EXAMEN 2018 - Técnicas de Agrupación y de Reducción de la Dimensión

Álvaro de Prada Martínez

31/1/2018

Introducción:

En el presente trabajo se parte de un dataset que recoge los datos financieros de 1239 junto con el valor de 8 indicadores financieros distintos, y con una columna que indica si la empresa manipuló sus cuentas o no lo hizo.

A lo largo del trabajo se llevarán a cabo los análisis pertinentes con el fin de:

- Comprobar las relaciones subyacentes entre variables en las observaciones fraudulentas y las no fraudulentas, para conocer si son las mismas en un grupo y otro.
- Elaborar un análisis clúster que trate de identificar a ambos grupos de empresas.

Preparación de los datos

En primer lugar, vamos a resumir de forma breve el significado y funcionamiento de cada una de las variables incluidas en el dataset, para posteriormente proceder a su formateo.

Uno de los primeros datos que arroja esta tabla es que no existen NAs en los datos. Por lo que el siguiente paso será hacer un breve entendimiento de las variables. El dataset está compuesto por un conjunto de observaciones referentes a distintas empresas. En él, la columna Manipulater indica si dicha empresa cometió fraude o no (manipulación de los libros contables.)

En cuanto al resto de variables presentes en el dataset:

1. DSRI: Days Sales to Receivables Index () DSRI = (cuentas por cobrar(t)/sales(t)) / (cuentas por cobrar(t-1)/ventas(t))
2. GMI : Gross Margin Index (Margen Bruto) GMI = (ventas(t-1)-coste del bien vendido(t-1))/ventas(t-1) / (ventas(t)-coste del bien vendido(t))/ventas(t)
3. AQI : Asset Quality Index (Indice de calidad de activos) AQI = (1-((activo corriente-inmovilizado material)/total activos)) / (1-((activo corriente-inmovilizado material)/total activos))
4. SGI : Sales Growth Index (indice de ventas) SGI = sales(t) / sales(t-1)
5. DEPI : Depreciation Index (indice de depreciación) DEPI = Depreciation(t-1)/(depreciation(t-1)+inmovilizado material) / (depreciation(t)/(depreciation(t)+inmovilizado material(t)))

6. SGAI : Sales General and Administrative Index (indice general de gastos en aministrción y ventas) SGAI = (SGMI(t)/ventas(t)) / (SGMI(t-1)/ventas(t-1))
7. ACCR : Accruals to Total Assets (% de devengo sobre el total de activos) ACCR = (Bº antes de impuestos(t) - efectivo de operaciones(t) (efectivo)) / total de activos(t)
8. LEVI : Levenge Index (Indice de apalancamiento)

El M-Score de Beneish:

Todas las variables incluidas en el dataset tienen un factor común; el M-Score. El M-Score fue desarrollado por el profesor Messod Beneish y consiste en aplicar un peso a cada una de las variables anteriores, obteniendo así una cifra final. Si esta cifra es superior a -1.74, existen elevadas probabilidades de que la empresa esté manipulando la contabilidad. (Beneish, 1999) La fórmula del M-Score es la siguiente:

```
M = -4.84 + 0.92*DSRI + 0.528*GMI + 0.404*AQI + 0.892*SGI + 0.115*DEPI -
0.172*SGAI + 4.679*ACCR - 0.327*LEVI
```

A priori, esto no tendrá ningún impacto en nuestro trabajo, pero es importante conocer la utilidad que pueden tener las variables para tratar de conocer mejor el entorno del trabajo.

Tratamiento de datos

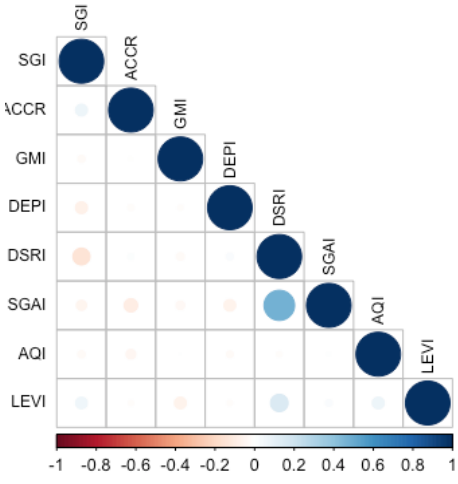
Una vez explicadas brevemente cada una de las variables, procedemos al análisis exploratorio y formateo de las variables.

1 - En primer lugar, transformamos la variable Manipulater a factor (1,0) para poder trabajar con ella.

	Manipulater = 0	Manipulater = 1
#	**1200**	**39**
%	**97%**	**3%**

Podemos comprobar en un primer acercamiento que nos encontramos ante un dataset poco balanceado, en el que las observaciones que manipulan es el 3% del total del dataset. En un principio no sabemos si esto tendrá impacto a la hora de llevar a cabo nuestro trabajo. Sin embargo, al final del trabajo llevaremos a cabo una técnica de balanceo del dataset para comprobar si los resultados varían significativamente.

