Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

FACULTAD DE CIENCIAS EXACTAS Y NATURALES
FACULTAD DE INGENIERÍA

Análisis Inteligente de Datos

Trabajo Práctico Final

ALUMNA

Opazo, F. Ayelén

EQUIPO DOCENTE

Chan, Débora

Balzarotti, Federico

Oliva, Cecilia

Agosto, 2021

Predicción de rotación de empleados y clusterización

Opazo, F. Ayelén

Agosto, 2021

RESUMEN

El presente trabajo tuvo por objetivo predecir qué características hacen a los empleados más propensos a la rotación. Adicionalmente se buscó clasificar a los mismos empleados en distintos grupos o clusters para observar alguna posible particularidad que sume al entendimiento de la rotación. La métrica utilizada para evaluar el rendimiento de los análisis de clasificación supervisados fue el *F2-Score*; mientras que para seleccionar la cantidad de clusters con el método *k-means* se usó un criterio de validación interna: la métrica *Silhouette*. Se encontró que el mejor análisis para predecir cuáles son los empleados con mayor probabilidad de abandonar la compañía es el discriminante cuadrático, con un *F2-Score* de 0,95, y que 3 clusters es el número óptimo con un score silhouette de 0,34.

DATOS

El dataset utilizado está compuesto por 14999 observaciones de empleados que han o no rotado de su trabajo. El mismo posee 10 atributos, de los cuales 5 son numéricos (*satisfaction, evaluation, number of projects, average montly hours* y *time spend company*) y 5 son categóricos (*work accident, promotion, department, salary* y *churn*, siendo esta última nuestra variable dicotómica target en el modelo de aprendizaje supervisado). No contiene valores faltantes. A continuación, en la *tabla 1*, se muestran las primeras filas del dataframe:

	satisfaction <dbl>	evaluation <dbl>	number_of_projects <int>	average_montly_hours <int>	time_spend_company <int>	work_accident <int>	churn <int>	promotion <int>	department <chr>	salary <chr>
1	0.38	0.53	2	157	3	0	1	0	sales	low
2	0.80	0.86	5	262	6	0	1	0	sales	medium
3	0.11	0.88	7	272	4	0	1	0	sales	medium
4	0.72	0.87	5	223	5	0	1	0	sales	low
5	0.37	0.52	2	159	3	0	1	0	sales	low
6	0.41	0.50	2	153	3	0	1	0	sales	low

Tabla 1: 6-head dataset de rotación de empleados.

Dentro de los features categóricos *work accident, promotion* y *churn* poseen 2 valores posibles, siendo 0 o "No" en los tres casos el valor predominante en términos de frecuencia; mientras que *department* y *salary* contienen 10 y 3 valores posibles, respectivamente. En cuanto a los features numéricos, *satisfaction* (mean=0.61; sd=0.25)*, evaluation* (mean=0.72; sd=0.17) y *average montly hours* (mean=201; sd=49.94) son variables continuas y *number of projects* (mean=3.8; sd=1.23) y *time spend company* (mean=3.5; sd=1.46)*,* en años, si bien pueden pensarse también como continuas, en la muestra utilizada toman valores discretos. En el ANEXO 1 se muestran las distribuciones univariadas.

METODOLOGÍA

Para una primera aproximación hacia las relaciones entre las variables y la selección de aquellas que mejor discriminan, se realizan tablas de contingencia y test de independencia para las categóricas, por un lado, y análisis de correlación de Spearman y ACP para las numéricas, por el otro; acompañadas de técnicas gráficas. Si bien la cantidad de observaciones es grande para la realización de los tests, se efectúan solo a modo orientativo y sus resultados no son concluyentes. De forma esperable, las variables más correlacionadas son *number of projects* y *average montly hours* (0.40), mientras que *satisfaction* y *churn* tienen una correlación negativa de -0.37 (ver ANEXO 2). Pese a la poca correlación de las variables originales, exceptuando las comentadas, se realiza un biplot exploratorio que permite acercarnos a una idea de distribución multidimensional entre las variables y el personal que ha y no ha rotado, así como buscar grupos o patrones antes de realizar las clasificaciones. En la *Figura 1* se observan componentes principales de contraste, o de forma. Con el 36.6% de variabilidad explicada podríamos asociar a la primera componente con el nivel de satisfacción de los empleados, mientras que a la segunda, con el 22.2% de variabilidad, con el tiempo dentro de la compañía y cantidad de proyectos asumidos, que podríamos pensar como variable latente *"engagement"*. En cuanto a los colaboradores que han rotado, en la *Figura 2* se concentran más hacia los extremos, en evaluaciones altas y menor tiempo promedio, así como también mayor tiempo en la compañía y menor satisfacción, que puede indicar personas con buen *desempeño* y necesidad de desafío pero baja satisfacción y "*engagement*", pero también los hay de bajo valor de evaluación, o peor desempeño, con baja satisfacción. Por otro lado, los que no rotan se concentran del eje de coordenadas hacia afuera, lo que indica porcentaje mayor de valores más promedio partiendo del centro.

