Análisis Inteligente de Datos
MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

TRABAJO PRÁCTICO FINAL
Clasificación Supervisada y No supervisada

Análisis Inteligente de Datos
Explotación de Datos y Descubrimiento de Conocimiento

Alumna: Macarena Roel

Docentes: Débora Chan, Federico Balzarotti, Cecilia Oliva

Agosto 2021

Índice

Introducción

La rotación de empleados es un problema costoso para las empresas, ya que el costo real de reemplazar a un empleado a menudo puede ser bastante grande. Un estudio realizado por el Center for American Progress encontró que las empresas suelen pagar alrededor de una quinta parte del salario de un empleado para reemplazar a ese empleado, y el costo puede aumentar significativamente si los ejecutivos o los empleados mejor pagados deben ser reemplazados.

Es por eso que cualquier departamento de recursos humanos tiene dentro de sus objetivos detectar posibles motivos de fuga de empleados, ya sea por motivos ajenos a la empresa o intrínsecos a la misma.

Búsqueda de mejores oportunidades o salarios, fatiga y mejores prestaciones son sólo algunas de las variables que entran en juego a la hora de tomar la decisión de abandonar un trabajo.

En este trabajo, se analizan los datos de 1470 empleados ficticios generados por IBM, con el objetivo de poder clasificar a los mismos según si dejaron o no la empresa (de aquí en adelante, "*Attrition*").

Para ello se tomó un subconjunto de observaciones y se realizaron en base a los mismos una serie de análisis de clasificación supervisados y no supervisados.

Pre procesamiento de datos y análisis exploratorio

Para comenzar el análisis, fue necesario reducir el conjunto de datos con el que se trabajó. En un comienzo, se habían realizado los análisis con las 1470 observaciones, obteniendo porcentajes de clasificaciones correctas cercanos al 90%, pero al observarse los mismos con mayor detenimiento fue claro que dicho porcentaje se debía a la sobreestimación resultante del desbalanceo entre ambas clases de la variable objetivo, *Attrition*. Originalmente, había 1233 (84%) observaciones con un valor de Attrition de "*No*" y 237 (16%) de "*Yes*", lo que generaba que los modelos tendieran a priorizar al no en sus clasificaciones, generando muchos falsos negativos pero obteniendo altos valores de precisión igualmente.

Para intentar corregir esta situación, se generó un nuevo set de datos, conservando todos los valores afirmativos de la variable *Attrition* y seleccionando aleatoriamente 350 observaciones de los negativos, obteniendo así un set de datos de 587 observaciones, 350 (60%) negativas y 237 (40%) positivas.

Se consideró igualmente necesario reducir el número de variables a considerar. De las 36 variables originales, 15 fueron removidas, por considerarse que las mismas no aportaban al análisis buscado o tenían definiciones poco claras.

Dentro de las variables conservadas, muchas resultaron ser valores numéricos del tipo rango, con valores del 1 al 5 significando diferentes grados de conformidad o satisfacción, de acuerdo a la variable. Se decidió considerar a las mismas variables categóricas, y analizarlas como si así lo fueran, a pesar de saber que se podría haber decidido tomarlas como numéricas.

Finalmente, las variables consideradas fueron las presentadas en la Tabla I. En la misma pueden verse las variables utilizadas en el análisis, con su mediana y rango inter cuartil (RIC) en el caso de las variables numéricas y el N(%) de cada grupo en el caso de las variables categóricas, considerando a todas las observaciones estudiadas por un lado, y luego dividiendolas según su valor para la

variable respuesta, *Attrition*. Se obtuvo también un p-valor indicando si ambos grupos eran significativamente distintos de manera univariada en dichas variables, utilizando los tests de Fisher y chi-cuadrado para las variables categóricas y Kruskal-Wallis para las numéricas.

