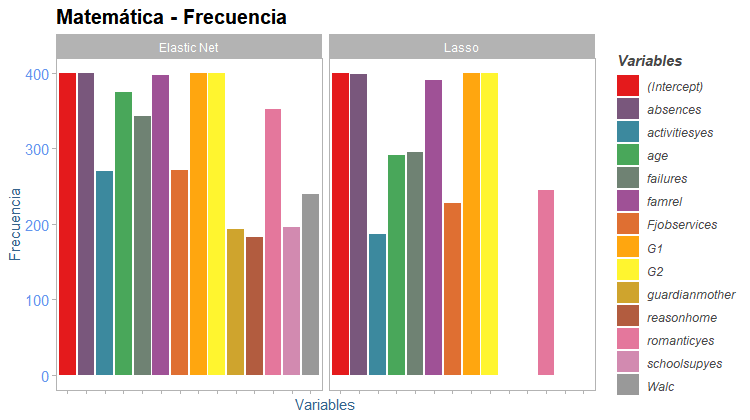
**Resultados**

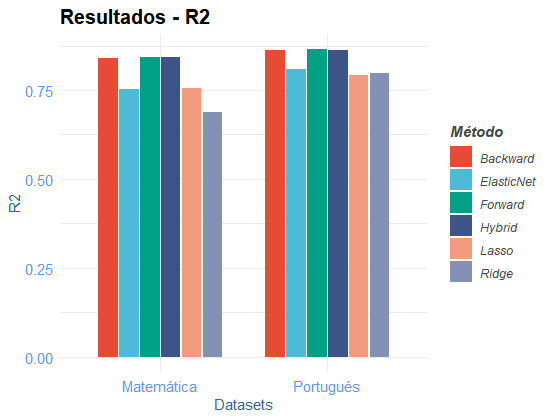
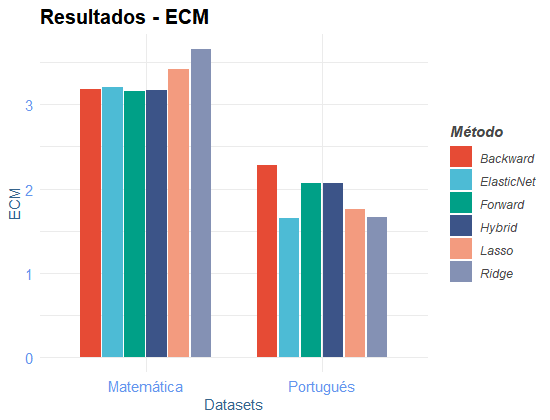
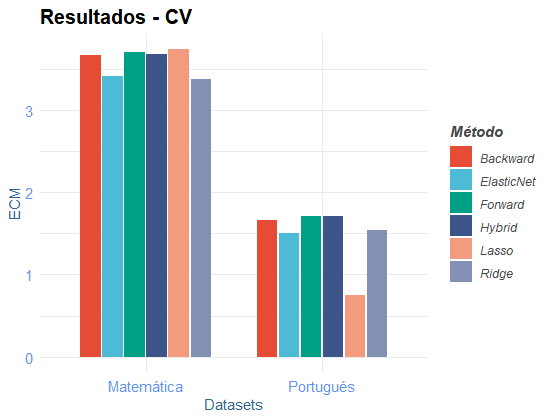
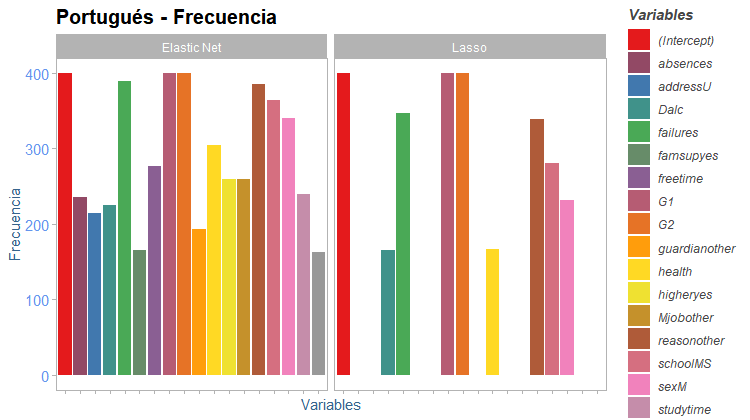
Para llevar a cabo el análisis de los *dataset* con los métodos descritos anteriormente se dividieron los datos en dos grupos: en entrenamiento (80%) y testeo (20%). Luego de haber entrenado el modelo de regresión ante los datos de entrenamiento se utilizaron los datos de testeo para determinar: el error cuadrático medio y el de cada modelo. Sin embargo, al realizar varios entrenamientos de los modelos con los métodos de selección *stepwise* se pudo ver que las variables que estos usaban para crear el modelo de regresión no eran consistentes. Es por esta razón, que se analizaron la frecuencia de cada una de las variables y se seleccionaron aquellas que cumplan un cierto umbral. Este umbral fue seteado a la mitad de la máxima frecuencia, es decir, que se seleccionaron aquellas variables que poseían una frecuencia mayor a la mitad de la frecuencia máxima. Los resultados en términos de frecuencia de aparición de las variables son los siguientes,



A partir de los datos obtenidos de las gráficas se determinó que el criterio establecido para ambos *datasets* es que el umbral utilizado sea de 200. Por lo que, se utilizaron para armar el modelo de regresión aquellas variables que tuvieron un rango de aparición mayor a 200 veces. El error cuadrático medio y el de los modelos fueron los siguientes,



En el caso de las regresiones mediante *Ridge, Lasso, y Elastic Net* únicamente se analizaron las frecuencia de selección de variables para determinar aquellas que fueron consideradas relevantes por los algoritmos. Se llevó a cabo esto debido a que los tres métodos utilizados se basan principalmente en encontrar un *tuning parameter* (𝜆)y luego a partir de este crear el modelo de regresión según las variables determinadas. Las frecuencias de aparición de las variables son,

Por último, se analizaron los errores cuadráticos medios, los ECM obtenidos por *cross validation* y el de todos las metodologías utilizadas,