UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

INGENIERÍA CIVIL INDUSTRIAL
OPTIMIZACIÓN II

Caso de estudio 2

Fernanda Muñoz, Felipe Rubilar, Ignacio Silva, Jeremías Vásquez, Natalia Velastín

Resumen

El presente informe muestra el análisis de regresión hecho a un grupo de estudiantes que cursan la asignatura de matemáticas, teniendo por objetivo, predecir el rendimiento académico según datos de cada uno de ellos, aplicando diferentes modelos de regresión, que posteriormente fueron combinados entre sí para obtener mejores resultados.

Palabras clave — Regresión lineal, predecir, modelos predictivos, algoritmos iterativos, optimización.

1. Introducción

A lo largo del tiempo, diversos autores se han interesado por conocer qué variables inciden en el rendimiento académico de los estudiantes. Se cree que cierta combinación de factores puede hacer que el estudiante se vea favorecido, o desfavorecido, al momento de enfrentar evaluaciones. A continuación se pretende explicitar los mecanismos utilizados para predecir el rendimiento académico de un grupo de estudiantes en la asignatura de Matemáticas en base a ciertos datos, que van desde su edad hasta si desean seguir estudios posteriores, llámese universitarios. Para poder construir el modelo predictivo del rendimiento académico se utilizó el lenguaje de programación R y el software R Studio.

2. Análisis de datos

2.1. Datos

Dentro de este trabajo se presenta una gran cantidad de variables posibles a utilizar, por lo que, una parte importante del desarrollo y creación del modelo predictivo es analizarlas, para así determinar cuáles son más o menos significativas, y, por tanto, cuáles son utilizadas a la hora de predecir.
En el desarrollo del problema se presentan tres tipos de datos, numéricos, nominales y binarios. Los primeros son representados por números en un cierto rango, como por ejemplo entre 1 y 5 o 1 y 20, por su parte los segundos son distribuidos en distintos grupos etiquetados, que no presentan valores o una jerarquía entre ellos, por ejemplo, al describir el trabajo del padre de un alumno las opciones son ser médico, profesor o un trabajo administrativo, lo cual es una variable nominal, al no presentar una escala entre las diferentes opciones. Por último, están los binarios, que presentan dos valores posibles
Las variables observadas son:

- Numéricas:

    1. Age: edad del estudiante, de 15 a 22.
    2. Medu: educación de la madre, de 0 a 4, menor a mayor.
    3. Fedu: educación del padre, de 0 a 4, menor a mayor.
    4. Traveltime: tiempo de viaje de casa al colegio, de 0 a 4, menor a mayor.
    5. Studytime: tiempo de estudio semanal, de 0 a 4, menor a mayor.
    6. Failures: número de fracasos en clases pasadas.
    7. Famrel: calidad de la relación familiar, de 1 a 5, mala a buena.
    8. Freetime: tiempo libre después del colegio, de 1 a 5.

9. Goout: salir con amigos, de 1 a 5.
10. Dalc: consumo de alcohol en la semana, de 1 a 5.
11. Walc: consumo de alcohol el fin de semana, de 1 a 5.
12. Health: salud actual, de 1 a 5, mala a buena.
13. Absences: ausencias del colegio, de 0 a 93.
14. G1: nota primer semestre, de 0 a 20.
15. G2: nota segundo semestre, de 0 a 20.
16. G3: nota final, de 0 a 20.

- Nominales:

1. Mjob: trabajo madre.
2. Fjob: trabajo padre.
3. Reason: razón para escoger este colegio.
4. Guardian: persona responsable del estudiante.

