Desempeño estudiantil

Gastón Barcelo, Tisiana Franco, Agustín Macció

Introducción:

Este trabajo tiene como objetivo investigar los factores que influyen en el rendimiento académico de los estudiantes de secundaria, utilizando el conjunto de datos disponible en el portal **UCI Machine Learning Repository.**

El dataset recopila información (durante el año lectivo 2005/2006) detallada sobre el rendimiento de alumnos de dos escuelas secundarias en Portugal, considerando diversas variables. El archivo original se divide en dos partes: una que contiene datos de la asignatura de Matemáticas y otra de Lengua Portuguesa. En este estudio, nos enfocaremos únicamente en el análisis del desempeño en Matemáticas.

A partir de este caso especifico, buscamos identificar posibles relaciones entre características personales, familiares y escolares y las calificaciones finales, sin la intención de generalizar los resultados más allá de este contexto particular. Para ello, se emplearán técnicas de análisis exploratorio de datos y se construirá un modelo estadístico que permita predecir la nota final de los estudiantes. Finalmente, se desarrollará una aplicación Shiny para presentar de manera interactiva los resultados más relevantes.

Esperamos que este trabajo sirva como un ejemplo de la aplicación de herramientas de ciencia de datos en el ámbito educativo y como un punto de partida para reflexionar sobre factores que podrían ser significativos en contextos similares.

Datos:

```
Rows: 395
Columns: 33
$ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP",~
$ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F",~
```

```
$ age        <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15,~
$ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U",~
$ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "LE~
$ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T",~
$ Medu       <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,~
$ Fedu       <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3,~
$ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "servic~
$ Fjob       <chr> "teacher", "other", "other", "services", "other", "other", ~
$ reason     <chr> "course", "course", "other", "home", "home", "reputation", ~
$ guardian   <chr> "mother", "father", "mother", "mother", "father", "mother",~
$ traveltime <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,~
$ studytime  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,~
$ failures   <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0,~
$ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no", "n~
$ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes",~
$ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes", ~
$ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye~
$ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "yes~
$ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye~
$ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes",~
$ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "no"~
$ famrel     <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,~
$ freetime   <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1,~
$ goout      <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3,~
$ Dalc       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
$ Walc       <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3,~
$ health     <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5,~
$ absences   <dbl> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, 16,~
$ G1         <dbl> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 14, ~
$ G2         <dbl> 6, 5, 8, 14, 10, 15, 12, 5, 18, 15, 8, 12, 14, 10, 16, 14, ~
$ G3         <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14,~
```

El dataset utilizado como ya mencionamos procede de **UCI Machine Learning Repository** (repositorio de datos abiertos internacional).

El archivo seleccionado corresponde a información recopilada en dos escuelas secundarias portuguesas, específicamente sobre estudiantes de la asignatura de Matemáticas. El dataset contiene un total de **395 observaciones**, cada una correspondiente a un estudiante individual.

En total se incluyen **33 variables**, que abarcan características académicas, personales, familiares y sociales. Entre ellas se encuentran variables como:

Información demográfica: edad, género, si vive en lugar urbano o rural.

Contexto familiar: nivel educacional de los padres, categoría laboral de los padres, relación familiar.

Aspectos academicos: tiempo de estudio, inasistencias, notas parciales y finales.

Variables socioeconómicas: actividades extracurriculares, apoyo educativo, comportamiento de consumo de alcohol, entre otras.

Este cantidad de variables ofrece la posibilidad de investigar varios factores los cuales pueden tener influencia en el comportamiento escolar y facilita el uso de técnicas de análisis exploratorio y modelado estadístico.

La variable `G3` es la nota final anual. Mientras que `G1` y `G2` corresponden a las notas de primer y segundo período respectivamente.

Para el mejor entendimiento de las variables de calificaciones recién mencionadas dejamos una breve explicación del funcionamiento del sistema de evaluación en estas escuelas portuguesas:

Se adoptan calificaciones de 0 a 20 de escala, donde se requiere 10 de calificación mínima para aprobar. El cierre del curso académico se califica a través de tres momentos a lo largo del año escolar, dados por las variables `G1`, `G2` y `G3`, correspondientes a calificación del primer período, del segundo período y a la calificación del cierre del curso, correspondientemente.

Es importante mencionar que, una calificación final de 0 podría reflejar que el estudiante no se presentó a clase o a evaluaciones clave. Sin embargo, el dataset no aclara si existe una regla administrativa de 0 en caso de inasistencia o abandono del curso..

Análisis Exploratorio:

El objetivo de este análisis exploratorio es determinar si las variables que, en un primer momento, seleccionamos por considerarlas explicativas de la nota final (`G3`), realmente cumplen con esa función.

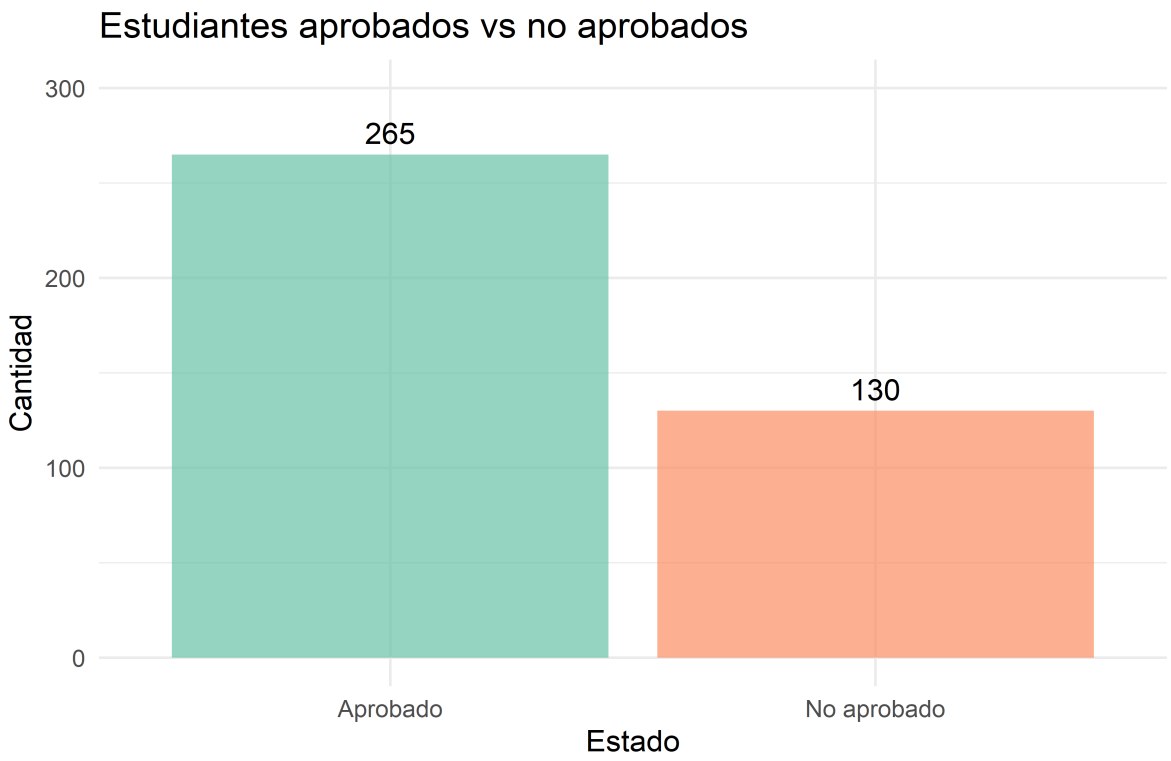


Figura 1: Estudiantes aprobados vs no aprobados

Este gráfico de barras compara la cantidad de estudiantes que aprobaron con los que no aprobaron, según la nota final (`G3`). Aunque hay una mayoría que logra aprobar, **un 33% aproximadamente no lo logra** (130 de 395 estudiantes en total).

