Informe del 2do Proyecto de Estadística - Equipo 6

Lázaro Raúl Iglesias Vera L.IGLESIAS@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Miguel Tenorio Potrony M.TENORIO@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Daniel Enrique Cordovés Borroto D.CORDOVESB@ESTUDIANTES.MATCOM.UH.CU
Grupo C411

Resumen

Aplicar análisis estadístico a un set de datos referentes al rendimiento de estudiantes de una asignatura determinada utilizando técnicas avanzadas de Estadística Inferencial.

Abstract

Apply an statistical analysis to a set of data referent to the performance of students of a subject using advanced tecniques of Inferential Statistics.

Palabras Clave: Correlación, Regresión, Variables.
Tema: Análisis Estadístico, Estádistica Inferencial.

1. Introducción

En este trabajo se planea realizar un análisis estadístico del desarrollo de un grupo de estudiantes en cierta asignatura, utilizando técnicas de regresión, ANOVA y reducción de dimensión; esta última incluye: análisis de componentes principales, clúster y árboles de decisión.

2. Ejercicios

Se utilizarán las técnicas en el orden que están propuestas en la orientación [1].
Tanto en 2.1, como en 2.2 y 2.3 se utiliza y procesa el set de datos orientado para el equipo (students-data.csv). Este contiene información para medir el rendimiento de estudiantes de educación secundaria en dos escuelas portuguesas. Contiene variables como los resultados en diferentes períodos, estado familiar y características económicas de cada estudiante. Todas estas variables están descritas en [2].

2.1 Regresión

2.1.1 Análisis de Regresión

En primer lugar hallamos la correlación entre todos los pares de variables en el data frame. Podemos apreciar que `G1.x` y `G2.x` están ambos correlacionados con `G3.x`, lo cual tiene sentido, dado que los primeros son notas de ambos semestres y el último es la nota del curso.
Luego separamos variables cualitativas de cuantitativas, para aplicarle a estas últimas una estandarización usando la función `scale`.

A continuación se creó el modelo teniendo como variable dependiente a $G3.x$ y usando al resto como variables dependientes, para luego aplicar un algoritmo de model selection para nuestros datos, en este caso usamos la función `MASS::stepAIC`, que se encarga de eliminar las variables independientes no necesarias.

El modelo obtenido es el siguiente:

```
lm(formula = G3.x   age + famrel.x +
absences.x + G1.x + G2.x, data = df)
```

Coefficients:	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-6.468e-17	2.111e-02	0.000	1.000000
age	-5.286e-02	2.169e-02	-2.437	0.015267 *
famrel.x	7.994e-02	2.124e-02	3.764	0.000194 ***
absences.x	7.986e-02	2.137e-02	3.738	0.000215 ***
G1.x	1.093e-01	4.172e-02	2.620	0.009152 **
G2.x	8.039e-01	4.204e-02	19.125	< 2e-16 ***

Y obtenemos estos datos del modelo:

```
Multiple R-squared:  0.8321
Adjusted R-squared:  0.8298
F-statistic:  372.6 on 5 and 376 DF
p-value:  < 2.2e-16
```

El parámetro `Adjusted R-squared` es 0.82 lo cual es bueno, es cercano a 1. El `p-valor` del estadígrafo de F es menor que 0.05 por lo que existe una variable significativamente distinta de 0 en el modelo.

©2020 Lázaro Raúl Iglesias Vera, Miguel Tenorio Potrony, Daniel Enrique Cordovés Borroto.