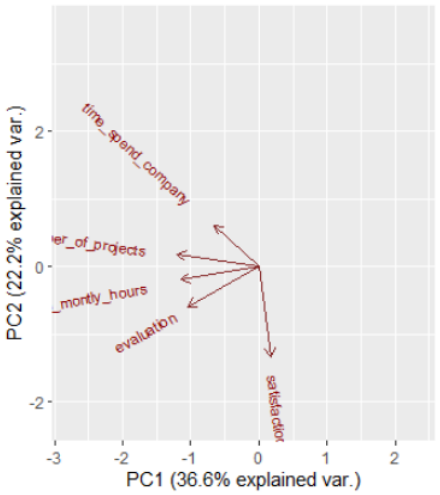

Figura 1: Biplot ACP variables numéricas.

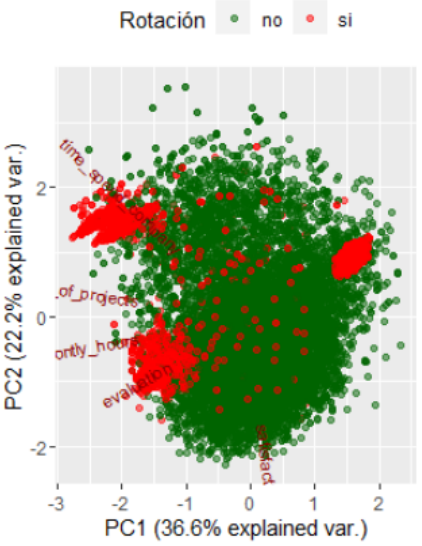

Figura 2: Biplot con distribución de clases.

A partir de esta primera exploración de los datos se decide continuar con las clasificaciones supervisada y no supervisada únicamente con los valores que son numéricos, sumando la variable clase (*churn*) en el caso de la supervisión. El motivo se relaciona con cierta independencia detectada entre la variable target y las categóricas (con un *p-value*>0.05), pero se adicionan razones de negocio: los features numéricos se pueden trabajar con mayor facilidad para provocar cambios desde aspectos de desarrollo humano. De esta manera quedan 6 variables para efectuar los análisis que siguen a continuación.

Clasificación supervisada

Para realizar la clasificación supervisada, a partir de la clase *churn,* se separan los datos en train y test en una proporción 70/30, respectivamente, y se prueban distintos métodos de aprendizaje. Como el objetivo de esta clasificación es la retención de colaboradores se tiene una preferencia por reducir la tasa de *falsos negativos* (error de tipo II), más que la de *falsos positivos* (error de tipo I), es decir, es preferible que los modelos clasifiquen a un individuo como un posible candidato a irse de la organización incorrectamente a que clasifiquen a uno que sí es propenso a rotar como negativo. En un caso la clasificación errónea es inocua y hasta propicia; mientras que en el otro tiene un impacto directo en términos de costo de negocio. Se utiliza, entonces, la métrica de exhaustividad *F2-Score*, dando peso al *recall,* para evaluar los resultados de los métodos de clasificación que informa sobre la *cantidad* que el modelo es capaz de identificar, teniendo en cuenta además que existe cierto desbalance de los datos target (76.19% *churn*=0 vs. 23.81% *churn*=1).

El modelo planteado para los distintos análisis es de nuestra variable *churn* contra las 5 numéricas que se mantienen. El primer análisis es un **discriminante lineal (LDA)** que, si bien los datos no satisfacen los supuestos de normalidad multivariada y homocedasticidad con los test de Shapiro-Wilk y Box's M-test respectivamente (ambos con un *p-value*< 2.2e-16), se aborda para evaluar el rendimiento de la clasificación. En la *Figura 3* se observa además que las clases no pueden dividirse linealmente; sin embargo, este método aporta a la identificación de variables con mayor importancia para la división de clases.