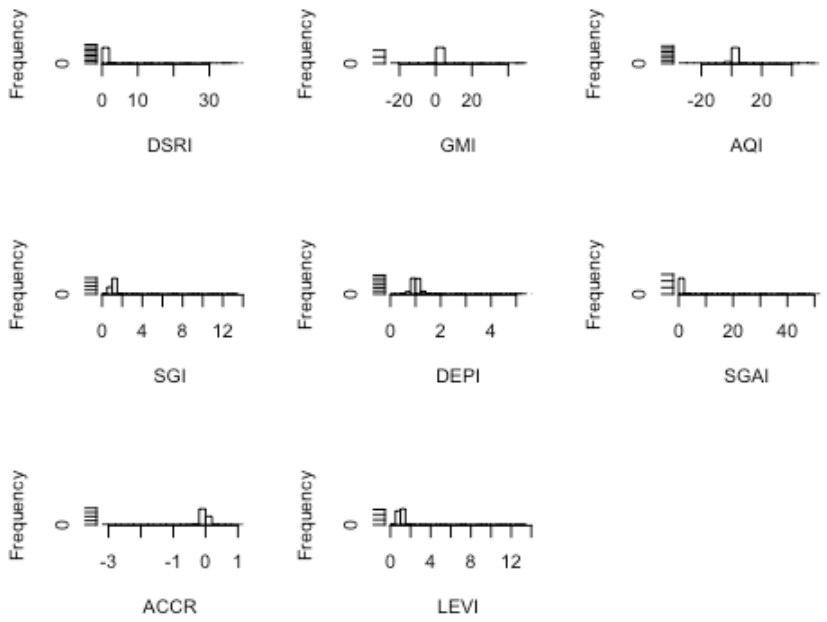
2 - Comprobamos la matriz de correlaciones de las variables de la totalidad del dataset, es decir, incluyendo grupo fraudulento y no fraudulento y representamos gráficamente:



Podemos observar en el gráfico superior que las correlaciones existentes entre variables son muy reducidas, siendo la mayor la correlación entre SGAI y DSRI con un 0.47. Como conclusión obtenemos que las variables se encuentran a nivel general muy poco correlacionadas entre ellas.

Podemos observar casi todas las variables tienen una media comprendida entre 0 y 1, con alguna excepción que se sale de este rango ligeramente. Sin embargo, comprobamos que los mínimos y máximos de casi todas las variables tienen valores muy alejados de este rango. Esto nos indica una alta probabilidad de que existan outliers, por lo que deberemos analizarlos para decidir cómo proceder con ellos.

Podemos también observar esta situación mediante una representación gráfica de las variables:

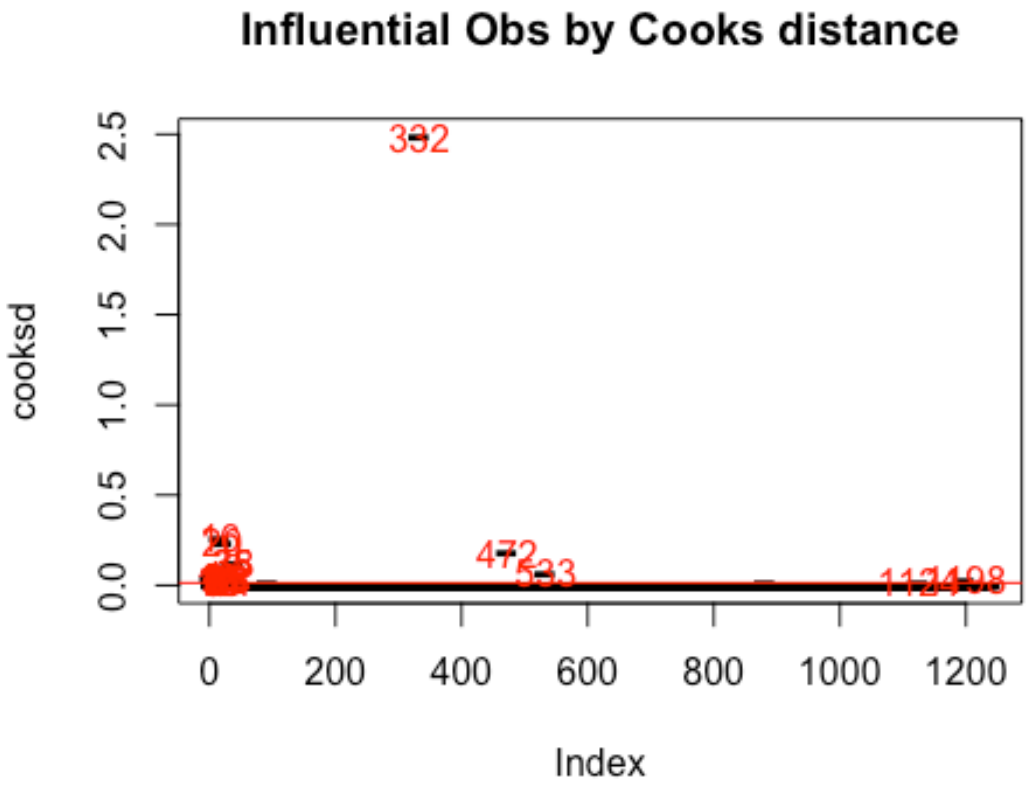


Comprobamos que, si bien la mayoría de las observaciones se comprenden en rangos reducidos, existen observaciones que toman valores alejados a estos, corroborando la hipótesis de la existencia de outliers.

Por este motivo, se llevará a cabo el análisis de outliers.

Análisis de Outliers:

Al no haber encontrado NAs, vamos a comprobar si existen outliers en el dataset que puedan hacer que los resultados de nuestro trabajo posterior se vean perjudicados por datos erróneos. Para ello emplearemos la distancia de Cook, y la representamos gráficamente. Todos los puntos por encima de la línea roja pueden ser considerados outliers, ya que difieren mucho del resto de observaciones en cuanto a la desviación típica media de las variables (concretamente hemos establecido el punto de corte en 4 veces la desviación típica media).



En la gráfica se observa que hay una serie de puntos que superan la línea de corte, indicando que son outliers. A continuación, extraeremos la tabla con las 6 primeras de estas observaciones para su análisis.