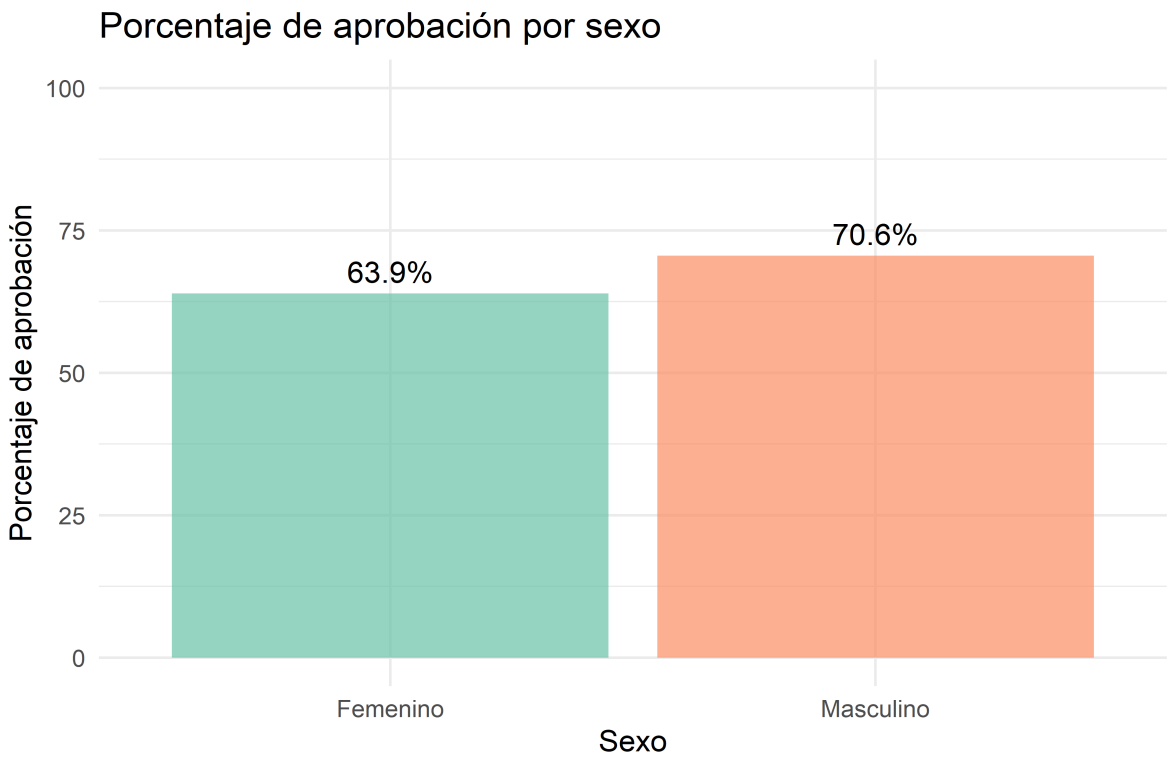


Figura 2: Sexo

Esta visualización muestra el porcentaje de estudiantes aprobados según su sexo: femenino o masculino.

En este grupo de datos, los estudiantes masculinos tienen una tasa de aprobación más alta que las estudiantes femeninas.

70.6% de los estudiantes masculinos aprueban, mientras que **63.9%** de las estudiantes femeninas aprueban.

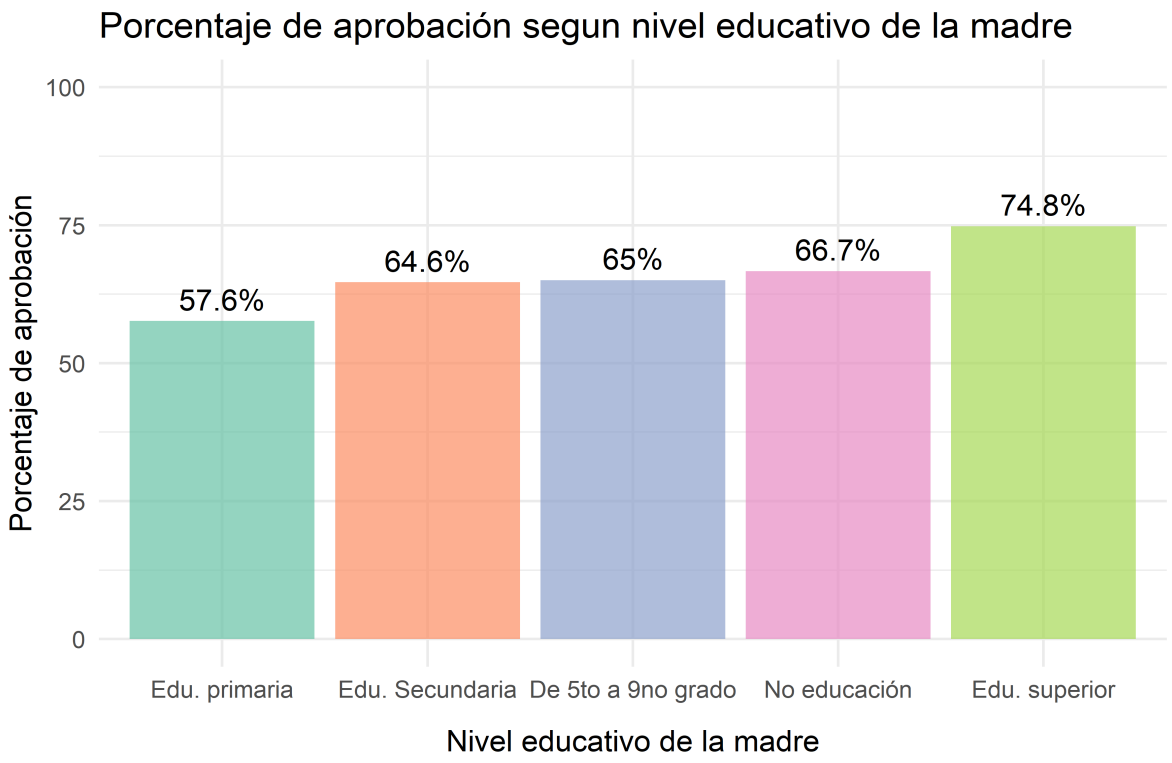


Figura 3: Educación de la madre

Compara el porcentaje de estudiantes que aprueban (nota final suficiente) en función del nivel educativo alcanzado por la madre.

- Los hijos de madres con educación superior tienen la tasa de aprobación más alta (**74.8%**).
- Los hijos de madres con educación primaria presentan la tasa más baja (**57.6%**).
- Entre medias, los niveles de secundaria, 5to a 9no grado, y sin educación rondan entre **64% y 67%**, sin una diferencia notoria.
- Llama la atención que el grupo "sin educación" **(66.7%)** no es el más bajo: supera a primaria y se parece a secundaria y 5to-9no grado.