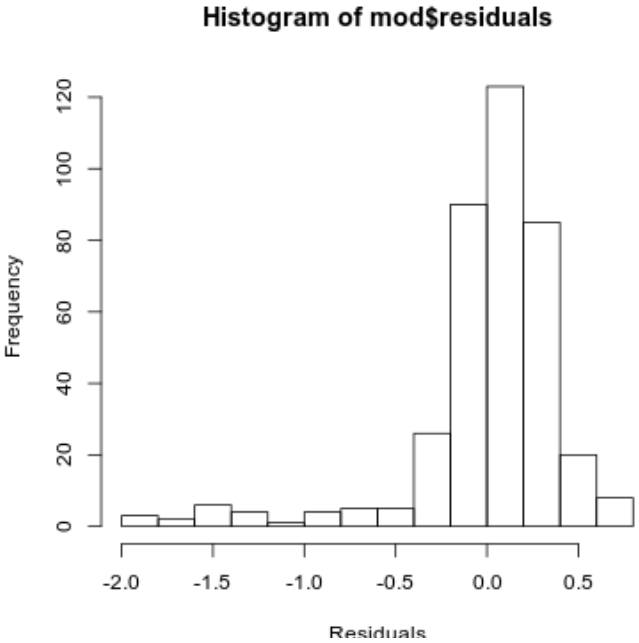
2.1.2 Análisis de Residuos

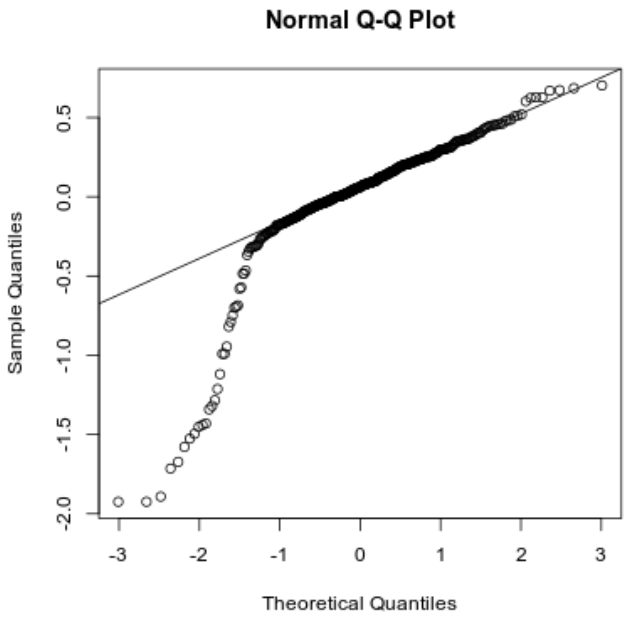
1. Media de errores:

```
Media de error residual -3.157526e-18
Suma de error residual  -1.20997e-15
```

   Por lo que se cumple que ambas son muy cercanas a 0.

2. Podemos ver el histograma de residuos y el gráfico QQ-Plot para asegurar que los errores están distribuidos normal:





3. Independencia de los residuos:

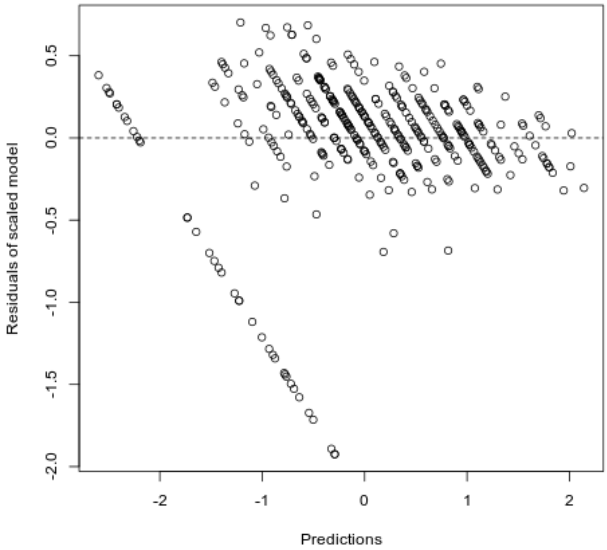
   Al realizar el test de Durbin-Watson obtenemos:

```
DW = 2.0634, p-value = 0.7137
```

   Como 0.7137 >> 0.05 no podemos rechazar la hipótesis nula por lo que los errores son independientes.