	DSRI <dbl>	GMI <dbl>	AQI <dbl>	SGI <dbl>	DEPI <dbl>	SGAI <dbl>	ACCR <dbl>	LEVI <dbl>	Manipulater <fctr>
1	1.624742	1.1289269	7.1850534	0.3662115	1.3815191	1.62414487	-0.166808698	1.1610817	1
2	1.486239	1.0000000	0.4655348	0.6728395	2.0000000	0.09288991	0.273434125	0.6809750	1
3	1.000000	1.3690378	0.6371120	0.8613464	1.4546757	1.74145963	0.123047699	0.9390472	1
4	0.905532	1.3609149	0.7839949	1.7933237	1.2782441	0.50526004	0.054642384	1.5431371	1
5	7.659976	0.5801841	1.0357910	1.4854401	0.6788332	0.65365114	0.003651969	1.1015913	1
6	1.346279	2.4735627	0.7025301	0.9656245	0.9852729	0.59451069	-0.013484949	1.0284588	1

6 rows

Podemos ver que las 6 son del tipo Manipulater=1, por lo que extraemos una tabla que nos indique el número de observaciones de tipo manipulater que aparecen entre los outliers:

	Manipulater=NO	Manipulater=YES
Nº observaciones	5	19
%	21%	79%

Las dos tablas anteriores arrojan mucha información, la más relevante es la siguiente: Las observaciones consideradas como outliers tienen en alguna de sus variables valores que se alejan excesivamente de la media. En total 24 outliers.

- ¿Debemos considerarla outliers y eliminarlas del dataset? Por un lado sí podríamos decir que se trata de outliers. Sin embargo, de las 24 observaciones obtenidas, 19 son de tipo Manipulater=1. Recordemos que en el dataset completo existían un total de 39 Manipulater=1/Yes. Es decir, más de la mitad de las observaciones de tipo Manipulater=1/Yes están compuestas de algún tipo de outlier.

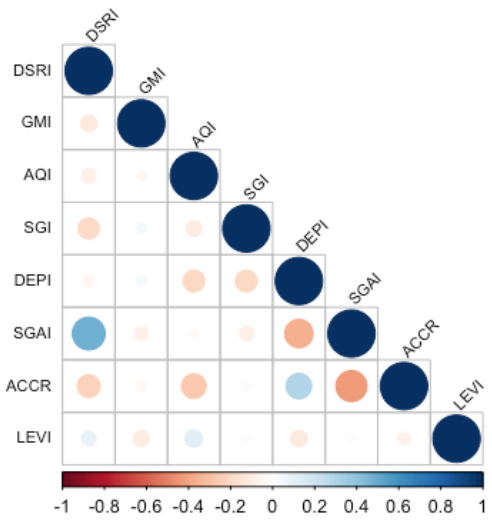
Por lo tanto, la conclusión que obtenemos es que, si bien pueden tratarse de outliers, lo más probable es que el propio hecho de que alguno de los ratios contenga un outlier eleva la probabilidad de que se esté manipulando la contabilidad. Por lo tanto, no podemos quitar los outliers del dataset ya que nos están aportando una información muy valiosa.

Concluimos aquí el análisis exploratorio de datos en el que hemos comprobado como existen numerosos outliers en los datos, pero tras hacer un análisis de los mismos hemos decidido mantenerlos en el dataset de trabajo ya que consideramos que el hecho de que una observación contenga algún tipo de outlier puede ser considerado como una característica común entre las empresas que cometen fraude. Por lo tanto, el hecho de que una empresa contenga outliers en alguno de sus ratios aporta una información muy valiosa a los datos.

PRIMERA PARTE: Relaciones de variables entre grupo fraudulento y no fraudulento

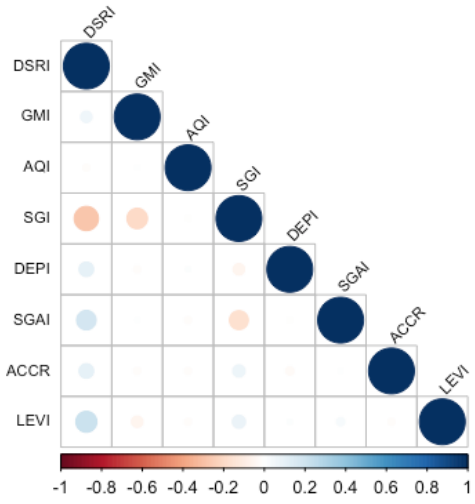
A continuación, una vez que hemos comprobado que no existen relaciones significativas entre las variables sobre el dataset total que incluye ambos grupos de empresas, llevaremos a cabo un análisis por separado de ambos grupos con la finalidad de hallar relaciones entre variables dentro de ellos, y comprobar si coinciden entre ambos grupos.

Para ello, inicialmente hemos dividido el dataset en dos subconjuntos, uno que incluye todas las observaciones que manipulan y otro que incluye las que no manipulan. Se ha representado la matriz de correlaciones entre variables de las observaciones que sí manipulan:



Comprobamos que, aunque las correlaciones son más fuertes en este subconjunto que en el dataset total, sigue sin existir una fuerza de correlación suficiente entre las mismas.

