Universidad de Buenos Aires
Facultad De Ciencias Exactas y Naturales
Maestría en explotación de datos y descubrimiento del conocimiento

Año 2020 - 1er Cuatrimestre

Análisis Inteligente de Datos

TRABAJO PRÁCTICO INTEGRADOR
TEMA: Clasificación
FECHA: 01 de septiembre de 2020

INTEGRANTES:
Massaro Rocca, Patricio Nicolás
<patomassaro@gmail.com>

Índice

1. Clasificación Supervisada

En esta sección, se realizará una clasificación supervisada de la base de *próstata*, que posee datos de un estudio médico del cáncer de próstata. El objetivo de la clasificación es predecir la ruptura capsular (variable CAPSULE).

Los datos fueron cargados en R y se realizó un sampleo mediante una función creada en R, tomando el 80 % del contenido del dataset en forma aleatoria usando el DNI como semilla.

Los algoritmos que se utilizaron para la clasificación son los de Regresión Logística y Máquinas de soporte vectorial.

1.1. Análisis de variables

En primer lugar, se hizo una inspección de los datos para encontrar valores anómalos o faltantes, se observo que tres individuos poseían valores faltantes en la columna GLEASON y valores prohibidos en la columna DCAPS. Dado que el conjunto es de alrededor de 300 elementos, se decidió eliminar estos registros.

La columna VOL, que representa el volumen de la próstata, contiene una cantidad importante de valores en 0. Esto puede querer decir que la próstata fue extirpada o puede ser un error en la carga. Para dilucidar este problema, se observó que ocurre con la ruptura capsular para los individuos con volumen igual a 0. Por otro lado, se evaluó como se comporta la variable objetivo para los individuos con un volumen distinto de cero. Al no observarse un comportamiento destacado y al no conocer la naturaleza de la variable, se decidió eliminarla.

Mediante Boxplots y tablas de frecuencia, se observaron las siguientes cuestiones:

- Los individuos con ruptura capsular presentan una mediana de PSA mas grande.
- Los individuos con ruptura capsular tienden a tener un valor de GLEASON más grande.
- No parece haber diferencias significativas en la edad entre los dos grupos.
- La proporción de individuos de esta población con rotura capsular para cada una de las razas es similar.
- Los resultados del examen prostático parecen influir en la rotura capsular. Cuando el resultado es que no hay nódulos, hay muchas mas individuos sin rotura que con ella. Esa relación se invierte para el caso en donde hay nódulos a ambos lados. En los casos de nódulos solo a la izquierda o a la derecha, las relaciones de positivos y negativos se encuentra en valores medios respecto de los extremos.

Por otro lado, para las variables numéricas se realizó un análisis de correlación, en donde se observaron coeficientes menores a 0.1. Este resultado muestra que no hay una interacción muy profunda entre las variables predictoras numéricas.

Se hizo un análisis de normalidad de las variables numéricas en forma gráfica o mediante test de Anderson-Darling. En todos los casos, el *p-value* resulto ser menor a 0.001, rechazando la normalidad de los datos.

1.2. Algoritmos de clasificación

Para probar la bondad de clasificadores, se utilizó el método de clasificación no ingenua, en donde se separó el conjunto de datos en dos partes, una para entrenamiento y la otra para validación. La proporción elegida fue de 80/20

1.2.1. Regresión logística

El primer algoritmo utilizado es el de Regresión logística, la principal ventaja de este método es que permite ponderar la importancia de la relación entre las variables predictoras y la variable dependiente (ruptura capsular). En primer lugar, se decidió aplicar el algoritmo utilizando todas las variables predictoras posibles, para poder entender la relación de cada una de ellas con la independiente. Se observó que como se anticipaba en la sección 1.1, las variables GLEASON, DPROS, DCAPS y PSA son las que mayor relación guardan con CAPSULE. El algoritmo fue ajustado con los datos de entrenamiento para luego ser puesto a prueba. La tabla de confusión de la predicción contra los datos de validación puede verse a continuación