Tabla 1: descripción de las variables de estudio

	Total N(%)	**No**	**Yes**	**p-valor**
	N=587	***N=350***	***N=237***	
Age	34.0 [29.0;42.0]	36.0 [30.0;44.0]	32.0 [28.0;39.0]	**<0.001**
BusinessTravel:				**<0.001**
Non-Travel	62 (10.6%)	50 (14.3%)	12 (5.06%)	
Travel_Frequently	122 (20.8%)	53 (15.1%)	69 (29.1%)	
Travel_Rarely	403 (68.7%)	247 (70.6%)	156 (65.8%)	
Department:				**0.036**
Human Resources	24 (4.09%)	12 (3.43%)	12 (5.06%)	
Research & Development	366 (62.4%)	233 (66.6%)	133 (56.1%)	
Sales	197 (33.6%)	105 (30.0%)	92 (38.8%)	
DistanceFromHome	7.00 [2.00;16.0]	6.50 [2.00;14.8]	9.00 [3.00;17.0]	**0.016**
Gender:				0.473
Female	227 (38.7%)	140 (40.0%)	87 (36.7%)	
Male	360 (61.3%)	210 (60.0%)	150 (63.3%)	
HourlyRate	67.0 [50.0;85.0]	70.5 [50.0;86.0]	66.0 [50.0;84.0]	0.163
JobInvolvement:				**<0.001**
1	44 (7.50%)	16 (4.57%)	28 (11.8%)	
2	158 (26.9%)	87 (24.9%)	71 (30.0%)	
3	332 (56.6%)	207 (59.1%)	125 (52.7%)	
4	53 (9.03%)	40 (11.4%)	13 (5.49%)	
JobLevel:				**<0.001**
1	259 (44.1%)	116 (33.1%)	143 (60.3%)	
2	191 (32.5%)	139 (39.7%)	52 (21.9%)	
3	83 (14.1%)	51 (14.6%)	32 (13.5%)	
4	31 (5.28%)	26 (7.43%)	5 (2.11%)	
5	23 (3.92%)	18 (5.14%)	5 (2.11%)	
JobSatisfaction:				**0.016**
1	132 (22.5%)	66 (18.9%)	66 (27.8%)	
2	121 (20.6%)	75 (21.4%)	46 (19.4%)	
3	172 (29.3%)	99 (28.3%)	73 (30.8%)	
4	162 (27.6%)	110 (31.4%)	52 (21.9%)	
MaritalStatus:				**<0.001**
Divorced	121 (20.6%)	88 (25.1%)	33 (13.9%)	
Married	254 (43.3%)	170 (48.6%)	84 (35.4%)	
Single	212 (36.1%)	92 (26.3%)	120 (50.6%)	

	Total N(%)	No	Yes	p-valor
	N=587	*N=350*	*N=237*	
MonthlyIncome	4478 [2674;7408]	5152 [3200;8450]	3202 [2373;5916]	**<0.001**
MonthlyRate	14382 [8332;20538]	13994 [7821;20276]	14618 [8870;21081]	0.490
NumCompaniesWorked	2.00 [1.00;4.00]	2.00 [1.00;4.00]	1.00 [1.00;5.00]	0.354
OverTime:				**<0.001**
No	381 (64.9%)	271 (77.4%)	110 (46.4%)	
Yes	206 (35.1%)	79 (22.6%)	127 (53.6%)	
TotalWorkingYears	9.00 [5.00;13.0]	10.0 [6.00;16.0]	7.00 [3.00;10.0]	**<0.001**
TrainingTimesLastYear	3.00 [2.00;3.00]	3.00 [2.00;3.00]	2.00 [2.00;3.00]	0.125
YearsAtCompany	5.00 [2.00;9.00]	6.00 [3.00;10.0]	3.00 [1.00;7.00]	**<0.001**
YearsInCurrentRole	2.00 [1.50;7.00]	3.00 [2.00;7.00]	2.00 [0.00;4.00]	**<0.001**
YearsSinceLastPromotion	1.00 [0.00;3.00]	1.00 [0.00;3.75]	1.00 [0.00;2.00]	**0.027**
YearsWithCurrManager	3.00 [1.00;7.00]	3.00 [2.00;7.00]	2.00 [0.00;5.00]	**<0.001**

Por otra parte, la gran mayoría de las variables numéricas consideradas no tenían una distribución normal, teniendo por lo general una tendencia a la asimetría a la derecha o comportamientos multi modales. Para evitar que esto influyera en las clasificaciones realizadas posteriormente, las variables fueron escaladas y centradas antes de particionar al set de datos en entrenamiento y validación.

También se realizaron análisis de correlación y PCA con las variables numéricas. Allí se vio que algunas variables estaban fuertemente correlacionadas, como los años dentro de la empresa y los años en el mismo departamento, y el salario mensual y el salario por hora. Resultó importante identificar estas correlaciones para poder interpretar mejor los resultados y evitar posibles errores introducidos al utilizar variables colineales, pero en un primer momento no fueron removidas.