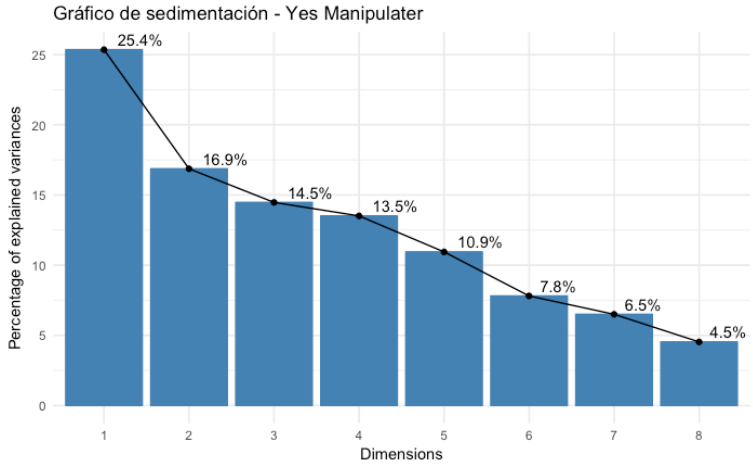
Realizamos el mismo gráfico sobre las observaciones que no manipulan:



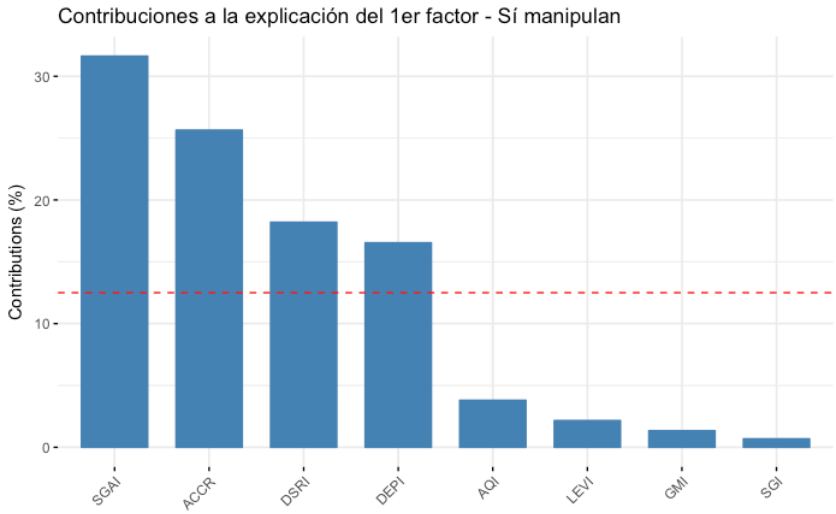
En esta ocasión podemos comprobar que esta matriz de correlaciones guarda mayor parecido con la matriz de todo el conjunto de datos. Esto es lógico ya que, como hemos visto anteriormente, el dataset está compuesto en su mayoría por este tipo de observaciones. Por lo tanto, concluimos que las correlaciones entre ambos subconjuntos no son demasiado fuertes.

A continuación, llevaremos a cabo un gráfico de sedimentación que nos permita identificar cual es capacidad explicativa de los factores en cada uno de los subconjuntos. En primer lugar, realizamos el gráfico con el subconjunto de "Yes":

Observamos que las tres primeras dimensiones son capaces de explicar más del 50% de la variabilidad, por lo que analizaremos la composición de dichas dimensiones para comprobar que variables son importantes en ellas.

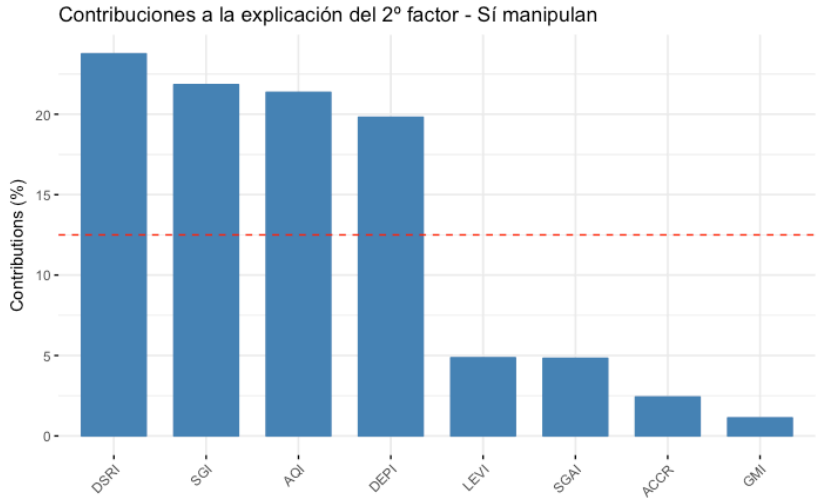


Primer Factor:



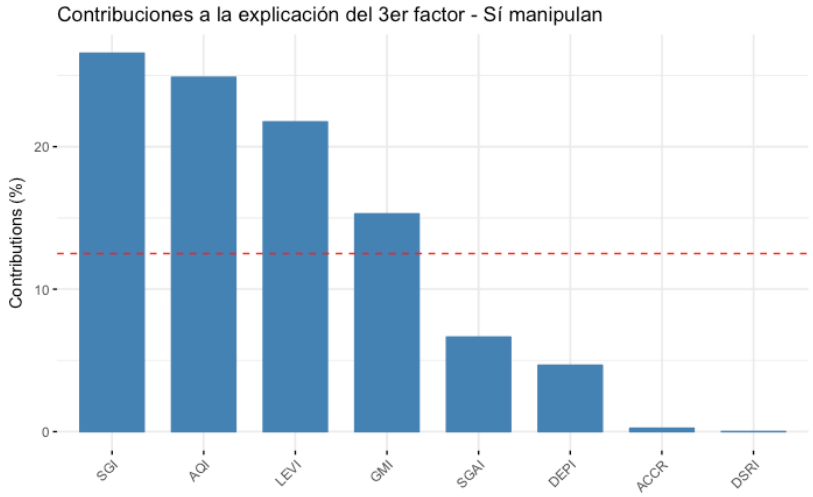
El primer factor, que es el que más variables explica, da un porcentaje de importancia elevado a las variables SGAI, ACCR, DSRI y DEPI.