- Binarias:

1. Sex: sexo del estudiante.
2. Address: tipo de dirección de casa de estudiante.
3. Famsize: tamaño de familia.
4. Pstatus: padres viven juntos o separados.
5. Schoolsup: apoyo educacional extra.
6. Famsup: apoyo educacional de la familia.
7. Paid: clases privadas extras.
8. Activities: actividades extra-curriculares.
9. Nursery: estudiante fue a sala cuna.
10. Higher: quiere estudiar en educación superior.
11. Internet: acceso a internet en casa.
12. Romantic: en una relación romántica

2.2. Correlación

Con un entendimiento más completo de los datos que se manejan y sus posibles características, se procede al siguiente paso para su análisis, que es observar la correlación entre las variables numéricas, y, a partir de esto, sacar conclusiones correspondientes sobre el comportamiento de unas con otras. Gracias a esta correlación se puede ver de qué forma se asocian las variables con G3 (nuestro objetivo a predecir), ya sea de manera positiva o negativa, y también se puede estudiar el comportamiento entre las variables explicativas para rescatar información relevante.

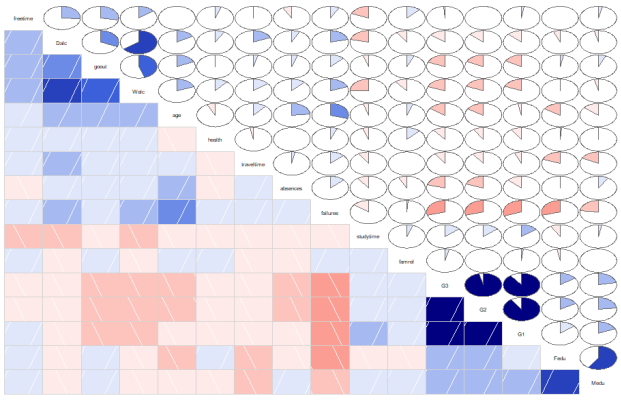

Figura 1: Gráfico de correlaciones

Primero que todo, analizando G3, se puede destacar que posee la correlación positiva más alta con G1 y G2, lo cual resulta bastante intuitivo ya que estas corresponden a las notas de los semestres de forma individual. Luego, las siguientes correlaciones positivas más altas con G3 vendrían a ser con Fedu y Medu. Por último, presenta correlación negativa con failures, absences, age, Walc y Goout.
A partir de esto, se puede estimar que, a primera vista, quedaría un modelo lineal del siguiente estilo:
$G3 = Y = \beta_0 + \beta_1 G1 + \beta_2 G2 + \beta_3 fedu + \beta_4 medu + \beta_5 failures + \beta_6 absences + \beta_7 age + \beta_8 Walc + \beta_9 goout + \xi$

Más información que se puede rescatar de este análisis se presenta a continuación:

- Se observa una correlación positiva muy alta entre Fedu y Medu.

- Existe correlación positiva entre studytime y G1, la cual es mayor que con G2 y G3.
- Correlación positiva ligera entre age, failures y absences.
- Correlación negativa entre studytime y Walc, Dalc,Freetime.
- Correlación positiva alta entre Walc, goout y Dalc.

2.3. Análisis Gráficos de Caja

Terminado el análisis de correlación, se crearon gráficos de caja con las variables numéricas, para analizar detenidamente los datos, donde se evaluaron tanto los rangos como los máximos, mínimos y los posibles outliers. El resultado que se obtuvo se muestra a continuación:

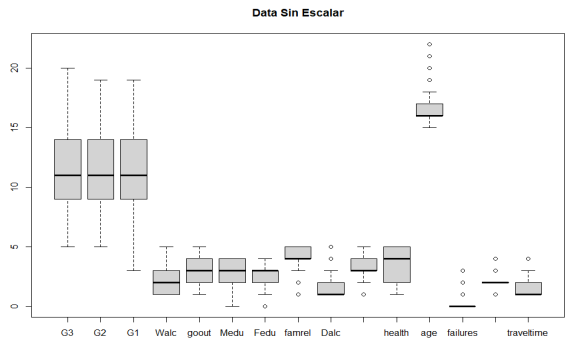

Figura 2: Data sin escalar

Cabe destacar que no se utilizaron las ausencias (absences) debido a que presentaban una cantidad muy grande de outliers, lo cual podría afectar negativamente a la visualización del resto de gráficos y a la interpretación de este mismo. A primera vista se puede ver que los gráficos de caja presentan rangos muy diferentes entre sí, por lo que se determina que se necesita escalar los datos, para que así se puedan analizar los gráficos y su información de mejor manera. Esto resulta en un nuevo conjunto de gráficos, que se pueden ver en la figura 2.