Figura 3: Distribución de variables numéricas según clase.

Como no se satisfacen los supuestos expuestos arriba se prueba con un **análisis discriminante cuadrático (QDA)** y también con un **análisis discriminante robusto (RDA)**. Se compara la rotación real con las clasificaciones de los 3 análisis por fuerza bruta, sin ajuste de parámetros, y se obtienen los resultados de la *Figura 4*:

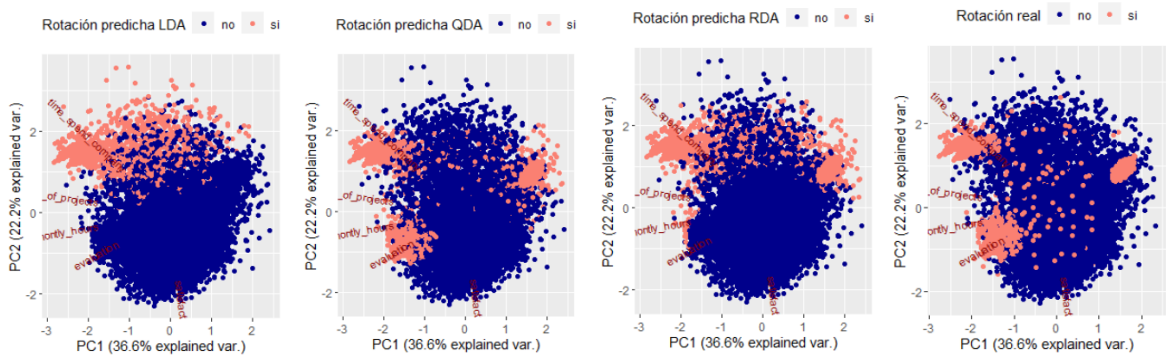

Figura 4: Biplots de clasificación de cada análisis y comparación con la realidad.

A simple vista se puede ver que el modelo que mejor predice es el QDA, dado que su representación es la más similar con la realidad. El RDA puede tomar un segundo lugar, pero no tendría mucho sentido su aplicación, ya que no existen grandes valores atípicos en este dataset. El lineal, que presentaba dificultades teóricas para su aplicación, tampoco clasifica adecuadamente en la práctica. Se opta, entonces, por el QDA pero se realiza

también una **regresión logística (LG)** para evaluar variables más influyentes en la probabilidad de ocurrencia del evento de rotación. Se ajusta un umbral óptimo mayor a 0,3 de todos los modelos y en resultados se muestra lo obtenido.

Clasificación no supervisada

Para la clasificación no supervisada se opta por la utilización del algoritmo **k-means**, manteniendo las mismas variables numéricas pero descartando *churn*, ya que esta vez no se trata de una predicción, sino que el objetivo en este caso es más abierto y a partir de la elección de distintos parámetros se interpretan los resultados que de ellos se derivan. Para el algoritmo k-means se parametriza la distancia *euclídea* y se define la cantidad de clusters a través de la métrica de validación interna: *Silhouette.* Para silhouette, el rango de score va de 0 a 1 y, a mayor score obtenido, se espera mejor separación entre los grupos generados. En la *Figura 5* se muestran el número de clusters (definido en el rango de 2 a 15) y el valor de esta métrica para cada uno de estos k-cluster:

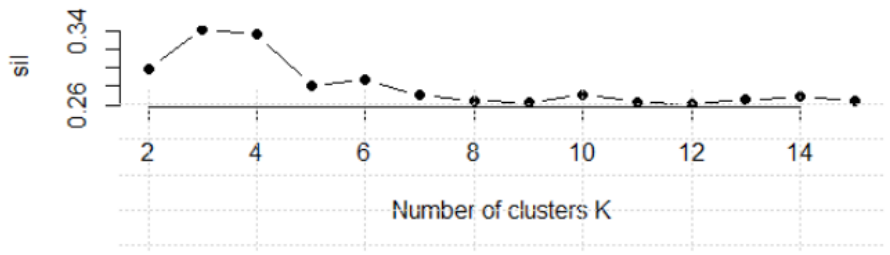

Figura 5: Número de clusters según métrica Silhouette.