A su vez, se realizó un test de Hotelling. Al obtenerse un p-valor menor a 0.01, puede afirmarse que existen diferencias en la media multivariada entre ambos grupos.

Se realizaron tests de homogeneidad de la matriz de covarianza (Box's M) y de normalidad multivariada (Shapiro-Wilk multivariado). En ambos casos, el test dio significativo (p-valor menor a 0.01), por lo que deben rechazarse las hipótesis de la normalidad multivariada y homogeneidad de matriz de covarianza.

Primera Parte: Clasificación Supervisada

Para la clasificación supervisada, se utilizaron los métodos de Análisis Discriminante y Máquina de Soporte Vectorial.

Previo al entrenamiento del modelo, se realizó una partición de los datos en sets de entrenamiento y validación, con el 80% de las observaciones perteneciendo al set de entrenamiento con el cuál se generaron los modelo y el 20% restante perteneciendo al set de validación, a partir del cual se obtuvieron las métricas de clasificación de cada modelo.

Aún considerando la no normalidad ni homocedasticidad de los datos, resultó interesante realizar análisis discriminantes para evaluar la clasificación de los datos, aunque su validez no sea la ideal

Dentro del análisis discriminte, se probaron 5 modelos, seleccionando diferentes grupos de variables. Las mismas fueron previamente centradas y escaladas de ser necesario

El primer modelo tomó a *Attrition* como variable respuesta y a todas las variables restantes como variables explicativas. Probó ser el mejor modelo de los evaluados, con un 76% de las observaciones bien clasificadas. Sin embargo, resultó interesante buscar modelos más simples y con menos variables que pudieran clasificar similarmente bien.

Luego se analizó un segundo modelo, que incluyó a las variables numéricas consideradas como más importantes para el análisis de componentes principales como variables explicativas. Al realizar el análisis de componentes principales con el total de las variables, el primer componente había explicado el 35.7% de la variación entre las observaciones, por lo que se consideró que tomar aquellas variables cuya contribución al mismo fuera mayor que 7 era una buena aproximación a la reducción de variables explicativas. Se conservaron 7 variables, (*Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion* y *YearsWithCurrManager*), y el modelo logró clasificar el 64% de las observaciones del set de validación correctamente.

El tercer modelo se basó en tomar aquellas variables que habían resultado significativamente diferentes entre ambos grupos en el análisis univariado, según lo analizado en la Tabla I. Se consideró aquellas variables con un p-valor menor a 0.05, y se conservaron 14 variables. Este modelo logró clasificar un 72.6% de las observaciones correctamente.

Deseando reducir aun más el número de variables evaluadas, se tomaron los resultados del modelo 3 y, con las variables con mayores coeficientes absolutos en el discriminante lineal único, se generó un cuarto modelo, que incluyó 9 variables (*BusinessTravel, Department, JobInvolvement, JobSatisfaction, MaritalStatus, OverTime, YearsAtCompany, YearsInCurrentRole* y *JobLevel*) y clasificó correctamente al 73.5% de las observaciones.

Por último, a modo de comparación, se generó un último modelo tomando aquellas variables que uno, a priori, podría considerar como buenos clasificadores de si una persona va a dejar o no su empleo. Se consideraron las variables *MonthlyIncome*, *JobSatisfaction*, *OverTime*, y *YearsAtCompany.* Este reducido número de variables, que tampoco incluyeron a variables altamente correlacionadas, pudo clasificar correctamente a solo el 62% de las observaciones del set de validación.

Considerando la no idoneidad previamente mencionada de los datos para un análisis que tenga como supuestos la homogeneidad de matriz de covarianzas y normalidad multivariada, se realizó también un análisis más robusto que no tuviera dichos supuestos, como ser una máquina de soporte vectorial.

Se generaron dos modelos, uno incluyendo todas las variables como explicativas, y otro utilizando las variables del modelo con mejor performance para el análisis discriminante (modelo 4).

En ambos casos, los modelos resultaron tener una performance aceptable. Ambos modelos de SVM tuvieron valores idéntico a aquel obtenido por el modelo con mejor performance del análisis discriminante, lo que es remarcable considerando la previamente mencionada falta de supuestos del SVM.

Una vez realizados los modelos, se comparó la performance entre los 5 modelos de análisis discriminante y los dos realizados utilizando SVM. En la Tabla II, puede verse la proporción de casos correctamente clasificados por cada modelo en el set de validación.

Tabla II: Porcentaje de clasificación correcta de cada modelo evaluado.