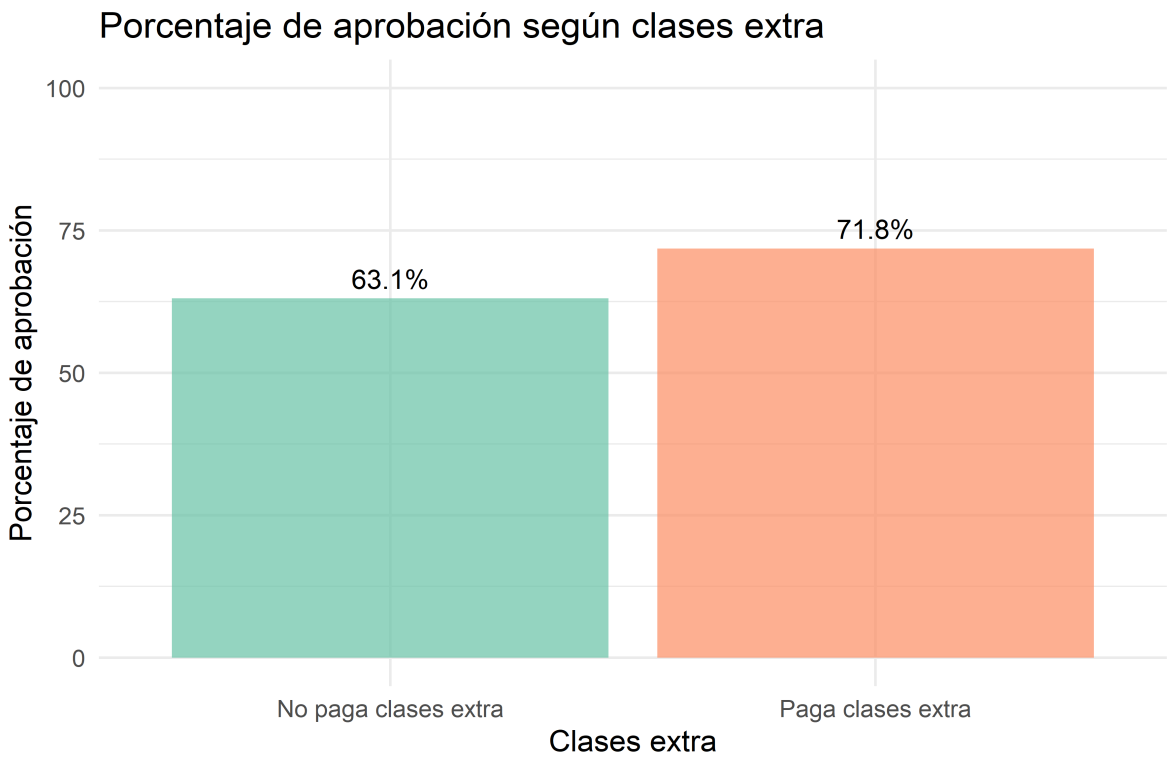


Figura 4: Clases extra pagas

Aqui podemos visualizar la tasa de aprobación de estudiantes segun si **pagan clases extra** o no.

- Los estudiantes que pagan clases extra tienen una tasa de aprobación de **71.8%**.
- Los que no pagan clases extra tienen una tasa de aprobación de **63.1%**.

Los estudiantes que reciben clases extra tienen más probabilidades de aprobar que quienes no lo hacen. Este resultado respalda la idea de que el refuerzo académico adicional (clases particulares o tutorías) tiene un impacto positivo en el rendimiento escolar.

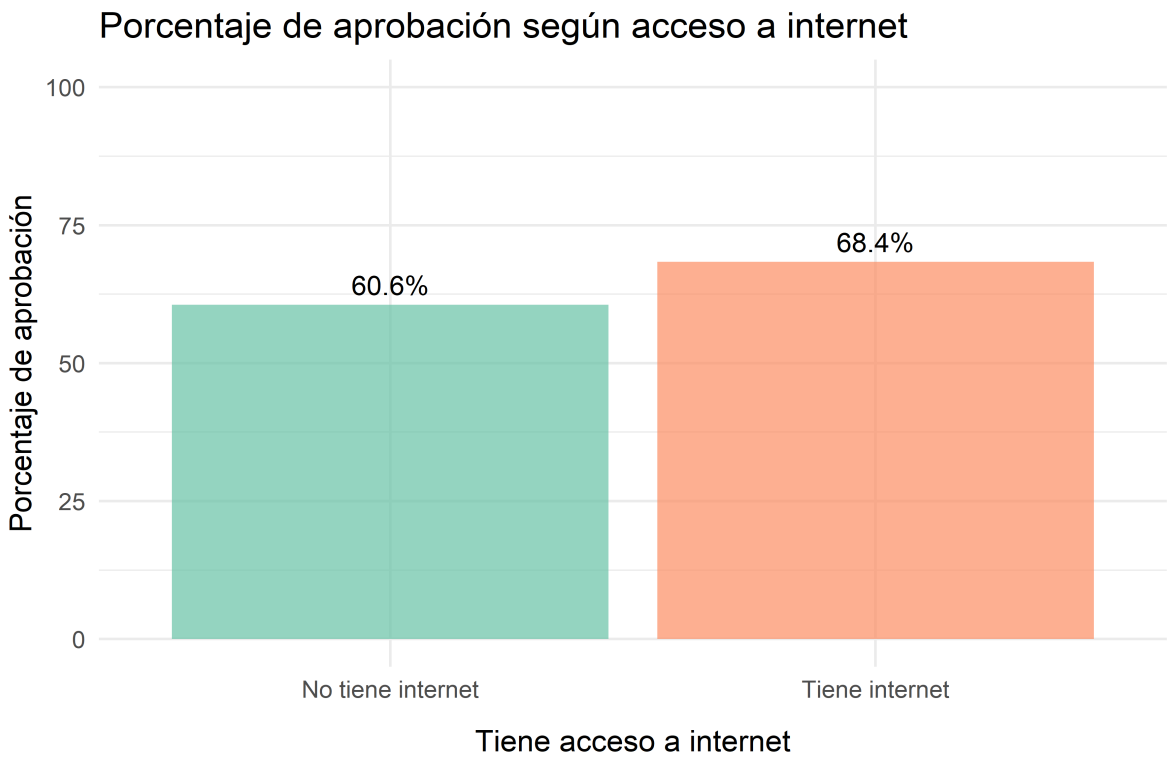


Figura 5: Acceso a internet

Este gráfico compara la tasa de aprobación de estudiantes con acceso a internet en casa frente a aquellos que no tienen acceso.

- Los estudiantes con internet tienen una tasa de aprobación de **68.4%**.
- Los estudiantes sin internet tienen una tasa de aprobación de **60.6%**.

Esto concluye a que tener acceso a internet, almenos en esta investigación, está asociado a una mayor probabilidad de aprobar, la diferencia es de **7.8** puntos porcentuales.

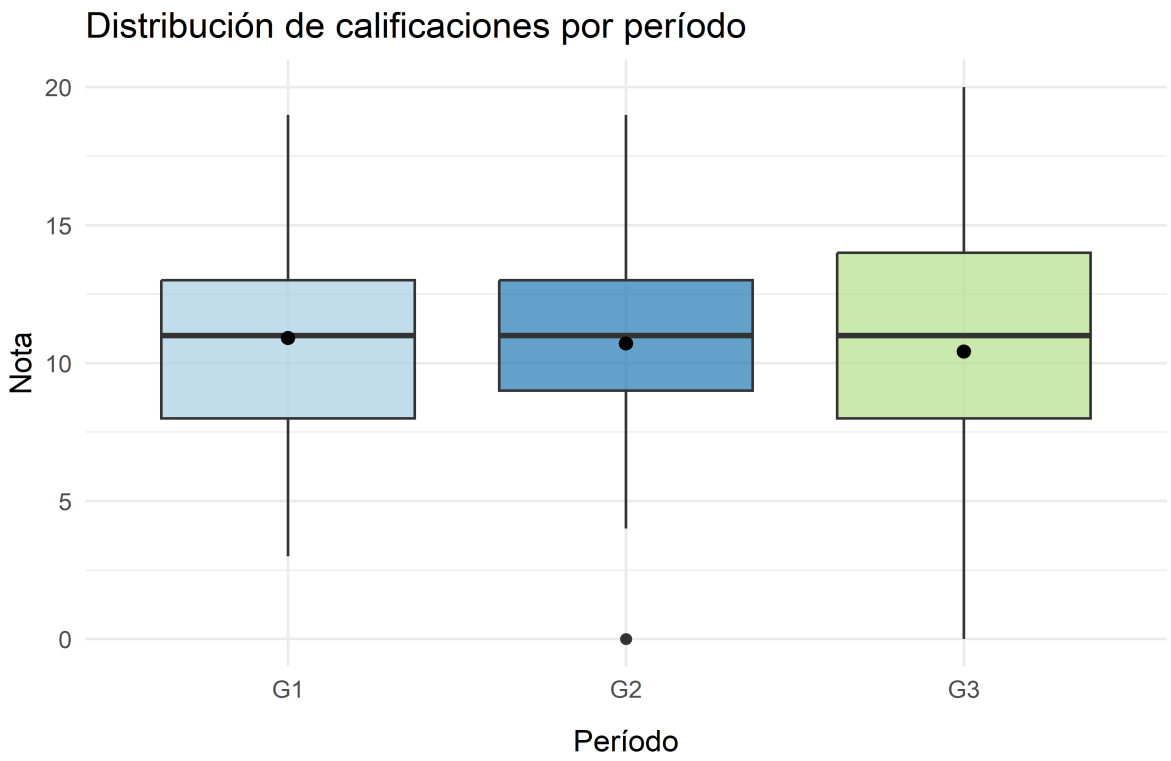


Figura 6: Box plot de clasificacion por periodo

Al observar la primera gráfica, vemos que la mediana y la distribución de las calificaciones de `G1`, `G2` y `G3` son bastante similares. Es decir, como grupo, los estudiantes no muestran un cambio drástico (ni mejora ni empeora fuertemente) a lo largo del año escolar. Por lo que las notas del primer y segundo periodo (`G1` y `G2`) pueden predecir bastante bien la nota final (`G3`).