Lo que nos muestra esta figura es que en 3 cluster hay mayor cohesión y menos interferencia entre clases, ya que coincide con el máximo score de silhouette para este conjunto, de 0,34. Se utiliza, entonces, un valor de 3 para generar los grupos de empleados. Un primer aspecto llamativo es que en el biplot de rotación (*Figura 2*) veíamos cierta concentración de colaboradores que rotaron también en 3 agrupaciones claras; por tal motivo se decide mostrar también los resultados del clustering en un biplot, a modo comparativo.

RESULTADOS

Clasificación supervisada

Comparamos en matrices de confusión las clasificaciones de los análisis y modelos del aprendizaje supervisado en datos de *test*. Los resultados se muestran en la *Figura 6:*

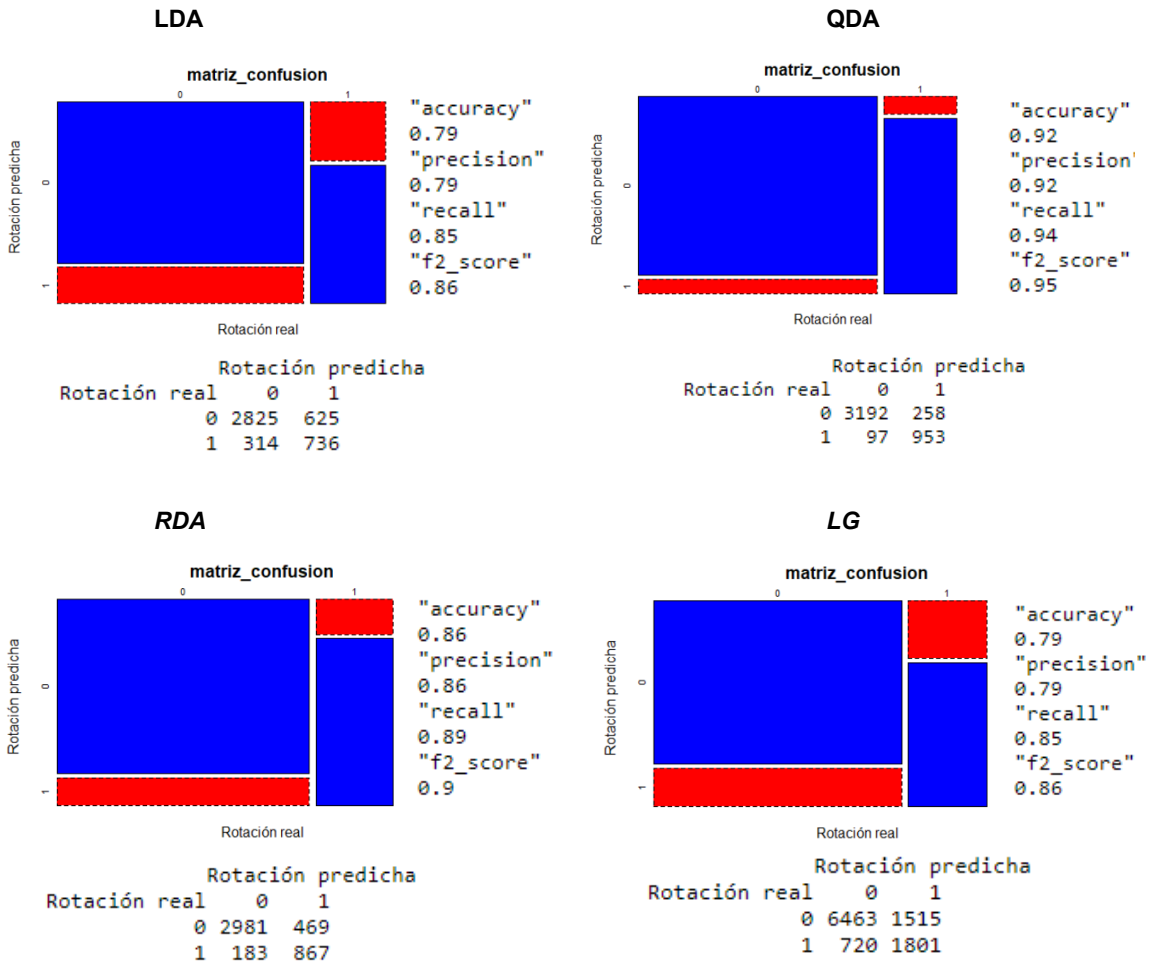

Figura 6: Matrices de confusión y métricas obtenidas en test.