Los datos ya escalados se presentan de mejor manera respecto a la imagen anterior, y sirven para crear una idea de su comportamiento y la dispersión de sus valores. Al terminar el análisis de las variables numéricas, se estudian los factores entregados (variables binarias y nominales). En el caso de estos valores se realizaron análisis de gráficos de caja individuales respecto a G3, para así poder determinar si había algún tipo de tendencia notable dependiendo del valor que toman, y llegar a la conclusión de si estos se comportan de forma significativa. Algunos ejemplos de los resultados obtenidos fueron los siguientes:

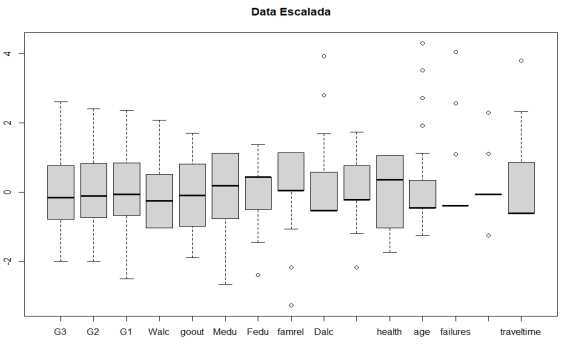

Figura 3: Data escalada

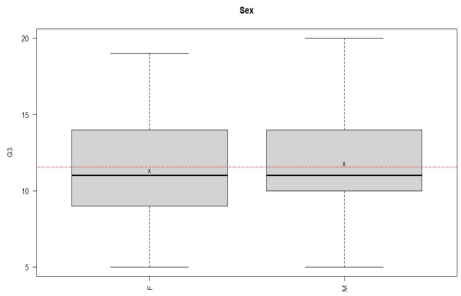

Figura 4: Boxplot sex

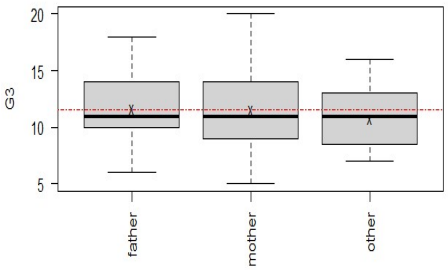

Figura 5: Boxplot guardian

Dentro de estos resultados se puede ver que, al menos de forma independiente, los factores no logran alterar considerablemente el resultado de G3, debido a que, sin importar la opción que tomen, estos se comportan de

forma relativamente similar, por lo que se puede concluir que no son significativos para la predicción que se quiere obtener. Ahora, vale la pena destacar, que esto solamente toma en cuenta su comportamiento de forma independiente, ya que aún puede existir la posibilidad de que estos factores en interacción sean significativos.

2.4. Selección de variables

Antes de verificar cómo se comportan distintas combinaciones de variables para encontrar las predicciones más adecuadas respecto a la variable G3, es necesario dejar constancia la forma en que se trabaja en la siguiente sección de este informe.

- Se separaron las variables entregadas en entrenamiento y testeo en proporción 70/30, con el fin de obtener una predicción generalizada y realística.
- La métrica usada para analizar la efectividad del modelo fue RMSE (Root Mean Squared Error, por sus siglas en inglés) Calculada en base a la salida del testeo y la predicción del modelo o algoritmo respectivo (OLS, Elastic-Net, Random Forest).
- El fin fue encontrar un modelo que redujese al máximo RMSE con los datos de testeo, sin considerar en demasía el que se pudiese obtener con los datos de entrenamiento.
- Los datos fueron escalados para no tener problemas de dimensiones y también se extrajeron los ceros de la variable de salida para no entorpecer las predicciones.
- Cuando se encuentre el grupo de variables adecuadas, se ensamblarán los modelos mencionados con el fin de que al unirlos se potencien las mejores partes de cada uno.
- Si este modelo final mejora respecto a los anteriores, se entrenan todos los datos para así entregar la predicción final.