		Referencia	
		0	1
Predicción	0	23	5
	1	9	23

Cuadro 1.1: Tabla de confusión para algoritmo de regresión logística

El algoritmo pudo clasificar correctamente un 75 % de los casos. Sin embargo, tratándose de un estudio médico, resulta problemático que en 5 casos el paciente tenía ruptura capsular cuando el algoritmo predijo que no. Es muy común en este tipo de casos médicos que se elija ser un poco más laxo a la hora de tomar un caso como positivo, dado que es posible hacer pruebas posteriores para determinar mejor el problema. Esto se puede hacer modificando el umbral en donde el algoritmo decide que un caso implica ruptura capsular. En este caso, el umbral fue corrido a 0,11 en lugar de 0,5. De esta forma, la nueva matriz de confusión se observa en el cuadro 1.2. A costa de mayores falsos positivos y menor porcentaje de aciertos, se consiguió reducir de 5 a 1 los falsos negativos. El umbral podría seguir disminuyendo, pero llega un momento en que el clasificador toma a todos los casos como positivos y se pierde el poder predictivo del clasificador.

Con el propósito de simplificar el modelo, se implementó un algoritmo paso a paso, en donde se parte del modelo completo y se eliminan variables que no poseen un poder predictivo alto. En concordancia con lo explicado anteriormente, al eliminar las variables

		Referencia	
		0	1
Predicción	0	13	1
	1	19	27

Cuadro 1.2: Tabla de confusión para algoritmo de regresión logística, ajustando el umbral de detección.

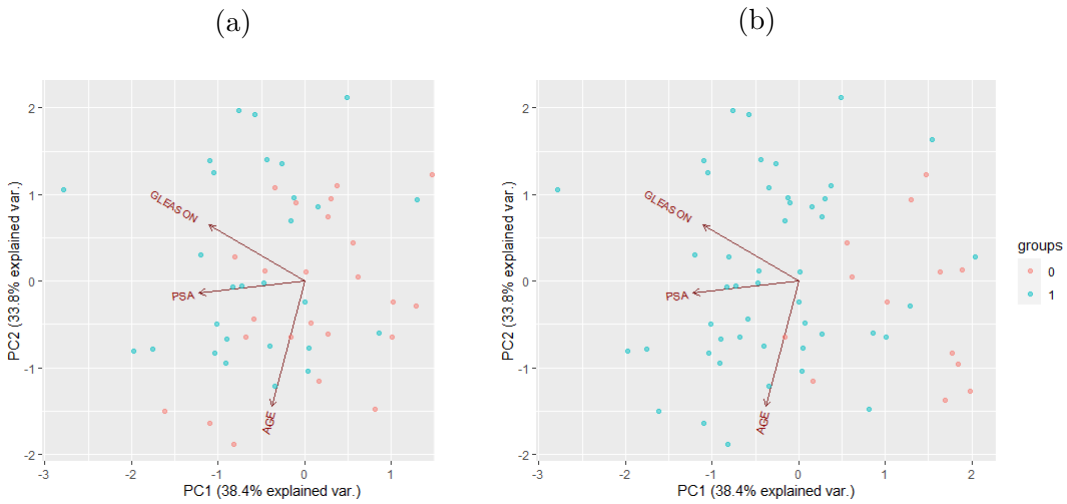

Figura 1.1: Biplot del conjunto de datos, agrupado por : (a) Valores actuales, (b) predicción del algoritmo con el umbral modificado.

de edad y raza no se encontró una degradación grande en la performance, permitiendo simplificar el modelo sin comprometer la bondad de la clasificación.

1.2.2. Máquinas de soporte vectorial

Dado que este tipo de algoritmos está dentro de la categoría de clasificadores lineales, resulta interesante evaluar la separabilidad lineal del conjunto de datos estudiado. Utilizando los datos numéricos, se puede observar en el biplot (a) de 1.1 que los conjuntos no son linealmente separables, por lo que será necesario el uso de un kernel de mapeo. Se probaron distintos kernels hasta encontrar el que mejores resultados produce.