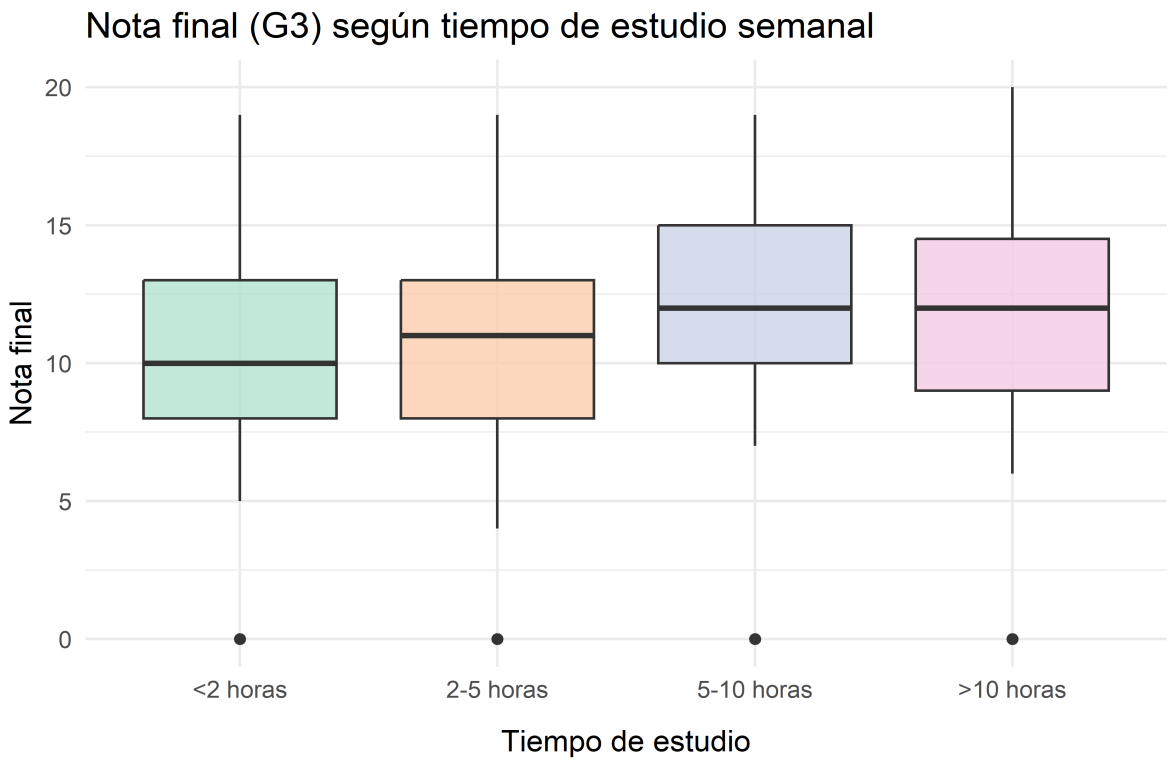


Figura 7: Box plot de nota final segun tiempo de estudio

Más horas de estudio tienden a relacionarse con mejores calificaciones finales, especialmente hasta el rango de 5-10 horas semanales. Estudiar más de 10 horas no garantiza mejores resultados, lo que sugiere que otros factores (como la calidad del estudio, apoyo docente o factores personales) también juegan un papel importante. Además, en todos los grupos hay alta variabilidad y valores bajos (0), lo que reafirma esa sugerencia.