Como se puede notar, se valida que el QDA es el análisis que más disminuye la tasa de *falsos negativos*, con un score de 0.95 de F2 y buenas métricas en general, por encima del 90%. Luego se encuentra el RDA con un score de 0.9 y, por último, la LG y el LDA con exactamente las mismas métricas. Sin embargo; más allá de las métricas resultantes, la mejor elección avalada por la teoría que puede hacerse es el QDA, ya que es un análisis que se aplica cuando no se cumplen los supuestos de normalidad multivariada y homocedasticidad, como en este caso, y además mejora en métricas a la LG, que también es un buen modelo para el objetivo de este estudio desde el aspecto teórico. Con respecto a las variables que mejor discriminan para clasificar e influyen en la rotación, tanto el LDA

como la LG coinciden en *satisfaction, time spend company* y *number of projects* (ver ANEXO 3).

Clasificación no supervisada

A continuación, en la *Figura 7*, se muestran los 3 grupos segmentados con k-means:

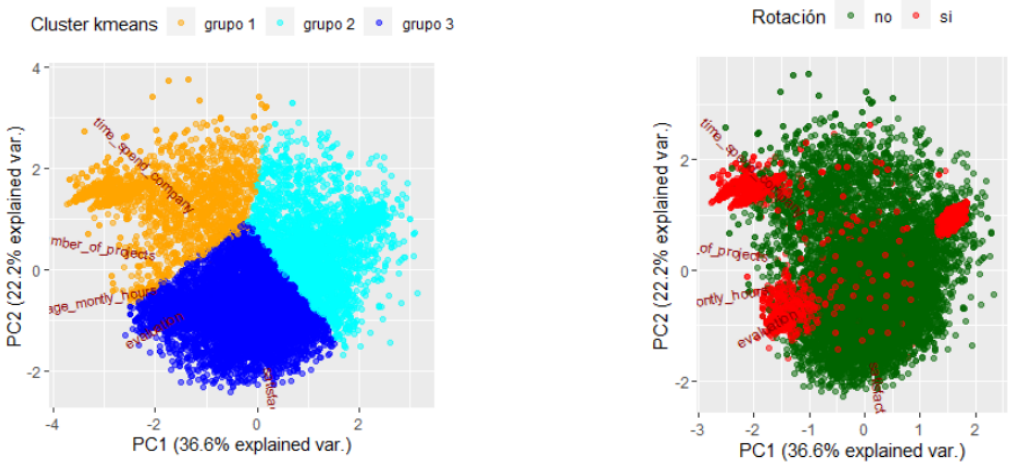

Figura 7: Biplot con 3 clusters k means y comparación con biplot de rotación.

Además de que en la *Figura 2* de rotación se ven 3 concentraciones claras de perfiles que abandonaron la compañía, los 3 cluster obtenidos en el mismo biplot contienen de manera separada a estos subgrupos que se encontraron en el ACP del análisis exploratorio. Siguiendo esta lógica, al *grupo 1* se lo puede relacionar con los colaboradores con mayor tiempo en la organización, al *2* con aquellos que tienen peor desempeño y al *grupo 3* con los que tienen mejor desempeño, en el cual existen individuos con alta satisfacción y con baja satisfacción, que coincide con los que rotan efectivamente. El *grupo 2*, entonces, es un perfil de colaboradores más “conflictivo” con los que se debería trabajar el desempeño. Este grupo, desde el punto de vista del negocio, es el que menos impacto tiene en caso de derivar en rotación. Así, se puede poner mayor foco a los grupos 1 y 3 en cuanto a la retención, ya que el primero es importante debido a la gestión del conocimiento que maneja, por la cantidad de tiempo en la organización, y el segundo es, quizás, el más importante porque allí se concentra hacia el extremo izquierdo los de mejor desempeño y que tienen probabilidad de rotación.