Modelo 1 AD	Modelo 2 AD	Modelo 3 AD	Modelo 4 AD	Modelo 5 AD	Modelo 1 SVM	Modelo 2 SVM
0.7607	0.641	0.7265	0.735	0.6239	0.735	0.735

Se analizó también la matriz de confusión para cada modelo. En su mayoría, tendieron a clasificar más individuos en la clase más representada, “No”, lo que generó un gran porcentaje de falsos negativos.

Para estudiar con mayor profundidad los casos mal clasificados, se generó una tabla con los mismos. Allí pudo verse que, de los 117 casos del set de validación, el 63 (53.8%) resultaron mal clasificados por alguno de los modelos.

Al analizar dichos casos con mayor profundidad, se descubrió que 31 de los 63 casos se encontraban mal clasificados en al menos 4 de los modelos analizados, lo que hace suponer que se trata de casos con un comportamiento realmente atípico que los modelos no supieron identificar.

Segunda Parte: Clasificación No Supervisada

Para la clasificación no supervisada, se realizaron análisis de clústers jerárquicos y no jerárquicos.

Se comenzó realizando el análisis con solo las variables numéricas. Como se había visto previamente, las variables originales contenían un número considerable de variables con alta correlación entre sí, por lo que se realizó una selección de las mismas, eliminando aquellas muy correlacionadas con muchas variables (como *MonthlyIncome*) y seleccionando sólo una de las variables de años (*YearsInCurrentRole*).

Posteriormente, se escalaron los datos y se generó la matriz de distancias, eligiendo la distancia euclideana para el cálculo de la misma. Para la generación de clusters jerárquicos se decidió comenzar con un enfoque aglomerativo. Se generaron una serie de árboles utilizando todos los métodos de aglomeración disponibles, y se calculó para cada uno el índice de correlación cofenética. El mismo no presentó valores demasiado altos en ninguno de los casos, pero consistentemente tuvo mejores valores en los métodos *single* y *average*. A pesar de ser ligeramente mayor el de single, se decidió utilizar el método del promedio para los análisis.

Tabla III: comparación del índice de correlación cofenética para cada uno de los métodos

complete	single	centroid	average	ward.D
0.4712	0.5974	0.5199	0.5961	0.3284

A continuación, se utilizaron diversos métodos para seleccionar el número óptimo de clusters. La mayoría coincidió que 7 era el número adecuado.

Para comprobar que este número fuera efectivamente el óptimo, se realizó el análisis para 2, 5, 7, 8 y 9 clústers.

Resultó interesante ver que un gran número de observaciones centrales no se separaban a pesar del aumento del número de clústers, si no que los valores extremos se dividieron cada vez en grupos más pequeños.

Al utilizar el índice de Dunn para cuantificar la idoneidad del método, se obtuvo un valor de 0.0034, que, considerando que se busca maximizar el valor del índice, no es exactamente un buen resultado.

Como se mencionó previamente, pudo notarse que se veían, a grandes rasgos, dos grupos: un gran grupo con la mayoría de las observaciones, posible de subdividir en grupos más pequeños pero generalmente muy similares, y otro grupo más pequeño, heterogéneo entre sí pero lo suficientemente diferente al grupo anterior como para resultar agrupado por descarte.

Al analizar estos casos, puede observarse que se suele tratar de personas con una edad mayor a la mediana general, por lo general con más años en la empresa. Curiosamente, también se encontró que tenían salarios mensuales más bajos que los generales (a pesar de su mayor experiencia), el cual puede ser uno de los motivos por los que se diferencian.

Tabla IV: comparación de las variables numéricas entre el grupo general y el especial

	Casos generales ***N=572***	**Casos especiales** ***N=15***
Age	34.0 [29.0;41.0]	56.0 [55.0;58.0]
DistanceFromHome	7.00 [2.00;15.2]	14.0 [6.50;24.0]
HourlyRate	67.0 [49.0;85.0]	80.0 [66.0;88.5]
MonthlyRate	14564 [8442;20594]	11924 [3744;15820]
NumCompaniesWorked	1.00 [1.00;4.00]	7.00 [5.00;8.50]
TotalWorkingYears	9.00 [5.00;13.0]	31.0 [16.0;36.0]
TrainingTimesLastYear	3.00 [2.00;3.00]	2.00 [1.00;3.00]
YearsInCurrentRole	2.00 [1.00;7.00]	9.00 [3.50;11.5]