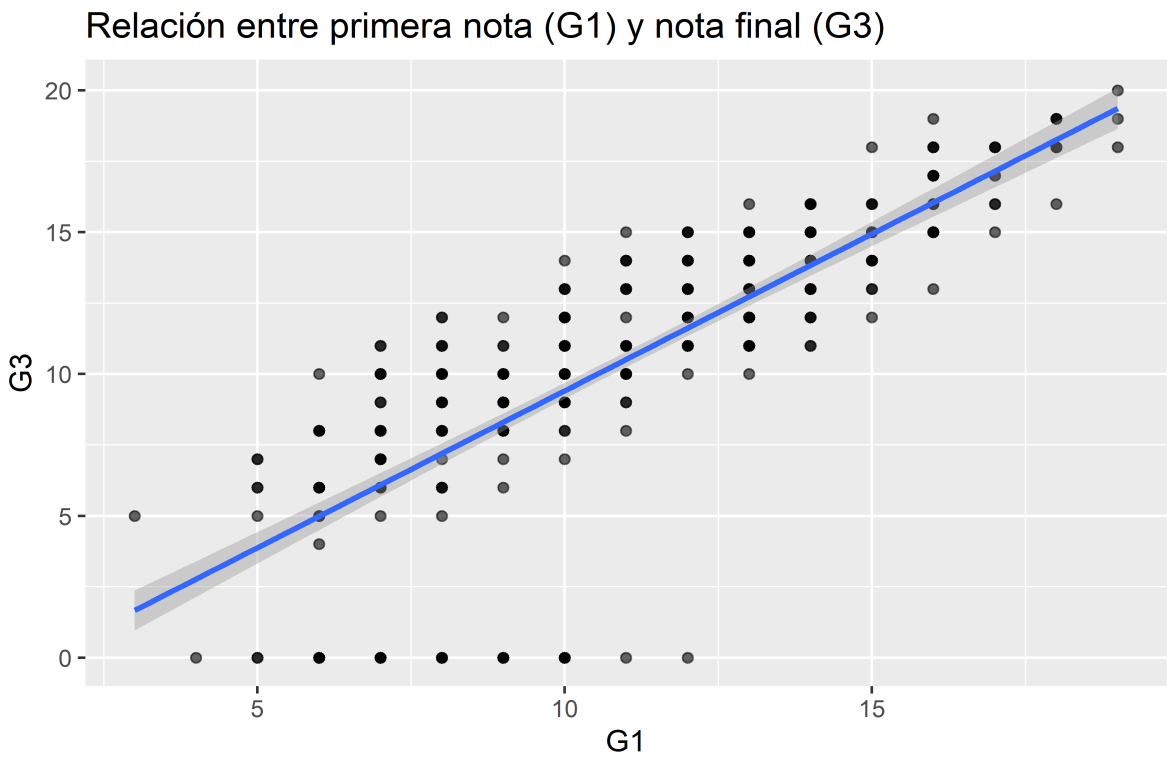


Figura 8: Relación entre G1/G2 y G3

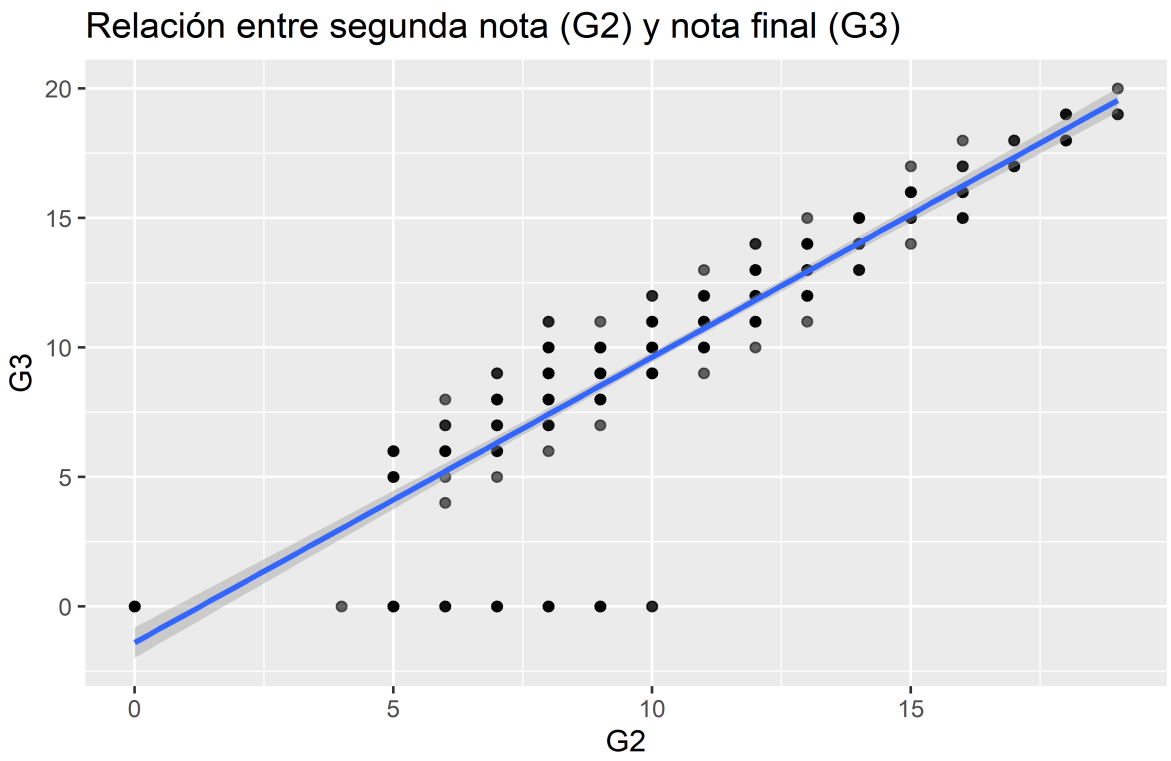


Figura 9: Relación entre G1/G2 y G3

JUSTIFICACION DE POR QUE SACAMOS G1 Y G2 DE VARIABLES PREDICTORAS:

Se observa una relación lineal positiva clara entre la primera calificación parcial (`G1`) y la calificación final (`G3`), lo que **indica que el rendimiento inicial de los estudiantes es un buen predictor de su rendimiento final.** La baja dispersión alrededor de la línea de tendencia refuerza la idea de que `G1` es una variable fuertemente predictiva. Lo mismo ocurre con `G2`. Por este motivo, **decidimos no incluir** `G1` y `G2` como variables explicativas al construir el modelo, además, en la misma página del dataset dice: *'It is more difficult to predict G3 without G2 and* ***G1****, but* ***such prediction is much more useful****'*. En base a esto, optamos por evaluar únicamente otras variables para explicar la nota final.

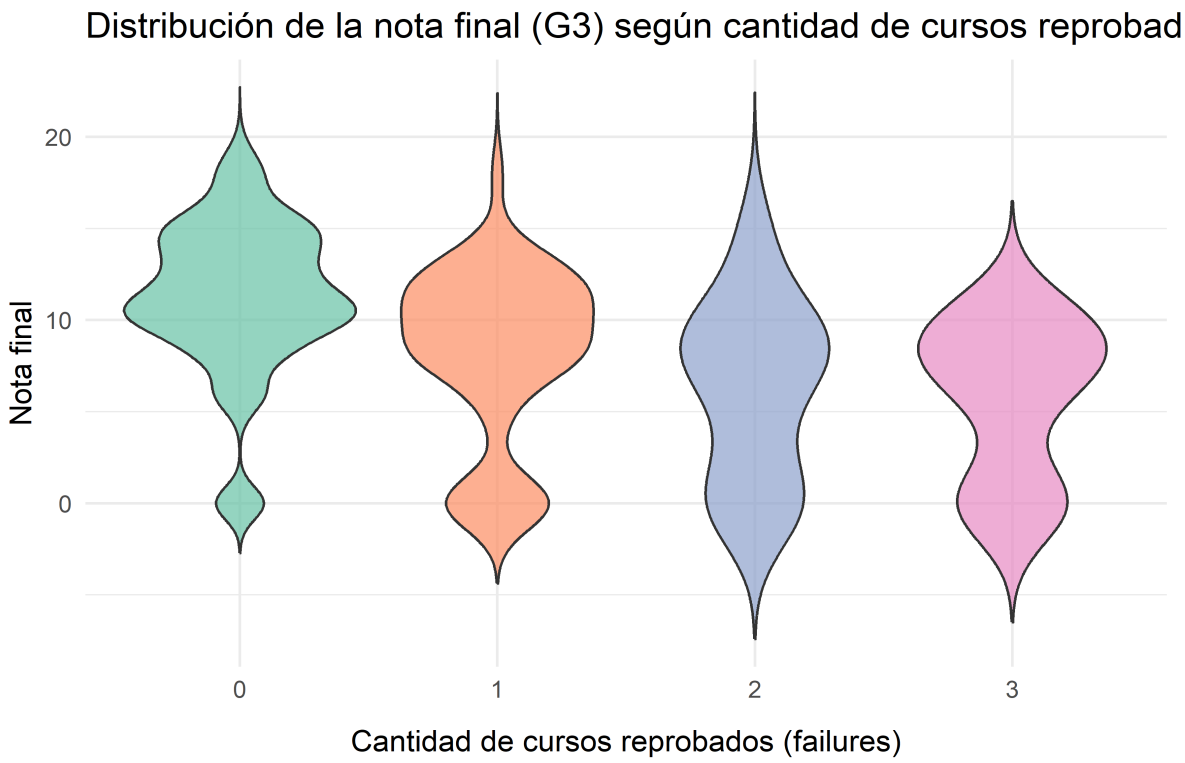


Figura 10: Violin plot de nota final segun cantidad de cursos reprobados

En este grafico de violin podemos observar conclusiones claras: A medida que aumenta la cantidad de cursos reprobados (failures), la distribución de la nota final (`G3`) tiende a bajar.

Los estudiantes sin cursos desaprobados (`failures` $= 0$) presentan una concentración de notas finales más alta, con la mayoría de los valores agrupados entre 10 y 15 puntos, lo que indica un buen desempeño general.

Para las puntuaciones de `failure` $= 2$ o 3, la distribución es mucho más dispersa y se extiende hacia notas muy bajas, llegando incluso a acumularse en valores cercanos a 0, lo que muestra una alta probabilidad de no aprobar.

El gráfico confirma el resultado esperado: no aprobar cursos anteriores es un buen predictor de bajo rendimiento final.

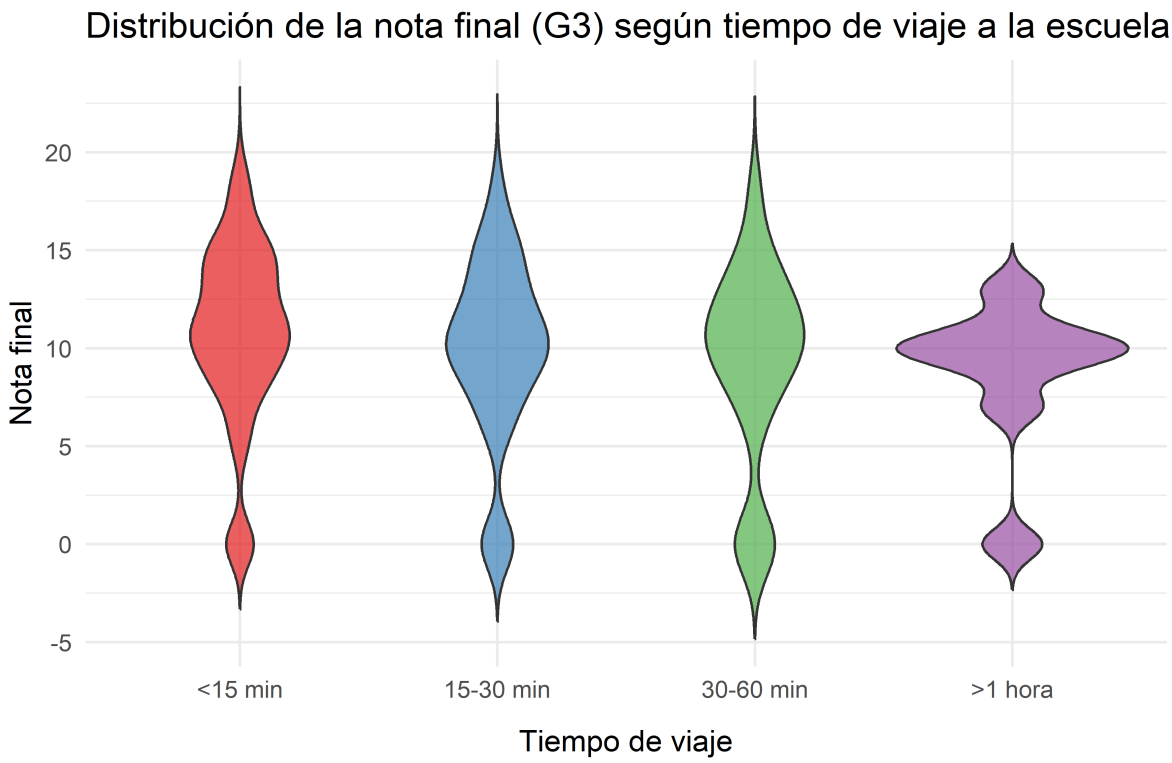


Figura 11: Violin plot de nota final segun tiempo de viaje

Los estudiantes con tiempos de viaje menores a 15 minutos, entre 15-30 minutos y entre 30-60 minutos presentan distribuciones de notas relativamente amplias y centradas alrededor de la media (aproximadamente entre 10 y 12). Esto significa que estos tiempos de viaje no se relacionan de forma fuerte o directa con grandes diferencias en el rendimiento académico.