DISCUSIONES

A partir del desarrollo de este estudio algunas conclusiones interesantes tienen que ver con reconocer la influencia de los atributos *satisfaction, time spend company* y *number of projects* a la hora de trabajar principalmente la satisfacción laboral y la cantidad de proyectos en los que un colaborador participa. Con el ACP pudimos ver que a menor satisfacción hay más empleados que rotan , al igual que es más probable que abandonen la compañía aquellos que tienen menos tiempo en ella y hayan participado en menor cantidad de proyectos; ubicando cierta variable latente "*engagement"*. Con un buen abordaje de estas características por parte de las organizaciones hacia sus colaboradores puede disminuirse la tasa de rotación; atendiendo de forma latente a perfiles de alto *desempeño* y sin descuidar el *engagement*. Para las organizaciones con gran número de nómina (caso de *big data*), en el que no se cumplen supuestos de normalidad debido a su magnitud ni tampoco homogeneidad de varianzas, posiblemente no se puedan separar las clases de rotación sí vs. rotación no, con lo cual un análisis discriminante cuadrático podría funcionar adecuadamente; siempre y cuando no existan valores atípicos en esa multivariedad. También podría funcionar bien un modelo de regresión logística usando otras muestras. Por otra parte, pero acompañando estas interpretaciones, los 3 cluster encontrados con k-means se pueden entender como 3 tipos de colaboradores, relacionados a los que tienen mayor experiencia en la empresa, los que tienen bajo desempeño y los que tienen alto desempeño con distintos niveles de satisfacción. El foco principal para la prevención de rotación debiera ponerse en el *grupo 3,* ya que es esta segmentación la que generaría mayor impacto si ocurre tal evento.

Limitaciones: Los hallazgos expuestos deben tomarse con cautela, ya que las variables originales utilizadas para los desarrollos, tanto del aprendizaje supervisado como del no supervisado, no presentaban grandes correlaciones entre sí para el ACP y, por lo tanto, si bien sus análisis aportan información interesante, los mismos no son concluyentes y es necesario seguir profundizando en futuras investigaciones.

REFERENCIAS

Fuente dataset:
https://assets.datacamp.com/production/repositories/1765/datasets/ae888d00f9b36dd7d50a4afbc112761e2db766d2/turnover.csv

ANEXO

1- Distribución de variables categóricas y numéricas

Figura 5: Frecuencia de variables categóricas.

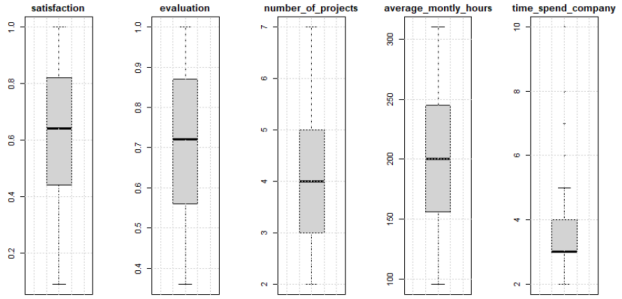

Figura 5: Distribución (boxplots) de variables numéricas.

2- Tablas de contingencia y correlación

number of projects	churn 0	churn 1
2	5.47	10.45
3	26.56	0.48
4	26.38	2.73
5	14.33	4.08
6	3.46	4.37
7	0.00	1.71

salary	churn 0	churn 1
high	7.70	0.55
low	34.30	14.48
medium	34.20	8.78

promotion	churn 0	churn 1
0	74.19	23.68
1	2.00	0.13

work accident	churn 0	churn 1
0	62.86	22.68
1	13.33	1.13

department	churn 0	churn 1
accounting	3.75	1.36
hr	3.49	1.43
IT	6.36	1.82
management	3.59	0.61
marketing	4.37	1.35
product_mng	4.69	1.32
RandD	4.44	0.81
sales	20.84	6.76
support	11.16	3.70
technical	13.49	4.65

time spend company	churn 0	churn 1
2	21.27	0.35
3	32.38	10.57
4	11.11	5.93
5	4.27	5.55
6	3.39	1.39
7	1.25	0.00
8	1.08	0.00
10	1.43	0.00

	churn
satisfaction	-0.37
evaluation	0.00
number_of_projects	-0.02
average_montly_hours	0.05
time_spend_company	0.27
work_accident	-0.15
churn	1.00
promotion	-0.06
department_accounting	0.02
department_hr	0.03
department_IT	-0.01
department_management	-0.05
department_marketing	0.00
department_product_mng	-0.01
department_RandD	-0.05
department_sales	0.01
department_support	0.01
department_technical	0.02
salary_high	-0.12
salary_low	0.13
salary_medium	-0.07

3- Variables importantes al particionar

LDA

	LD1
satisfaction	-1.05
evaluation	0.13
number_of_projects	-0.28
average_montly_hours	0.19
time_spend_company	0.30

LG

(Intercept)	satisfaction	evaluation	number_of_projects
-1.43	-1.04	0.15	-0.39
average_montly_hours	time_spend_company		
0.21	0.30		