4. Homocedasticidad

   Se realiza el gráfico de predicciones contra errores



   Como se puede apreciar, en la mayor parte de la imagen los valores son aleatorios por lo que se cumple la Homocedasticidad.

2.2 ANOVA

Utilizando el set de datos students-data.csv se quiere conocer si la dirección de un estudiante, ya sea en área rural o urbana, tiene alguna relación con la nota promedio obtenida por el mismo en las 3 evaluaciones realizadas.

Para ello se seleccionaron los datos y se colocaron en una tabla de la siguiente forma:

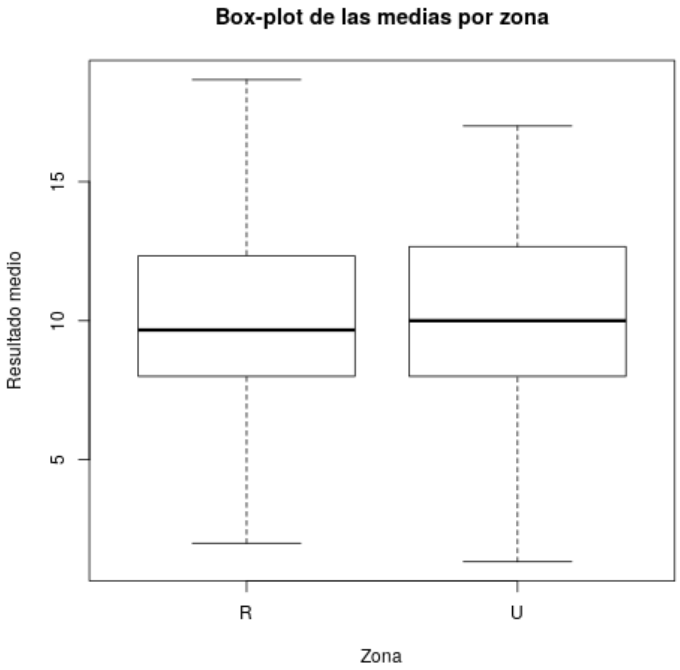
Dirección	Promedio
R	15.6
R	12.9
R	9.0
R	10.7
...	...
U	11.2
U	14
U	8.9
U	19
...	...

Los datos anteriores no son reales, solamente muestran la estructura utilizada.

Una vez agrupados los datos de esta forma se procede a hacer el análisis **ANOVA** para dar respuesta a la siguiente interrogante: **¿Existen diferencias en la nota promedio de los estudiantes de diferentes**

zonas?

Un primer acercamiento a la pregunta en cuestión lo brinda el análisis de las medias de factor.



Como se puede apreciar, no hay una diferencia notable dentro de los datos procesados, por lo que es posible que no se pueda rechazar la hipótesis (H_0) de que no existe diferencia.

> **NOTA:** Para los análisis posteriores se fijará un nivel de significación $\alpha = 0.05$

Los resultados del análisis de varianza de **ANOVA** indican exactamente lo visto anteriormente.

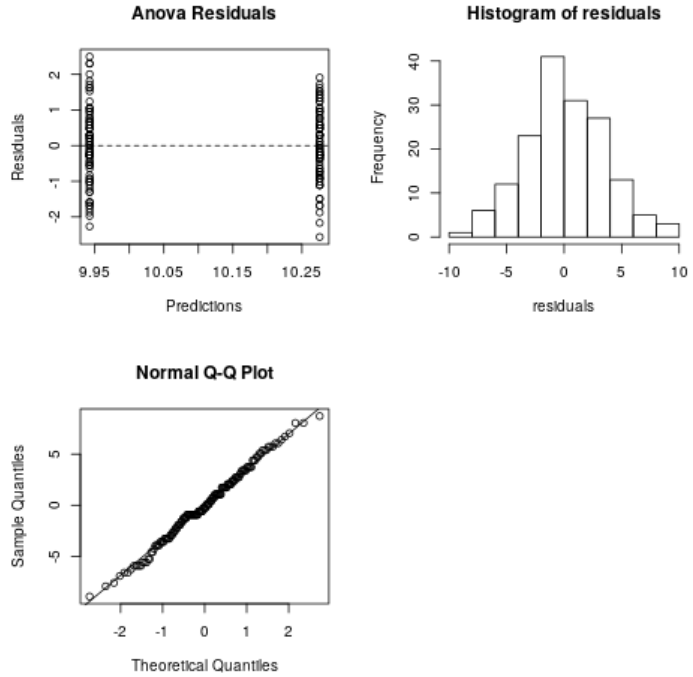
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dirección	1	4.5	4.50	0.353	0.553
Residuals	160	2040.6	12.75	-	-