Los estudiantes con tiempos de viaje mayores a 1 hora muestran una distribución más estrecha y concentrada en torno a notas cercanas a 10, con menor dispersión y una leve tendencia a agruparse en calificaciones más bajas que los grupos con viajes más cortos. Esto podría indicar que los tiempos de viaje muy largos se asocian con un rendimiento algo más bajo o con menor variabilidad (menos probabilidad de alcanzar notas muy altas).

Los grupos con trayectos cortos (<15 min) o medianos (15-60 min) presentan distribuciones más amplias en las notas altas (15-20). Esto podría sugerir que estudiantes con trayectos más cortos pueden llegar a tener mayores posibilidades de obtener calificaciones más altas, quizás porque les queda más tiempo para estudiar o porque el trayecto es menos agotador y están más descansados en clase.

Estas conclusiones nos generó una nueva pregunta; ¿Más tiempo de viaje es indicador de mayor cantidad de faltas?

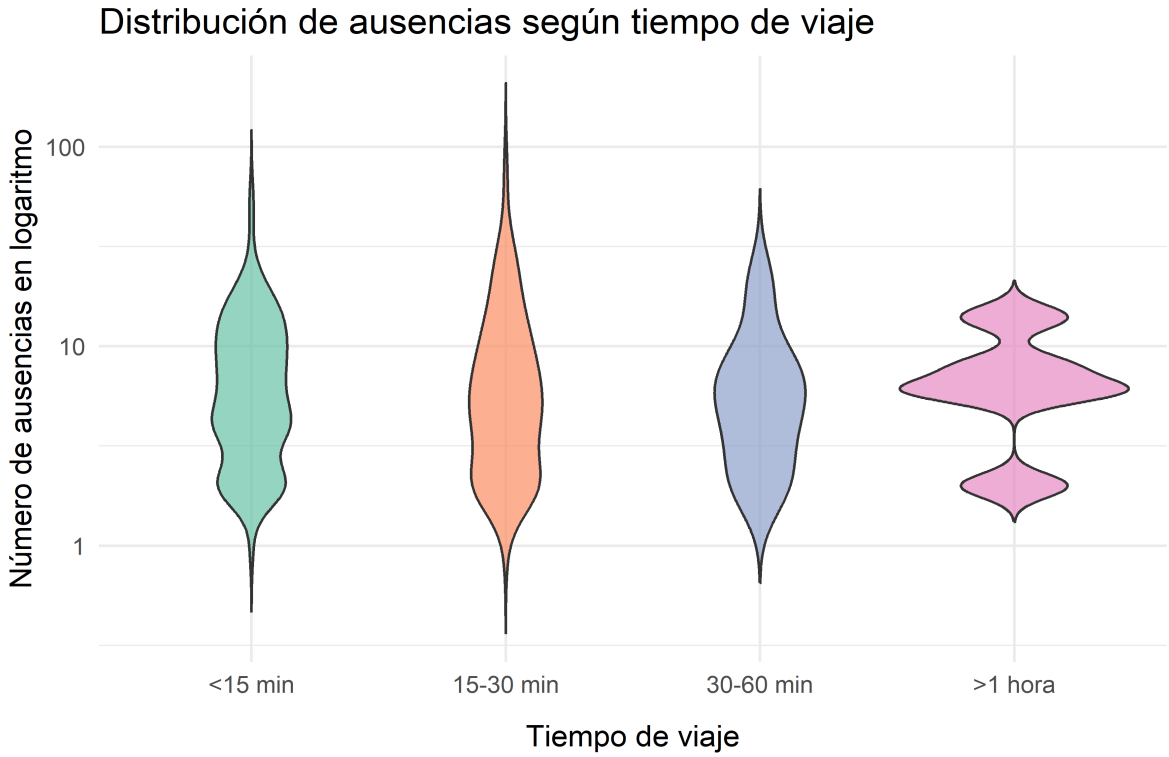


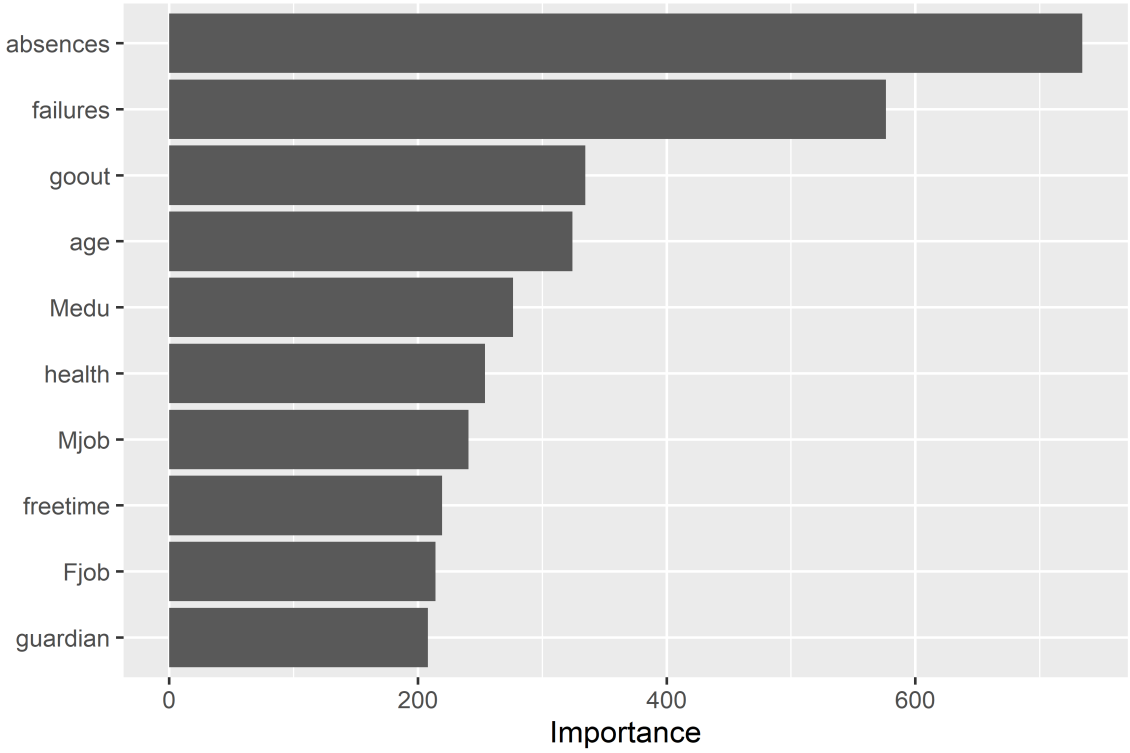
Figura 12: Violin plot de ausencias (escala log) segun tiempo de viaje

El gráfico de violín nos da la distribución del número de ausentes (escalada logarítmicamente) con el tiempo de viaje del estudiante en cuatro intervalos: menos de 15 minutos, entre 15–30 minutos, entre 30–60 minutos y más de 1 hora. A gran escala se observa que quienes residen a mayor distancia (>1 hora) tienden a acumular más ausencias con una distribución concentrada en valores más altos, también parece haber mas densidad en la parte superior del violín, indicando que hay más individuos con altas ausencias en este intervalo, mientras quienes viven a poca distancia (<15 min) suelen tener menos ausencias y presentan menor dispersión. Los intervalos intermedios (15–30 min y entre 30–60 min) muestran mayor variabilidad, lo que sugiere que hay otros factores que influyen en el ausentismo. En conclusión, se identifica una relación positiva entre mayor tiempo de viaje y mayor tendencia a acumular ausencias, si bien este patrón sugiere que el tiempo de viaje podría estar asociado con mayores niveles de ausentismo, otros factores pueden influir y la relación no es necesariamente causal.