Habiendo especificado las bases de este análisis, se procede a presentar el desarrollo.

2.4.1. Análisis 1: Variables numéricas

En este primer análisis se usaron solamente las variables numéricas (16) para crear un modelo de regresión lineal y estos fueron los resultados.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.893e-16  1.715e-02   0.000 1.000000
G2           8.241e-01  4.397e-02  18.743  < 2e-16 ***
G1           1.534e-01  4.343e-02   3.532 0.000522 ***
Walc         3.081e-03  2.564e-02   0.120 0.904480
goout       -1.875e-02  2.138e-02  -0.877 0.381734
Medu        -1.799e-03  2.252e-02  -0.080 0.936431
Fedu         3.307e-03  2.237e-02   0.148 0.882634
famrel       5.482e-02  1.841e-02   2.977 0.003297 **
Dalc        -1.481e-03  2.354e-02  -0.063 0.949896
freetime    -2.536e-02  1.931e-02  -1.313 0.190650
health      -3.509e-02  1.777e-02  -1.975 0.049736 *
age          1.349e-02  2.044e-02   0.660 0.510027
failures     2.223e-02  1.995e-02   1.115 0.266448
absences    -2.088e-02  1.855e-02  -1.126 0.261617
studytime    4.299e-03  1.855e-02   0.232 0.817016
traveltime  -2.231e-02  1.821e-02  -1.225 0.222122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2432 on 185 degrees of freedom
Multiple R-squared:  0.9453,    Adjusted R-squared:  0.9409
F-statistic: 213.2 on 15 and 185 DF,  p-value: < 2.2e-16
```

Figura 6: Resumen Modelo Lineal

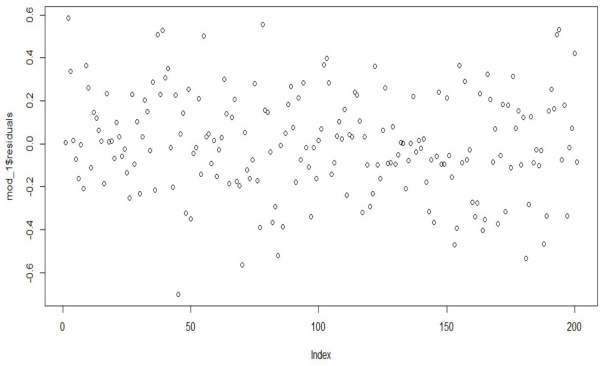

Figura 7: Residuos Modelo Lineal

De este modelo se puede desprender:

- Tiene un potencial predictivo pues

$$R^2 = 0,9409$$

- El modelo es significativo globalmente ya que

$$P - valorglobal = 2{,}2 * 10^{-16}$$

- Solo G1 y G2 son variables significativas.
- Hay tres variables que tienen una significancia leve (famrel, freetime, absences)
- Los residuos tienen comportamiento homocedástico, validado por el test de Breusch-Pagan donde se obtuvo un p-valor de 0,41, lo cual no es evidencia suficiente para rechazar dicha hipótesis nula.

- Los residuos se comportan de manera normal, lo que está validado por el test de Shapiro-Wilk donde se obtuvo un valor de 0,61 que no entrega evidencia suficiente para rechazar la hipótesis nula.
- También se observó que no existe correlación entre los residuos, basándose en los test de Durbin-Watson y Breusch-Godfrey.

Pese al buen comportamiento del modelo respecto a los supuestos estadísticos, este no se puede usar directamente debido a que no todas las variables son significativas. Por lo cual, se intenta llegar a un modelo lineal donde todas lo sean.

Independiente de ello, como se explicó anteriormente, se presentan los resultados de todos los modelos debido a que, como los restantes son algoritmos, no necesitan cumplir los supuestos estadísticos.

Comparación de errores:

	RMSE	OLS	Elastic	Forest
1	Train	0.2332789	0.2432261	0.2810490
2	Test	0.9033030	0.8595866	0.9634865
3	Error %	6.9754753	6.7029333	7.4262170

Figura 8: Comparación de errores análisis 1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.202e-16  1.709e-02   0.000 1.000000
G2          8.293e-01  3.963e-02  20.925  < 2e-16 ***
G1          1.516e-01  3.964e-02   3.825 0.000176 ***
famrel      4.674e-02  1.714e-02   2.726 0.006980 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2423 on 197 degrees of freedom
Multiple R-squared:  0.9422,    Adjusted R-squared:  0.9413
F-statistic:  1070 on 3 and 197 DF,  p-value: < 2.2e-16
```