La tabla de confusión se muestra en la tabla 1.3, se consiguió acertar el pronóstico en un 71,5 % de los casos, observándose una performance ligeramente inferior al caso de regresión logística.

		Referencia	
		0	1
Predicción	0	21	6
	1	11	22

Cuadro 1.3: Tabla de confusión para algoritmo de máquina de soporte vectorial.

1.3. Conclusión

A partir de las pruebas y los datos obtenidos, el algoritmo más adecuado para ser usado es el de regresión logística. La posibilidad de entender la relación de las variables predictoras y la variable dependiente es muy importante. Por otro lado, se obtuvo una performance ligeramente superior y se puede modificar fácilmente el umbral de detección para evitar falsos negativos, un fenómeno que puede ser muy peligroso en medicina.

2. Clasificación no Supervisada

En esta sección, se realizará una clasificación no supervisada de la base de *seguros*, que posee datos de pólizas contratadas. El objetivo es encontrar grupos que compartan características.

Los datos fueron cargados en R y se realizó un sampleo mediante una función creada en R, tomando el 75 % del contenido del dataset en forma aleatoria usando el DNI como semilla.

2.1. Análisis de variables

El conjunto de datos consiste en 6 variables numéricas y 2 variables categóricas. A continuación, se presenta un análisis realizado a las variables numéricas, observando las diferencias entre las distintas categorías de las variables categóricas. El detalle de los boxplots se puede ver en el código anexo.

	Outliers	Sexo	Region
Edad	No presenta	No se observan grandes diferencias	Mediana menor en región 1
BMI	0,6 %	No se observan grandes diferencias	No se observan grandes diferencias
Hijos	18 %	No se observan grandes diferencias	No se observan grandes diferencias, Región 0 posee más outliers
Fuma	No presenta	No se observan grandes diferencias	No se observan grandes diferencias
Cargos	No presenta	Sexo 0 posee mediana más grande	Región 0 posee mediana más grande
Prima del seguro	8 %	No se observan grandes diferencias	Región 1 posee mediana mucho mas grande (casi 4 veces)

Cuadro 2.1: Análisis de las variables del conjunto de datos, teniendo en cuenta las categorías.

Las variables Hijos y Prima del seguro poseen una importante cantidad de outliers. Dado que los métodos de clasificación no supervisada son muy susceptibles a éstos, se decidió eliminar los elementos más alejados del rango intercuartil, considerados outliers severos. En este trabajo se presentan los resultados aplicando este procedimiento de limpieza. Sin embargo, el análisis de agrupamientos se ejecutó también con el conjunto de datos completo para comprobar los efectos de los outliers.

Adicionalmente, se realizó un análisis de correlación de las variables numéricas. No se encontraron relaciones muy fuertes, el mayor coeficiente encontrado fue entre la variable hijos y la variable cargos, con un valor de 0.46.

2.2. Algoritmos de agrupamiento

Los algoritmos que se utilizaron para la clasificación son los de *k-means* y jerárquicos, a continuación se muestra el detalle de cada uno de ellos.

2.2.1. Cantidad de Clusters y métrica de distancia

Para aplicar una técnica de agrupamiento, es necesario determinar el número óptimo de clusters a utilizar. Se utilizó la función *fviz_nbclust* para determinar el número óptimo de clusters, empleando el criterio de maximizar la métrica *mean silhouette*.

En el caso de *k-means*, el criterio establece que el número óptimo de clusters es de 9. Sin embargo, se observa que 5 clusters posee un valor prácticamente igual y priorizando

la interpretabilidad de los resultados, se eligió ese número. En el caso de clustering jerárquico, el número de clusters óptimo coincide con el caso anterior.

La métrica de distancia fue elegida con el mismo criterio en el caso de clustering jerárquico, se buscó la combinacion que maximice a *mean silhouette*, siendo en este caso la distancia de Manhattan. En el caso de *k-means*, fue elegida la distancia euclídea.