Modelo:

Como ya mencionamos anteriormente, nuestra variable objetivo o de respuesta ~y~ es la nota final (`G3`) y usaremos como variables predictoras todas menos `G1` y `G2`.

El modelo elegido es Random Forest ya que nos parece un excelente modelo predictor. En este contexto usaremos una cantidad de 1000 árboles. No realizamos ningún podado.



A continuación un pantallazo de la tabla de predicción:

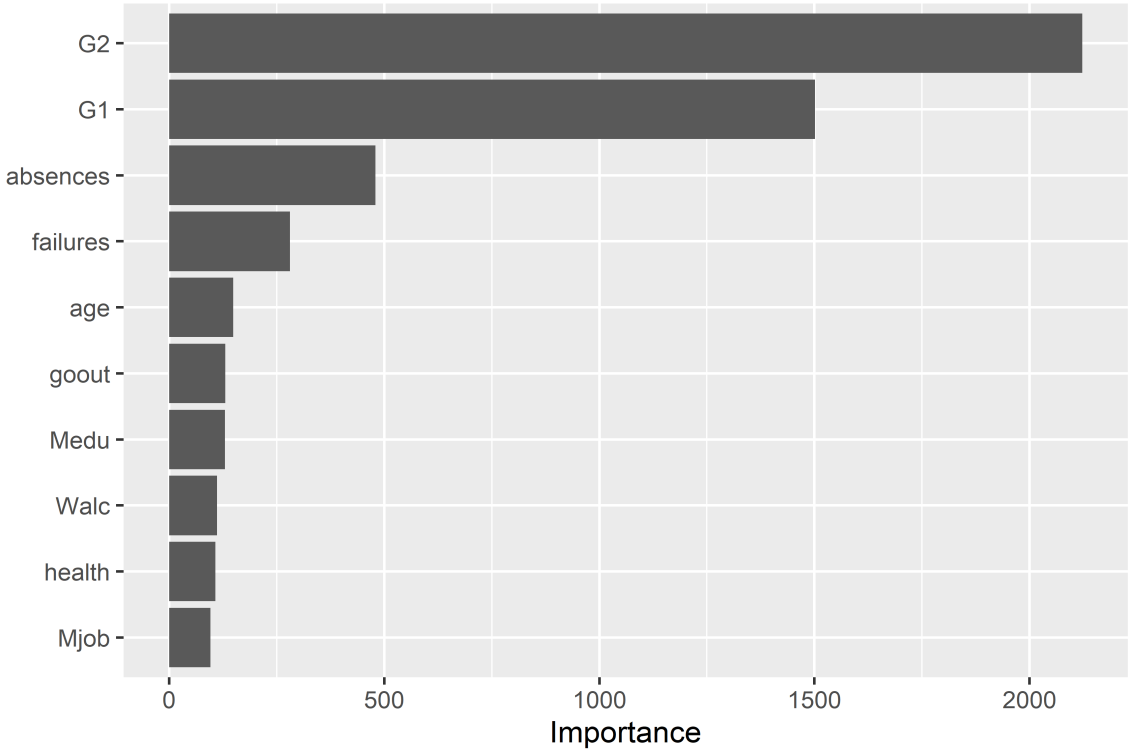
```
# A tibble: 6 x 2
     G3 .pred
  <dbl> <dbl>
1     6 10.8
2    10 10.8
3    19 12.6
4    14 12.5
5    16  8.66
6    11 12.0
```

Y las métricas correspondientes:

```
# A tibble: 3 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       4.05
2 rsq     standard       0.176
3 mae     standard       3.01
```

El R^2 obtenido en el conjunto de testeo fue de aproximadamente 0.18, lo que confirma que, sin considerar `G1` y `G2`, la capacidad predictiva de las demás variables no es muy buena. Este resultado coincide con el comentario de la página original del dataset, que señala que la nota final (`G3`) depende en gran parte de los parciales anteriores.

A continuación haremos un modelo teniendo en cuenta `G1` y `G2`.



Aquí un pantallazo de la tabla de predicción:

```
# A tibble: 6 x 2
     G3 .pred
  <dbl> <dbl>
1     6  6.37
2    10  8.95
```

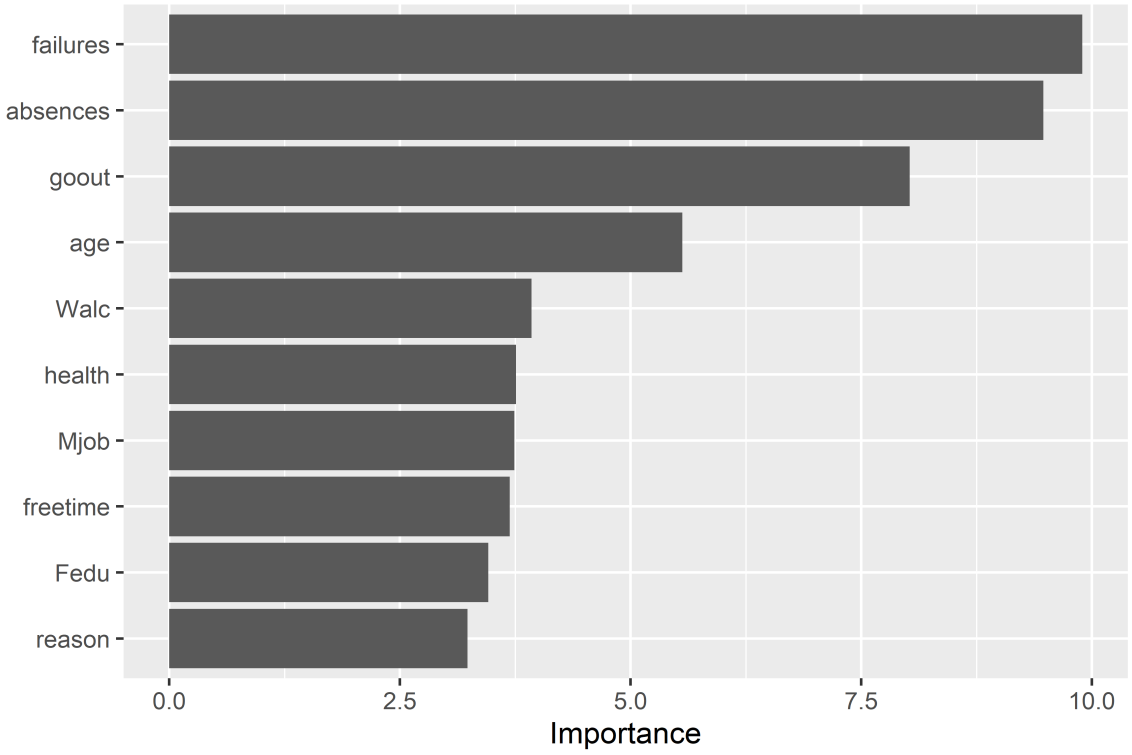
```
3    19 16.0
4    14 13.9
5    16 13.6
6    11 11.5
```

Y las métricas correspondientes:

```
# A tibble: 3 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard       2.12
2 rsq     standard       0.815
3 mae     standard       1.47
```

El incremento del R^2 de 0.176 a 0.815 muestra claramente que `G1` y `G2` concentran casi todo el poder predictivo del modelo. Sin ellas, las variables restantes no logran predecir la nota final (`G3`) de forma adecuada.

Para indagar un poco más, crearemos otro modelo de random forest pero ajustando nuestra variable `G3` de una numerica de 0-20 a una categorica de no aprobado (<10 puntos) y aprobado (>= 10 puntos) y nuestro modelo pasará ahora a un random forest de clasificación.