Figura 9: Resumen Modelo Lineal 2

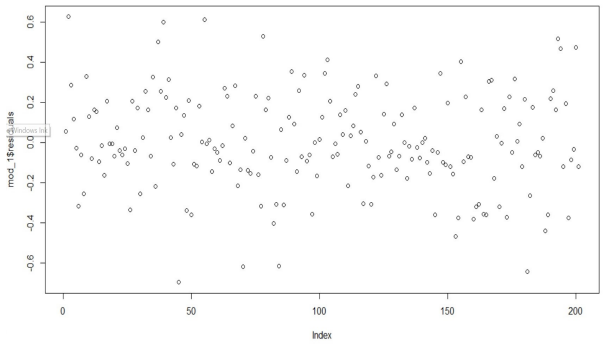

Figura 10: Residuos Modelo Lineal 2

2.4.2. Análisis 2: Variables numéricas significativas

En este análisis se extrajo una por una cada variable que no era significativa, quedando un modelo de esta forma.

De este modelo se puede desprender que:

- Al igual que el anterior, este modelo es significativo de forma global y presenta un R2 levemente mayor al anterior.
- En el test de Shapiro-Wilk se obtiene un p-valor de 0,08, el cual, por una leve diferencia, no entrega información suficiente para rechazar la hipótesis nula.
- Sin embargo, a pesar de que los residuos parecen ser homocedásticos respecto a su gráfica, el test de Breusch-Pagan entrega un p-valor de 0.03 que está bajo el umbral, por lo que se rechaza la hipótesis nula y los residuos no tienen comportamiento homocedástico.
- Respecto a los test de autocorrelación, estos valores se acercan más al umbral respecto a los calculados anteriormente, pero aún así no entregan información suficiente para rechazar la hipótesis nula.

A pesar de que todas las variables elegidas son significativas en este modelo lineal, no cumple el supuesto de homocedasticidad, por lo que tampoco sería correcto darle uso a este modelo directamente. Independiente de ello, como los otros modelos no tienen problemas con los supuestos, se muestra la tabla de comparación de errores.

	RMSE	OLS	Elastic	Forest
1	Train	0.2399037	0.2400048	0.2743016
2	Test	0.8713185	0.8707856	1.0053663
3	Error %	6.6497447	6.6499432	7.4243745

Figura 11: Comparación de errores análisis 2

2.4.3. Análisis 3: Elección arbitraria

En este análisis se eligieron las variables de forma arbitraria basándose en suposiciones generales sobre que podría tener incidencia en la nota final de una asignatura.

A continuación se presentan las variables elegidas y la razón de su elección.

- G1 y G2 : Debido a su alta correlación con la variable de salida ya que son notas de pruebas rendidas anteriormente durante el período de estudio.
- Studytime: Generalmente las personas que estudian más tiempo obtienen mejores calificaciones.
- Absences: Se elige bajo el supuesto de que las personas que van más a clases podrían tener acceso a información sobre qué sería importante estudiar para la prueba.
- Failures: Suele ocurrir que el mal rendimiento es una tendencia.
- Age: Elegida debido a que las personas de mayor edad podrían tener más experiencia para enfrentarse a estas situaciones.
- Traveltime: Esto se debe a que las personas que tardan más tiempo en llegar a su lugar de estudio tendrían por consecuencia menos tiempo para estudiar, sumado al cansancio que conlleva desplazarse en distintos transportes.