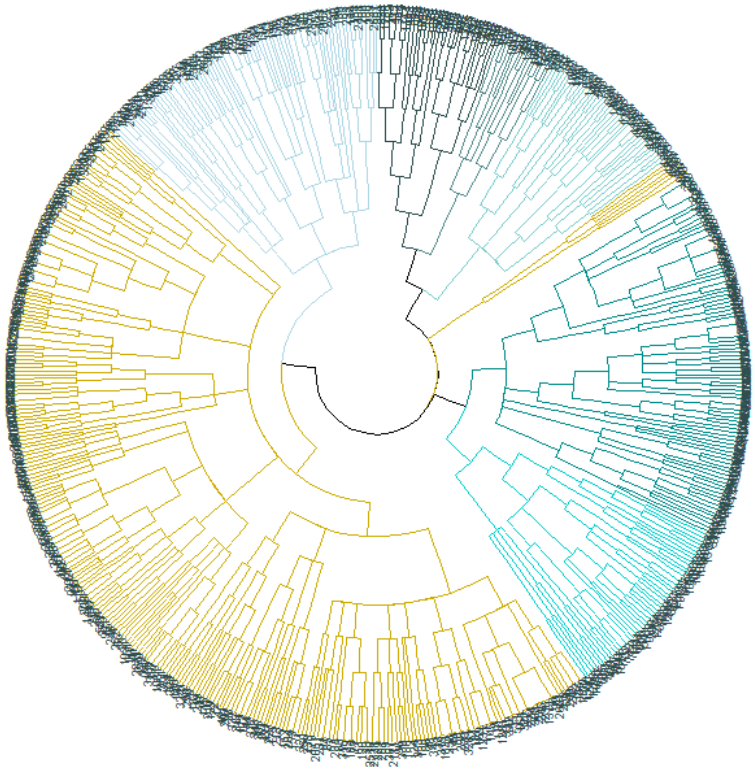
Se realizó nuevamente el análisis de clusters jerárquicos removiendo los 15 casos especiales, pero dicho cambio no mejoró la división de las observaciones y se siguió con el mismo problema de los casos extremos.

Algo similar ocurrió al realizar el análisis por k-means, seleccionando k=8. El método no logró separar realmente a los grupos, viéndose muchas observaciones superpuestas con otras pertenecientes a un clúster diferente.

Es por eso que, aunque en un caso ideal resultaría sumamente interesante estudiar las variables para cada clúster, en este se consideró mejor estudiar el común de las observaciones que no podían separarse versus las observaciones que eran tan diferentes que en algunos análisis constituían su propio clúster.

Para el análisis de las variables categóricas, se utilizó la distancia de Gower como medida de dissimilitud de las observaciones. A su vez, se utilizó un algoritmo divisivo tipo DIANA, aprovechando que se trataba en su mayoría de variables con 2 o 3 categorías.

Figura 1: clusters generados con las variables categóricas



Aún así, puede verse en la Figura 1 que nuevamente se observó el efecto de un clúster conteniendo a la gran mayoría de las observaciones, mientras que otros contienen muy pocas observaciones. Se consideró que explica que la mayoría de los empleados tienen combinaciones predecibles de las variables de estudio, y que aquellos que no son los que resaltan al intentar clasificarlos con estos métodos.

Conclusiones

La premisa inicial de este trabajo fue que la decisión de dejar o no un empleo depende de una multitud de razones.

A lo largo del desarrollo del mismo, se aplicaron diversos tipos de algoritmos de clasificación para intentar identificar y dilucidar estos motivos.

A pesar de no haber encontrado métodos de clasificación ideales para el problema, se generaron aproximaciones, como los obtenidos por la máquina de soporte vectorial, clasificando cerca del 75% de las observaciones correctamente.

Por otra parte, surgieron otros posibles temas de estudio en los análisis de clústering, dando a entender que, independientemente de si finalmente decidían abandonar un trabajo o no, la mayoría de los empleados eran relativamente parecidos.

Como conclusión general, podemos reafirmar la suposición de que dejar un empleo depende de una multitud de factores, agregando ahora que los mismos no siempre pueden ser representados con datos y plasmados de una manera fácilmente comprensible y reproducible. También puede remarcarse que las personas más jóvenes, con menos años en la empresa y menores salarios tienden a dejar su empleo con mayor frecuencia.

Resultaría interesante probar otros métodos de análisis sobre los datos, para intentar delucidar algún otro patrón no visto.