Como se puede observar el `p-value` resultó ser mayor que la significación α prefijada, por tanto se puede decir que no varía la nota promedio de los estudiantes según su dirección.

Para poder concluir con certeza es necesario realizar una verificación del cumplimiento de los 3 supuestos de este modelo. Para ello se utilizarán los residuos obtenidos en el paso anterior. Los supuestos antes mencionados son:

1. Los residuos siguen una distribución normal con media cero.
2. Los residuos son independientes entre sí.
3. Los residuos de cada tratamiento tienen la misma varianza σ^2.

Tal como en el procedimiento anterior, se puede tener una primera idea sobre el comportamiento de los supuestos si se analiza su comportamiento gráficamente como se muestra a continuación.



Este grupo de gráficos, indican que el test es válido ya que cada uno de ellos tiene el comportamiento esperado. Dado que esta información no es 100% verídica, se procederá a realizar los test para comprobar los supuestos utilizando los residuos obtenidos del análisis de varianza.

```
Shapiro-Wilk normality test:
W = 0.99342, p-value = 0.6761

Bartlett test of homogeneity of variances:
Bartlett's K-squared = 0.75208, df = 1,
p-value = 0.3858

Durbin-Watson test:
DW = 2.0016, p-value = 0.4726
alternative hypothesis:  true autocorrelation is
greater than 0
```

Los resultados de los test previos confirman la veracidad de el análisis de varianza realizado, dado que todos los supuestos se cumplen, hecho notable en cada uno de los test pues nuevamente todos los `p-value` obtenidos son mayores que la significación α.

2.3 Reducción de Dimensión

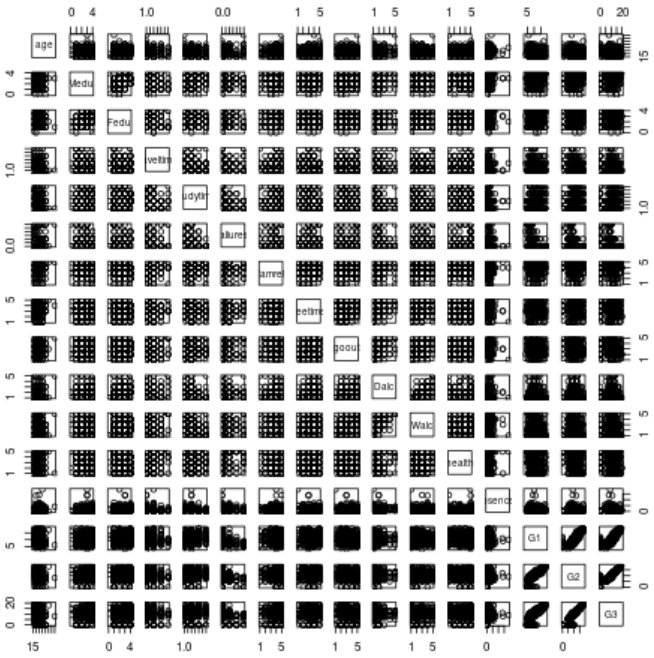
2.3.1 ACP

A partir del data set students-data.csv fueron extraídas variables numéricas de interés sobre los estudiantes para así realizar estudios sobre ellos. Estos datos se encuentran recogidos en sub-students-data.csv.