En lo que resta del informe no se hará especial énfasis en conclusiones respecto a los supuestos estadísticos del modelo lineal, ya que se está en conocimiento de lo difícil que es encontrar un modelo que cumpla al pie de la letra todos los supuestos. Esto se pudo observar en los análisis realizados anteriormente donde luego de lograr que todas las variables fuesen significativas, dejó de ser válido el supuesto de homocedasticidad.

Por lo tanto, se presentan los resultados de la predicción de los distintos modelos.

	RMSE	OLS	Elastic	Forest
1	Train	0.2420515	0.2455819	0.2846697
2	Test	0.8766246	0.8529097	0.9973476
3	Error %	6.6940360	6.6310951	7.5090193

Figura 12: Comparación de errores análisis 3

Se ve que hay una leve mejora en el RMSE test del modelo aplicando Elastic-net.

2.4.4. Análisis 4: Elección arbitraria con interacciones

A las variables elegidas anteriormente se les suma las interacciones entre G1 y G2, studytime y traveltime, y por último failures con absences, también basado en supuestos. Los resultados obtenidos son los siguientes: No hay mucho cambio respecto a los anteriores.

	RMSE	OLS	Elastic	Forest
1	Train	0.2393831	0.2464809	0.2865659
2	Test	0.8913502	0.8529097	1.0254089
3	Error %	6.7916539	6.6310951	7.6618719

Figura 13: Comparación de errores análisis 4

2.4.5. Análisis 5: Variables Categóricas

Para hacer este análisis se toman en consideración los factores excluidos anteriormente. Al ser variables categóricas se les tiene que hacer una transformación a variable dummy.

Por temas de extensión, no se hará mayor énfasis en la transformación de los datos para realizar los modelos por lo que se muestran los resultados finales donde no se observa una mejora respecto a los análisis hechos anteriormente.

	RMSE	OLS	Elastic	Forest
1	Train	0.2147333	0.2483449	0.2893552
2	Test	0.9775728	0.8662617	0.9966778
3	Error %	7.1924085	6.7116086	7.7760053

Figura 14: Comparación de errores análisis 5

2.4.6. Análisis 6: Variables Categóricas Arbitrarias

Se hace un filtro a las variables categóricas existentes en el modelamiento basado en cuáles podrían realmente tener una significancia en la nota final.

Las variables elegidas son:

- Father's Job: Debido a que en el análisis de los gráficos de caja realizados anteriormente se pudo observar que existían diferencias en la calificación final respecto a los distintos trabajos que tenían los padres.
- Mother's Job: Esta variable tenía un comportamiento similar a la anterior pero con diferencias respecto a cuáles eran los grupos con notas mayores.
- Internet: Suele darse que el acceso a internet puede tener influencia en las calificaciones finales.
- Higher: Esta variable hace referencia a los deseos de los estudiantes de tener estudios superiores, por lo tanto podría considerarse relevante a la hora de obtener mejores resultados.
- Romantic: Al analizar el gráfico de caja de esta variable respecto a la nota final, se observaron diferencias entre los rangos de ambas opciones.

Los resultados fueron los siguientes:

	RMSE	OLS	Elastic	Forest
1	Train	0.2328334	0.2470472	0.2833615
2	Test	0.9363863	0.8541250	1.0063491
3	Error %	7.0842014	6.6383999	7.5727548

Figura 15: Comparación de errores análisis 6

En base a los resultados obtenidos de los distintos análisis, se procede a elegir un grupo de variables para estudiar más en detalle.

2.5. Elección

De todos los análisis se considera que el tercero es el que podría predecir de forma más adecuada, ya que las predicciones hechas con el grupo de variables elegidas de forma arbitraria tenían valores de RMSE más pequeños y homogéneos respecto a los otros análisis.
Se procede a ensamblar las tres predicciones en un solo modelo para presentar los resultados en la siguiente figura.