Comprobemos el segundo factor:



Podemos ver que en el segundo factor son las variables DSRI, SGI, AQI y DEPI las que contribuyen a la explicación de la variabilidad. Coinciden con el primer factor las variables DSRI y DEPI

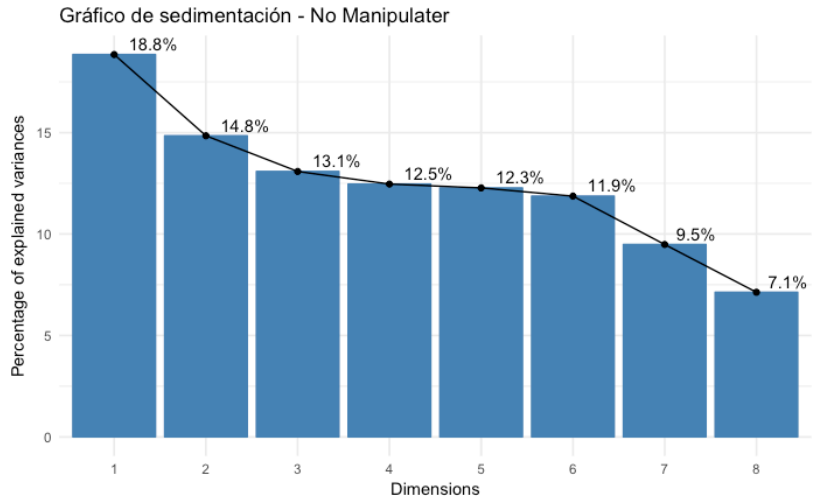
Comprobemos el tercer factor:



En este tercer factor son las variables SGI, AQI, LEVI y GMI las que lo explican, coincidiendo con los anteriores factores las variables SGI y AQI, que explicaban el segundo factor.

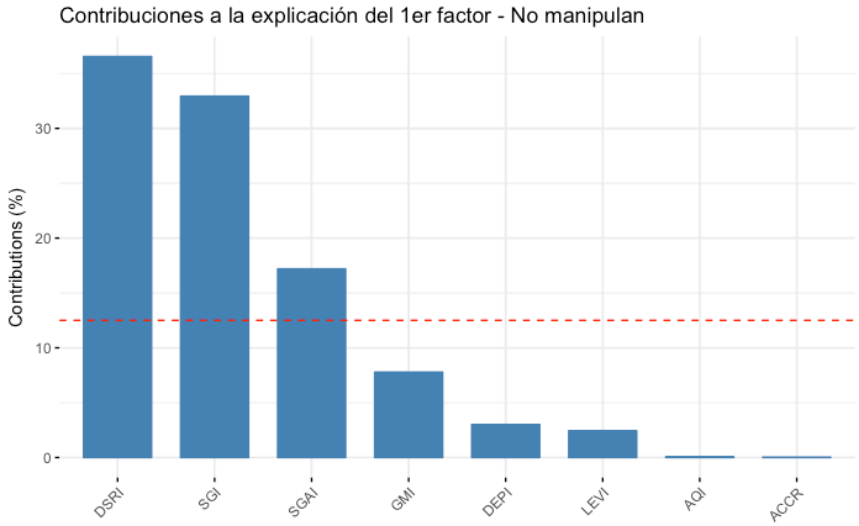
Por lo tanto, obtenemos como conclusión que en el subconjunto de empresas que sí manipulan su contabilidad, las variables DEPI, DSRI, SGI y AQI contribuyen en una medida superior al resto a explicar que estas empresas manipulen la contabilidad.

A continuación, llevaremos a cabo el mismo análisis sobre el subconjunto de empresas que no manipulan su contabilidad.



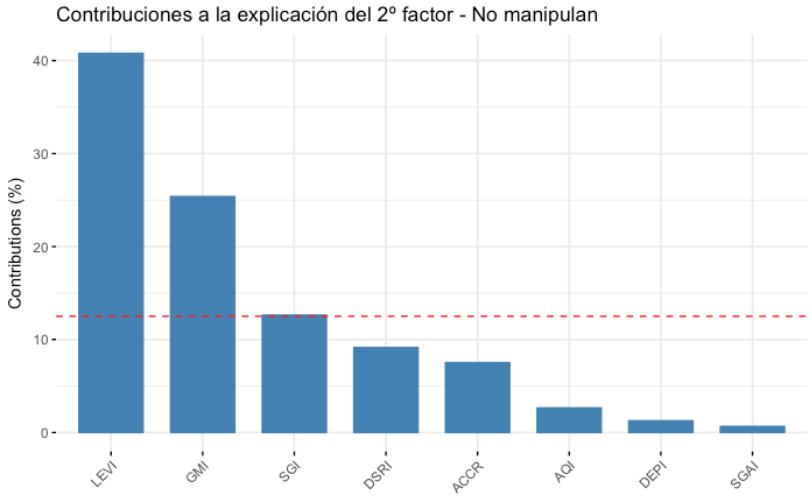
Comprobamos que en este caso, entre los 3 primeros factores se explica casi el 50% de la variabilidad, por lo que se analizará la composición de estos dos factores.

Primer Factor:



Comprobamos que la totalidad de la variabilidad explicada por el primer factor se basa en las variables DSRI, SGI y SGAI.