A continuación un pantallazo de la tabla de predicción:

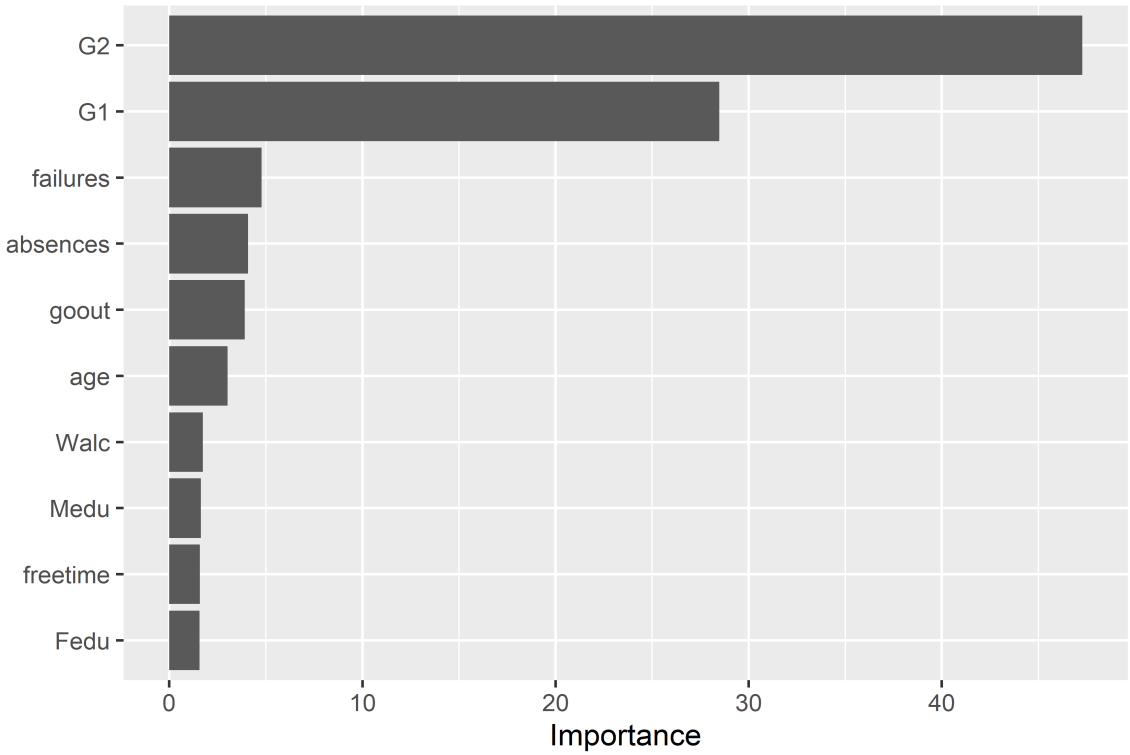
```
# A tibble: 6 x 2
  Estado      .pred_class
  <fct>       <fct>
1 No aprobado Aprobado
2 Aprobado    Aprobado
3 Aprobado    Aprobado
4 No aprobado Aprobado
5 Aprobado    Aprobado
6 Aprobado    Aprobado
```

Y las métricas correspondientes:

```
# A tibble: 2 x 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.734
2 kap      binary         0.260
```

Se observa una gran mejora en el rendimiento del modelo, alcanzando una precisión del 73 %. Al transformar el problema en uno de clasificación binaria (Aprobado/No aprobado), el modelo logra una capacidad predictiva considerablemente mejor (aproximadamente 7 de cada 10 son predecidas correctamente).

Ahora consideraremos tambien a `G1` y `G2` pero manteniendo la variable de respuesta siendo binaria (aprobado/no aprobado).



A continuación un pantallazo de la tabla de predicción:

```
# A tibble: 6 x 2
  Estado      .pred_class
  <fct>       <fct>
1 No aprobado No aprobado
2 Aprobado    Aprobado
3 Aprobado    Aprobado
4 No aprobado Aprobado
5 Aprobado    Aprobado
6 Aprobado    Aprobado
```

Y las métricas correspondientes:

```
# A tibble: 2 x 3
  .metric  .estimator .estimate
  <chr>    <chr>          <dbl>
1 accuracy binary         0.899
2 kap      binary         0.775
```

Vemos que el modelo mejora notablemente una vez más, alcanzando una precisión del 90% y estimando correctamente aproximadamente 9 de cada 10 casos.

Hay algo que se observa de forma consistente en todos los modelos: las variables más importantes para predecir la nota final, sin considerar `G1` y `G2`, son principalmente:

- `failures`: número de cursos reprobados previamente (valores de 0, 1, 2, 3 o 4; donde 4 indica cuatro o más cursos reprobados).
- `absences`: número de inasistencias registradas (de 0 a 93).
- `goout` y `age`: estas variables suelen intercambiarse en el ranking de importancia según el modelo. `goout` indica la frecuencia de salidas con amigos (de 1 = muy baja a 5 = muy alta), mientras que `age` corresponde a la edad del estudiante (de 15 a 22 años).

Aplicación Shiny:

Para complementar el análisis, se diseñó una aplicación Shiny que permite visualizar los principales resultados del análisis exploratorio de forma interactiva. La aplicación organiza la información en tres pestañas: la primera muestra gráficos de barras, la segunda contiene boxplots y la tercera presenta gráficos de violín, todo esto para brindar una perspectiva más detallada de la distribución de las variables.

El código de la app shiny se encuentra en el repositorio como «app» y aquí está el link.

Comentarios finales:

Relación entre calificaciones parciales y finales:
En primer lugar, confirmamos que las calificaciones parciales (`G1` y `G2`) explican gran parte del rendimiento final (`G3`). El desempeño del estudiante al inicio y a mitad de año es un indicador fuerte de la nota final. Esto se refleja en la diferencia de `R2` sin `G1` y `G2`, el modelo apenas alcanza un `R2` de 0.18 mientras que al incluirlas sube a 0.81, lo que demuestra su gran peso explicativo.

Factores de apoyo académico y contexto familiar:
Detectamos que ciertas variables como pagar clases particulares, tener acceso a internet en casa y el nivel educativo de la madre están asociadas a una mayor probabilidad de aprobación. Por ejemplo, los estudiantes que reciben clases extra tienen una tasa de aprobación del 71,8% frente al 63,1% de quienes no lo hacen. De forma similar, quienes cuentan con internet presentan mejores resultados (68,4% frente a 60,6%) y los hijos de madres con educación superior superan claramente a aquellos con madres con nivel primario (74,8% frente a 57,6%).

Historial académico y desempeño:
El número de cursos reprobados previamente resulta ser un fuerte indicador de bajo rendimiento final. A mayor cantidad de asignaturas desaprobadas, menor es la probabilidad de obtener una nota suficiente.

Factores logísticos:
El tiempo de viaje hasta la escuela también muestra cierta relación indirecta: quienes recorren más de una hora tienden a tener más faltas y rendimientos algo más bajos. Los estudiantes con trayectos más cortos suelen tener mejor desempeño y menor ausentismo, probablemente por disponer de más tiempo para estudiar y llegar menos fatigados.

Mejora del modelo al usar clasificación binaria:
Cambiar el enfoque a un modelo de clasificación (aprobado/no aprobado) permitió mejorar la precisión. Sin `G1` y `G2`, la precisión alcanzó un 73%, mientras que al incluirlas se elevó a un 90%, lo que vuelve a reforzar su relevancia.

Posibles lineas de trabajo futuras:

Analizar la calidad del estudio: profundizar en cómo influyen no solo las horas de estudio, sino la forma en que se organiza y la calidad del apoyo recibido.

Incluir variables complementarias: incorporar nuevos factores como características del entorno escolar por ejemplo la calidad docente, infraestructura, recursos. También aspectos psicológicos como motivación, estrés del estudiante.

Realizar estudios más extensos en el tiempo: seguir a los MISMOS estudiantes a lo largo de varios años o en otras varias materias para observar la evolución del rendimiento (en este caso, contábamos con datos de otra asignatura que era lengua portuguesa, pero no era posible identificar a los mismos estudiantes en ambas materias, por lo que no se pudo relacionar la información individual de cada uno entre asignaturas).