Dada la gran cantidad de datos se dificultan los estudios que se desean hacer por ello en este trabajo se

hará un análisis de las componentes principales(**ACP**) para así intentar facilitar futuros procedimientos.

De manera inicial es útil conocer la correlación existente entre las variables. Esta información puede ser obtenida analizándola gráfica y numéricamente. De la gráfica es difícil obtener información dada la cantidad de variables y el tamaño de la muestra, pero puede ser apreciada a continuación.



De igual modo extraer datos de la matriz de correlación se hace complicado por las mismas razones anteriores. Para ver dicha matriz puede ejecutar el siguiente comando desde la raíz del proyecto:

```
make reduct
```

A continuación se muestra la forma reducida de la matriz para una sencilla interpretación de la correlación de las variables utilizadas.

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1															
Medu		1														
Fedu		,	1													
traveltime				1												
studytime					1											
failures						1										
famrel							1									
freetime								1								
goout									1							
Dalc										1						
Walc									.	,	1					
health												1				
absences													1			
G1						.								1		
G2						.								+	1	
G3						.								+	*	1

Leyenda:

símbolo		.	,	+	*	B	1
significación	0	0.3	0.6	0.8	0.9	0.95	1

Como se muestra, se está en presencia de datos que no son altamente correlacionados. Dado lo anterior las variables son independientes y podemos proceder a realizar el análisis **ACP** para lograr un reducción de la dimensión.

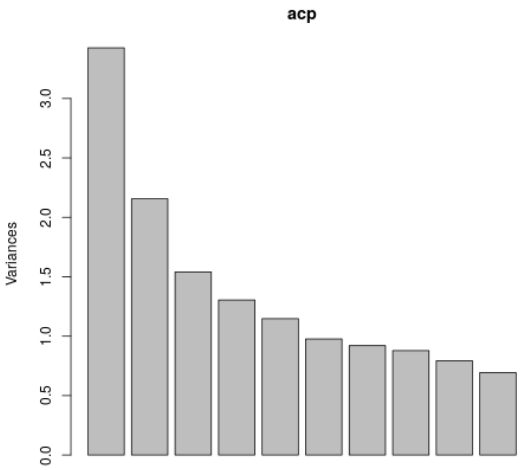
Como resultado se obtiene la importancia de las componentes:

> Solo se mostrarán las 5 primeras componentes, para verlas todas ver el log del comando antes mencionado

	PC1	PC2	PC3	PC4	PC5
SD	1.8509	1.4684	1.2411	1.1422	1.0711
PV	0.2141	0.1348	0.0962	0.0815	0.0717
CP	0.2141	0.3489	0.4451	0.5266	0.5984

Donde SV representa *Standard deviation*, PV *Proportion of Variance* y CP *Cumulative Proportion*.

Dado los valores principales de las componentes y utilizando el criterio de **Kaiser** podemos tomar las primeras 5 componentes como las principales de los datos muestrados, lo cual implica una reducción considerable de los mismos. Esta selección puede ser reforzada gráficamente analizando el siguiente gráfico.



Solo resta sacar resultados de los datos obtenidos; para ello es necesario conocer la matriz de valores propios.