2.3. Asignación y análisis de clusters

Para este análisis, se decidió utilizar el algoritmo jerárquico, utilizando el método de ward, los clusters fueron asignados y se realizó un biplot que se muestra a continuación.

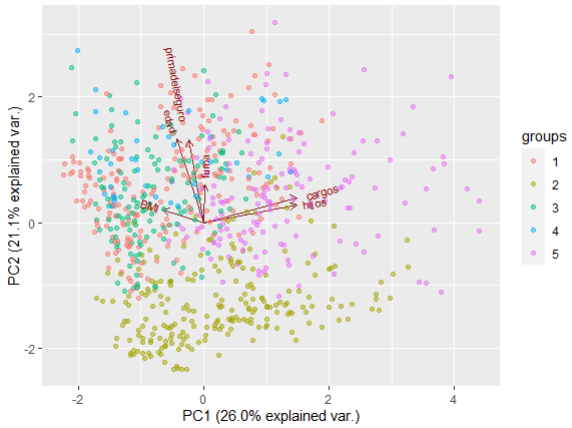

Figura 2.1: Agrupamiento jerárquico mostrado en un biplot con las componentes principales del conjunto de datos

Es importante resaltar que las componentes más influyentes del biplot captan menos del 50 % de la variabilidad. Aunque es la mejor manera de ver los efectos del agrupamiento, es muy posible que varios aspectos del mismo no se vean reflejados en el gráfico. Debido a esto, se presenta la tabla 2.2, en donde se analiza los valores medios de las características de los individuos que componen los grupos. Para las variables categóricas, al poseer solo dos categorías, se las convirtió en numéricas para entender la proporción de individuos pertenecientes a cada clase al calcular la media.

hcluster	edad	sexo	BMI	hijos	fuma	region	cargos	primadelseguro
1	52.46114	0.492228	29.55959	25.35233	0.3367876	0.134715	2.051813	13999.513
2	24.07556	0.4977778	29.14222	38.73333	0.32	0.1244444	2.315556	5438.284
3	44.0303	0.5151515	33.68182	19.18182	2.1818182	0.0530303	1.848485	10195.205
4	25.97674	0.6511628	33.11628	29.7907	0.6744186	0.9302326	2.093023	35628.674
5	38.76667	0.5	27.69333	91.03333	2.1333333	0.1	3.106667	9760.567

Cuadro 2.2: Valor medio de las variables en cada uno de los clusters

Puede observarse en la tabla que la proporción de sexos en los grupos se mantiene pareja. En cuanto a la región, se observa que el grupo 4 posee una alta proporción de individuos pertenecientes a la región 2, con un valor de prima muy alto. Estas características pueden verse reflejadas en lo descrito en la tabla 2.1.

El algoritmo estableció cuatro segmentos distintos en la prima del seguro, a continuación se realiza un análisis de cada uno de los grupos.

- Prima del seguro más baja (grupo 2): Franja de edad joven, no fumadores y con valores promedio de BMI, gastos en hijos y cantidad de cargos promedio, en su mayoría de la región 0.
- Prima del seguro media (grupo 5) : Franja de edad media, fumadores con alto gasto en hijos y muchos cargos, pero con bajo indice BMI, pertenecientes a la región 0 en su mayoría.
- Prima del seguro media (grupo 4) : Franja de edad media-grande, fumadores, poco gasto en hijos y pocos cargos, pero con alto indice BMI, pertenecientes a la región 0 en su mayoría.
- Prima del seguro alta (grupo 1) : Franja de edad grande, no fumadores, con gasto en hijos, cargos e indice BMI promedio, pertenecientes en su mayoría a la región 0.
- Prima del seguro muy alta (grupo 4): Franja de edad joven, no fumadores, promedio en hijos y cantidad de cargos, alto BMI y pertenecientes en su mayoría a la región 1.