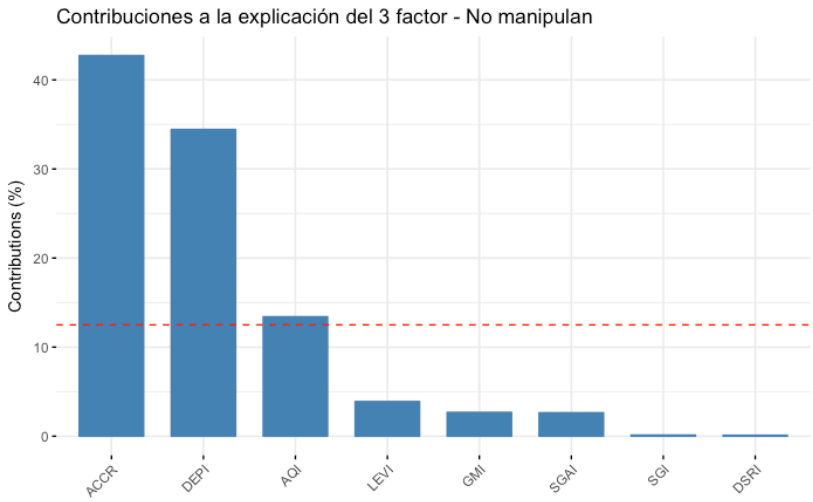
Vemos que sucede con el segundo factor:



En este caso el factor se compone en su mayoría por las variables LEVI, GMI y SGI.

Vemos que en la explicación tanto del primer y segundo factor aparece la variable SGI.

Tercer factor:



Por último vemos que en el tercer factor son las variables ACCR, DEPI y AQI las que lo explican.

Por lo tanto, en el subconjunto de empresas que no manipulan su contabilidad obtenemos que el valor que tomen las variables DSRI, SGI y SGAI tienen gran peso en esta categorización ya que forman parte de la explicación del primer factor, que es el que más porcentaje de variabilidad explica.

Finalmente comparamos los resultados de los análisis de ambos subconjuntos:

- Variables importantes de empresas que sí manipulan: DEPI, DSRI, SGI y AQI
- Variables importantes de empresas que no manipulan: DSRI, SGI y SGAI

(Las variables indicadas han sido consideradas ‘importantes’ por repetirse en 2 o más factores)

Por lo tanto, como conclusión a este primer apartado en el que se busca conocer las relaciones e importancia de las variables dentro de cada subconjunto, obtenemos que sí bien

en ambos subconjuntos existen ciertas variables que cobran más importancia que otras a la hora de asignar la pertenencia a una categoría u a otra, en ambos grupos el análisis coincide que la importancia de las variables DSRI y SGI es elevada. Por lo tanto, concluimos que estas variables tienen una importancia superior al resto, y pueden ser importantes a la hora de clasificar una empresa en un grupo u otro.

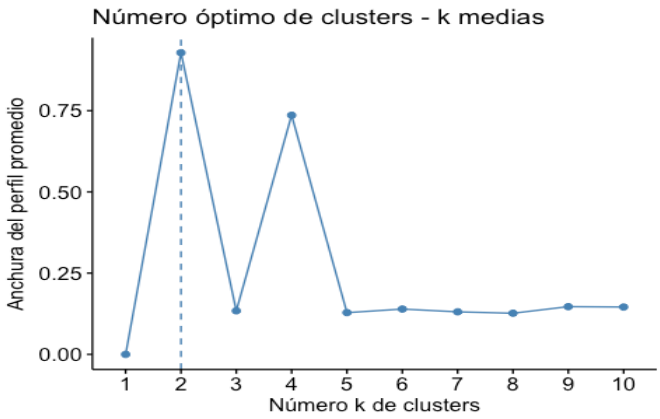
SEGUNDA PARTE: Análisis Clúster:

Una vez descartada la realización del PCA, vamos a llevar a cabo un análisis clúster sobre las observaciones del dataset.

A continuación, de cara a realizar el análisis clúster, vamos a efectuar una sería análisis sobre el dataset para poder averiguar cuál puede ser el método óptimo a la hora de realizar el análisis clúster.

En el siguiente gráfico se observa el número de clúster óptimo en el que se tendría que dividir los datos:

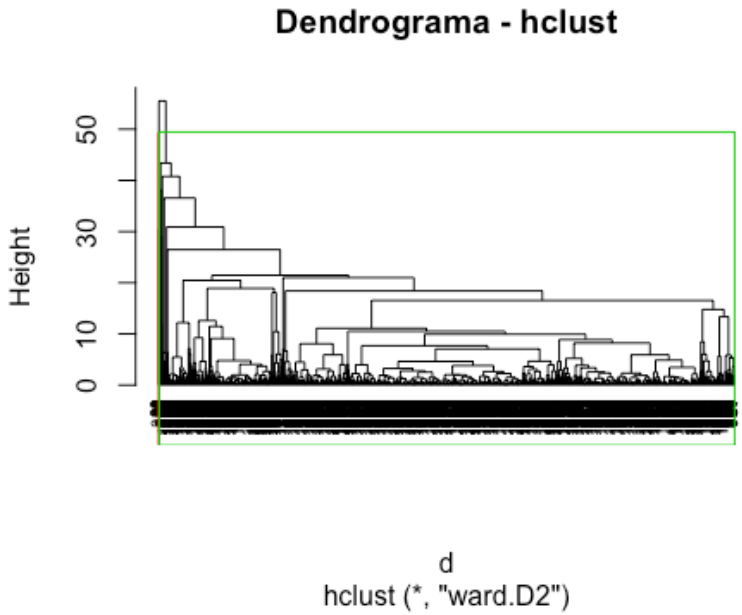
Podemos comprobar que el número óptimo de clúster siguiendo el método k-means sería de 2 clúster.