	PC1	PC2	PC3	PC4	PC5
age	0.17169428	-0.096375675	0.122900815	-0.116727605	-0.52230504
Medu	-0.23077461	-0.230813249	-0.543715147	-0.088915536	-0.01232128
Fedu	-0.21676166	-0.204272191	-0.559993441	-0.076942344	0.08553494
traveltime	0.16002591	-0.038139364	0.319148379	-0.084267084	0.08041764
studytime	-0.14784513	0.207102688	-0.034911390	-0.022401906	-0.47054192
failures	0.32507813	-0.006531465	0.068332064	0.049664978	0.11642778
famrel	-0.03352910	0.029605126	-0.064101343	0.586021189	-0.24574182
freetime	0.05203081	-0.263597314	0.001685637	0.563524636	-0.05156450
goout	0.14935467	-0.376609471	-0.075532835	0.221715560	-0.35687080
Dalc	0.14553046	-0.498027111	0.118927504	-0.091027971	0.09965108
Walc	0.16657207	-0.527056881	0.138340828	-0.124483003	0.06177011
health	0.06370752	-0.034531007	-0.056065007	0.301857324	0.50209746
absences	0.04236162	-0.219023277	-0.062977661	-0.366276132	-0.14114199
G1	-0.45823864	-0.128501492	0.246108714	0.035532688	-0.01979275
G2	-0.47075051	-0.140318134	0.277523544	-0.002009708	0.02699303
G3	-0.45236139	-0.167155590	0.283082240	0.044897256	0.03167643

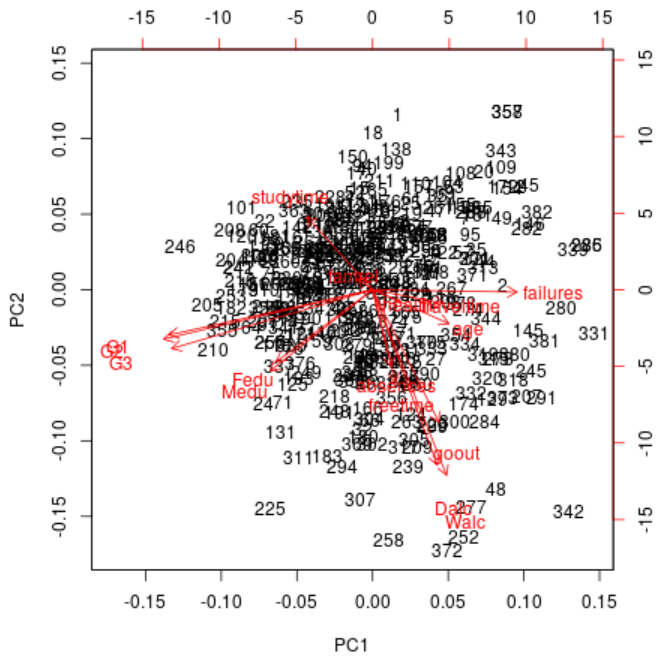
NOTA: Solo se muestran las componentes seleccionadas dado que son las que aportan datos a la investigación

Por cada componente es necesario obtener su mayor valor propio para luego seleccionar las variables de la componente.

- **PC1:** El mayor valor propio es el asociado a los `failures` por tanto es una componente marcada por las fallas anteriores de los estudiantes, su edad, la educación del padre y la madre, sus notas y el consumo de alcohol de los mismos en sus fines de semana.
- **PC2:** El mayor valor propio es el asociado al `studytime` por tanto es una componente marcada por tiempo que dedican los estudiantes a su estudio individual, la educación del padre y la madre, su tiempo libre y frecuencia de salidas, su consumo de alcohol, ausencias a clase y las notas obtenidas.
- **PC3:** El mayor valor propio es el asociado al `traveltime` por tanto es una componente marcada por las notas obtenidas por el estudiante, su tiempo de viaje hacia la escuela y la educación de sus padres.
- **PC4:** El mayor valor propio es el asociado al `famrel` por tanto es una componente marcada por las relaciones familiares del estudiante, su tiempo libre, ausencias a clase y su estado de salud.
- **PC5:** El mayor valor propio es el asociado al `health` por tanto es una componente marcada únicamente por el estado de salud del estudiante.

A través de estos datos podemos apreciar cómo no solamente las notas son factores de importancia en el análisis de su desempeño o cualquier otra valoración que se desee hacer respecto a los estudiantes, dado que tanto su estado de salud, relaciones familiares, la planificación de su tiempo y la educación de su padres influyen en sus indicadores principales.