	RMSE	OLS	Elastic	Forest	Ensemble
1	Train	0.2420515	0.2455819	0.2846697	0.2420515
2	Test	0.8766246	0.8529097	0.9973476	0.8737910
3	Error %	6.6940360	6.6310951	7.5090193	6.3667541

Figura 16: Comparación de errores con ensamblaje

También se puede observar la correlación entre la variable de salida y las predicciones respecto a los distintos modelos o algoritmos en la figura 17.

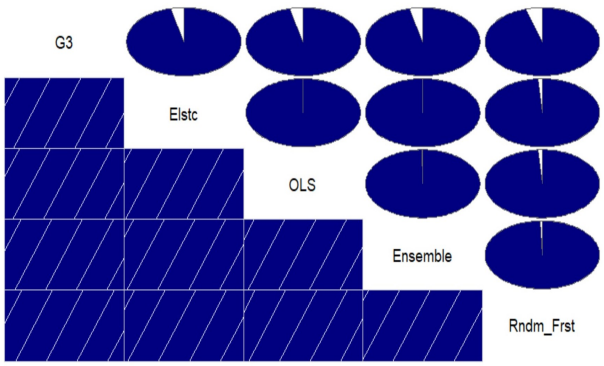

Figura 17: Correlación entre modelos

Por último, se presenta un gráfico que enfrenta las predicciones v/s el valor real del testeo en la figura 18.

Como la elección del modelo ya está hecha, el último paso es modificar un poco el código para entrenar el

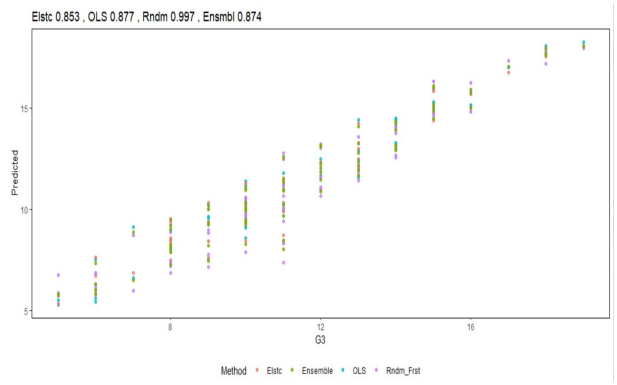

Figura 18: Gráfico

total de los datos con las variables escogidas para luego predecirlas con los datos de testeo entregados en el enunciado.

3. Conclusión

Debido al número de experimentaciones que se realizaron se seleccionó el tercer análisis, ya que este presentaba el menor error en comparación con otros estudios realizados, por lo tanto, no son utilizadas interacciones entre las variables ni los factores, ya que en los análisis no mostraron una mejora en la predicción. Un hallazgo interesante se puede observar al comparar el análisis número 3 y 4, que evidencia que el método elastic-net no sufre alteraciones al agregarle las interacciones. En base a todo esto se concluye que las variables numéricas elegidas de forma arbitraria basadas en supuestos racionales fueron las que presentaron mejores resultados, los cuales se ven potenciados con la implementación de la herramienta ensemble learning, la cual al reunir los diferentes algoritmos de aprendizaje produce un rendimiento predictivo mejor.

Por último, si se observa desde una perspectiva general e intuitiva, el modelo perfecto sería ponderar directamente las notas obtenidas en los períodos anteriores para obtener una predicción de la tercera calificación, pero gracias a la experimentación se logró concluir que los resultados son mejores al agregar más variables al estudio, ya que gracias a éstas, se puede obtener un modelo más generalizado y con una variabilidad mayor, teniendo siempre en consideración que es imposible encontrar un modelo con 100 % de efectividad.

4. Bibliografía

- Alvear, J. O. (2018). Arboles de decision y Random Forest.
- Amsantac. (2016). Mejorando la exactitud en la clasificación mediante ensamble de modelos. Bogotá.