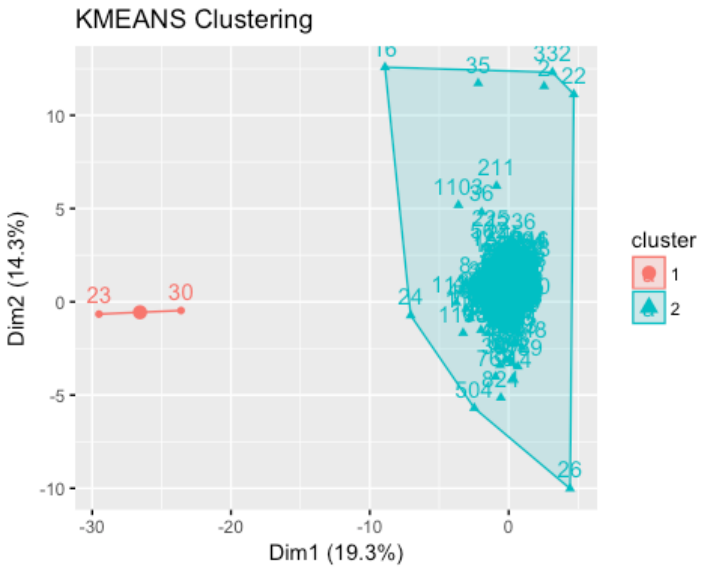
Por lo tanto, llevaremos a cabo una clusterización del dataset en 2 conjuntos. (ara ello hemos calculado previamente la distancia euclídea).

Ejecutamos la función hclust sobre el dataset, especificando el método "ward.D2" y la matriz de disimilaridades previamente calculada:

Y representamos gráficamente el dendograma:



Finalmente observamos los 2 clusters:



Vemos que un análisis clúster sobre la totalidad de los datos no aporta apenas información, ya que la división que se consigue haciendo clúster no está muy balanceada.

Éste sería el resultado de la clusterización del dataset de original.

La siguiente tabla muestra la media de cada variable en cada uno de los clúster, pero el hecho de que el clúster 1 esté formado sólo por 2 observaciones, limita mucho estos resultados.

clusters	DSRI	GMI	AQI	SGI	DEPI	SGAI	ACCR	LEVI
1	25,81735	-0,218409	0,9792959	0,04611154	0,6860753	30,548033	-0,2902734	1,097189
2	1,12923	0,9898121	0,9978485	1,12883368	1,0407149	1,059629	-0,0320039	1,057029

Comprobamos que en el clúster 1 las variables DSRI y SGAI tienen medias muy superiores a las del clúster 2, por lo que intuimos que es un clúster que contiene un porcentaje elevado de outliers.

También podemos intuir gracias al análisis exploratorio de datos previo, que el modelo ha podido orientar su división en base a aquellas observaciones que cometen fraude, por lo que vamos a comprobar cuantas veces aparece en cada clúster una observación que manipula:

	Manipulater=YES	% Yes	Manipulater=NO	% No
Cluster 1	2	100%	0	0%
Cluster 2	1200	97%	37	3%

Si bien en el segundo clúster la división no está tan clara, en el clúster número 1 sí queda claro que se incluyen únicamente valores de tipo Manipulater = Yes.

Esto nos hace plantearnos la siguiente pregunta: ¿Habría mejorado el análisis clúster de haber tenido un dataset más balanceado entre observaciones que manipulan y observaciones que no manipulan?

Para responder a esta pregunta hemos llevado a cabo una investigación de posibles soluciones, hallando una que puede ser muy útil y que será empleada a continuación.

Balanceando los datos:

En la búsqueda de soluciones a los problemas de balanceo, descubrimos que se trata de algo "típico" en problemas de fraude como el que nos ocupa el tiempo actualmente. Se han desarrollado numerosos métodos para tratar de diluir este problema de balanceo. La solución que emplearemos se encuentra en la librería mlr y consiste en procesos de "oversampling" y "undersampling".

- Oversampling: Consiste en generar observaciones adicionales de la clase minoritaria (en nuestro caso las observaciones con Manipulater=Yes), mediante la creación de copias exactas de observaciones minoritarias existentes, mientras que la clase mayoritaria no se manipula, sigue siendo constante.
- Oversampling-Smote: Se trata de una variación del oversampling que sigue la misma metodología, salvo que en lugar de duplicar observaciones minoritarias (que podría ocasionar overfitting), con este método las nuevas observaciones de la clase minoritaria son creadas siguiendo el método de los vecinos más cercanos.
- Undersampling: Esta técnica seria la contraria a las anteriores, ya que consiste en la eliminación de variables mayoritarias para conseguir de esta manera equilibrar las proporciones.

De entre los métodos descritos, hemos decidido emplear el método de smote al considerar que su forma de aumentar las observaciones de la clase minoritaria es más correcta. Hemos descartado también el proceso de undersampling ya que consideramos que el dataset no contiene un número suficiente de observaciones para que este método pueda ser llevado a cabo sin alterar los resultados de forma significativa sobre la realidad.

SMOTE:

Una vez aplicada la técnica de oversampling-smote, la proporción quedaría de la siguiente manera (aumenta 8 veces el número de observaciones tipo Manipulater=Yes):

```
##   0    1
## 1200  312
```

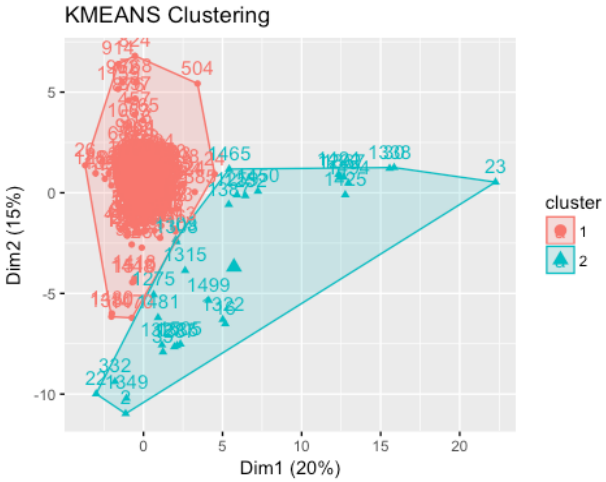
Como podemos comprobar, las observaciones minoritarias pasan de ser 39 a 312. Podríamos haber elegido que la proporción fuera menor aun (50-50 por ejemplo), pero consideramos importante que el dataset original sea lo más fiel posible a los datos originales.