Por último, podemos ver el biplot resultante de combinar las componentes, para observar la influencia de las variables en las componentes gráficamente. A continuación se muestra el biplot asociado a las 1ra y 2da componentes.

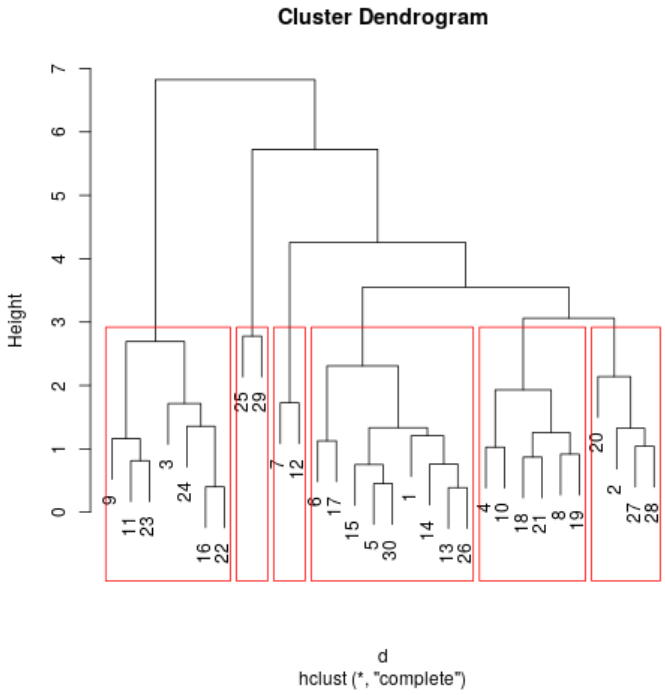


2.3.2 Cluster

De los estudiantes presentes en el set de datos students-data.csv se conoce que 30 de ellos están interesados en mejorar su rendimiento académico. Para satisfacer su interés, el profesorado quiere encontrar una distribución de aulas para agruparlos según sus necesidades, para que todos los estudiantes de un aula tengan un nivel semejante.

Con el objetivo anterior se desea hacer un análisis de clusters sobre los estudiantes para agruparlos según sus resultados académicos previos. Para ello, se utilizarán las variables `G1.x, G2.x, G3.x, G1.y, G2.y, G3.y`.

Una vez estandarizados los datos se puede hacer un análisis jerárquico completo sobre ellos como se muestra a continuación:



Como se puede apreciar es posible crear una distribución de 6 aulas para agruparlos según sus resultados. Este cantidad es válida siempre que se dispongan profesores para impartir las clases. Como son 30 estudiantes y 6 profesores, la repartición inicial serían 5 estudiantes por aula, pero este análisis nos permite conformar de una mejor manera los grupos -que casualmente se obtuvo una media de 5 estudiantes por aula- para intentar que su aprendizaje sea más rápido dada la semejanza entre los integrantes, la cual aún desconocemos.

Dada la incertidumbre referente al nivel de semejanza de los clusters obtenidos, utilizaremos el algoritmo **k-means** de clusters no jerárquicos para encontrar este valor desconocido.

> **NOTA:** El resultado completo de aplicar **k-means** se puede ver en el log del comando `make cluster`

Este algoritmo desprende resultados favorables para la investigación:

```
Within cluster sum of squares by cluster:
[1] 0.6340087 7.9790560 11.6178608 7.0645402
1.4887801 0.0000000
(between_SS / total_SS = 83.5 %)
```

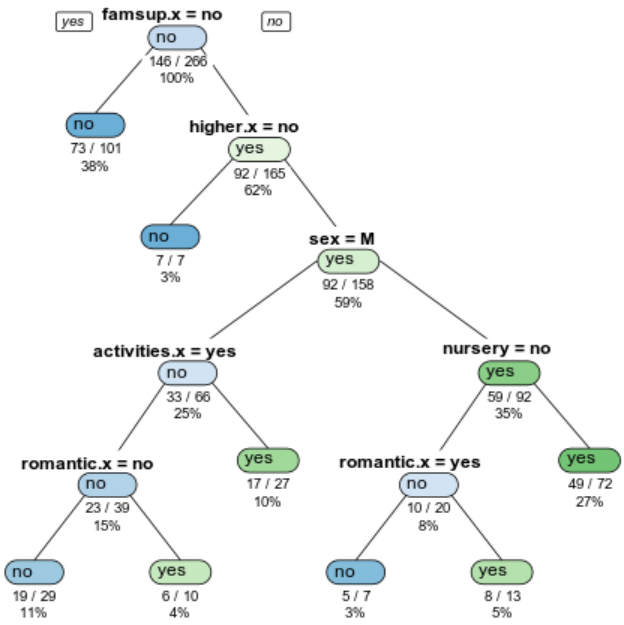
Como es apreciable, se obtuvo que el factor de semejanza de los clusters es de 83.5% lo cual implica que los estudiantes han logrado demostrar conocimientos parecidos en las evaluaciones por tanto se ha encontrado una buena distribución.

2.3.3 Árboles de Decisión

Construimos un árbol para analizar si un estudiante determinado debe optar por tener profesores particulares (además de los de la escuela) o no. Para esto utilizamos las siguiente variables cualitativas:

- `famsup.x` (apoyo educacional por parte de la familia)
- `higher.x` (quiere tener una educación universitaria)
- `nursery` (atiende además a la escuela de medicina)
- `activities.x` (hace actividades extracurriculares)
- `internet` (tiene internet)
- `sex` (sexo M o F)

Obtenemos el siguiente árbol, en el cual se pueden ver las probabilidades de si un estudiante debe optar por profesores particulares o no:



El error de entrenamiento es de 0.3879.

3. Conclusiones

Con las técnicas de regresión, se halló un modelo que nos va a permitir predecir dado un nuevo estudiante con distintos atributos, cual es su nota en el curso. Además, se puede concluir que el modelo cumple con los supuestos de la regresión lineal múltiple.

El análisis de los datos realizado con **ANOVA** demuestra que el aprendizaje y los resultados de los estudiantes en cuestión no depende de la zona donde esté ubicada su vivienda, por tanto para buscar datos que afecten su avance hacia mejores calificaciones será

necesario realizar nuevos tests.

De la reducción realizada (**ACP**) se desprenden datos interesantes, tales como la importancia de la edad(**PC1**) de un estudiante en sus estadísticas, lo cual implica que su nivel de madurez y responsabilidad representa un factor importante en sus resultados. También de manera perceptible tiene un efecto sobre su desempeño el nivel de estudios de su padre y madre (**PC2, PC3**), pues dicho nivel conlleva un mayor o menor nivel de exigencia de la familia hacia el estudiante.

Por último vale notar cómo la planificación del tiempo es vital, dado que tanto el tiempo de viaje (**PC4**) como el de estudio (bf PC5) tienen una marcada huella en la calidad del estudiante.

Las estrategias de generación de clusters permiten generar grupos con criterios de peso de manera sencilla, cosa que sería complicada manualmente dado que revisar un conjunto de 6 notas de 30 estudiantes para agruparlos no es una tarea trivial, por lo que en general no se recurre a ella en casos comunes y se conformarían los grupos a partes iguales o por afinidad entre profesores y alumnos, lo cual elimina la posibilidad de crear grupos balanceados donde todos los estudiantes puedan llevar el mismo ritmo de aprendizaje.

Los árboles de clasificación nos permiten estimar la probabilidad de que un evento determinado ocurra. En este caso, pudimos analizar la probabilidad de que un estudiante determinado deba optar por profesores particulares, dependiendo de varias variables.

References

[1] *Proyecto Evaluativo Estadística Fase 2.* (abrir)

[2] Artículo original