Es importante aclarar que no se van a repetir los pasos de análisis de variables llevado a cabo en la primera parte del trabajo, ya que el nuevo dataset no modifica las variables, simplemente las 'refuerza' aumentando la población, por lo que podemos concluir que en este sentido las correlaciones no varían de forma significativa.

Con este nuevo dataset sí tiene sentido realizar un análisis clúster, pues al haber más observaciones de la clase minoritaria, puede haber separaciones más claras, ya que como hemos comentado anteriormente, existían indicios de que en la división de clúster llevada a cabo tuviera un peso significativo el si la observación era o no Manipulater.

Por todo lo expuesto, a continuación se volverá a llevar a cabo el análisis clúster con el dataset SMOTE.

Análisis CLUSTER - SMOTE:



Comprobamos como claramente el método de oversampling-smote llevado a cabo ha conseguido mejorar de forma importante la clusterización, habiendo una clara diferenciación entre un clúster y otro.

Comprobamos la media de cada variable por cluster:

sclusters <dbl>	DSRI <dbl>	GMI <dbl>	AQI <dbl>	SGI <dbl>	DEPI <dbl>	SGAI <dbl>	ACCR <dbl>	LEVI <dbl>
1	1.179429	1.1510052	1.1709541	1.158036	1.0396046	1.079385	-0.01425769	1.021609
2	9.266997	0.4023128	0.8173671	2.971611	0.7811722	8.538937	-0.05224357	4.442833

Y la desviación típica:

sclusters <dbl>	DSRI <dbl>	GMI <dbl>	AQI <dbl>	SGI <dbl>	DEPI <dbl>	SGAI <dbl>	ACCR <dbl>	LEVI <dbl>
0	0.7646608	2.273999	3.1069333	0.5145116	0.3398927	0.5011904	0.1511696	0.3650831
0	9.9271957	1.543890	0.5672763	4.0245418	0.2663623	11.3555257	0.2066019	4.1756187

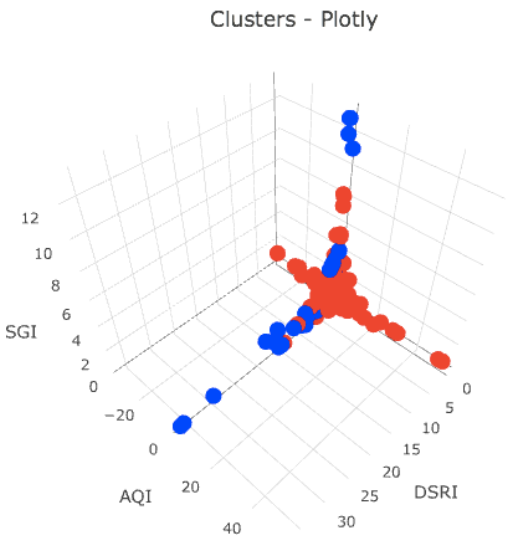
Como podemos comprobar, la desviación típica de la mayoría de las variables es considerablemente mayor en el clúster 2 que en el 1.

	Manipulater=YES	% Yes	Manipulater=NO	% No
Cluster 1	1198	81%	284	19%
Cluster 2	2	7%	28	93%

Conclusión:

Finalmente, comprobamos lo que habíamos podido intuir en la elaboración del clúster inicial, que existen 2 clúster diferenciados principalmente en base a si una empresa manipula o no manipula la contabilidad. En el caso del clúster 1, está compuesto tanto por empresas que manipulan como por empresas que no manipulan, aunque el porcentaje de empresas que no manipulan es del 81%. Sin embargo, en el segundo clúster únicamente dos empresas del clúster no manipulan, lo que supone un 7% en este caso. Podemos afirmar que gracias al método Smote empleado hemos podido facilitar la separación de los clúster, ya que con este método se ve más claramente lo que se podía intuir en el primer análisis; que el clúster 2 lo forman principalmente empresas que manipulan la contabilidad.

Para concluir nuestro trabajo, queremos relacionar el análisis realizado en la primera parte con el trabajo realizado en la segunda. Por ello, lo que queremos comprobar es si las variables que hemos considerado importantes en la primera parte por ser explicativas en los factores, tienen también importancia a la hora de clusterizar los datos, por lo que se ha representado en el gráfico las observaciones (del conjunto oversampled-smote) según las variables que hemos obtenido como importantes: DSRI, SGI y AQI(esta variable aparecía como importante en las empresas que sí manipulan)



Podemos comprobar que efectivamente parece haber una separación entre clúster (rojo, azul) según las variables consideradas importantes en el presente trabajo. Si bien el clúster 1(rojo) parece tomar valores de DSRI y SGI cercanos al 0, sobre la variable AQI tiene observaciones que toman valores alejados de 0. Por la parte del clúster clúster 2(azul), la variable AQI toma valores cercanos a 0 siempre.

En el clúster 2 (azul), compuesto por una mayoría que manipula, observamos que en los indicadores DSRI y SGI se toman valores muy amplios, cosa que no sucede en el clúster 1(rojo).

Bibliografía:

Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24-36.)

Librería MLR: *https://mlr-org.github.io/mlr-tutorial/release/html/over_and_undersampling/index.html*