**Análisis de la aplicación de un modelo de regresión logística para clasificar trastornos del hígado y cáncer de pulmón**

**Huamani Loredo, Alvaro**

20151966@aloe.ulima.edu.pe

Universidad de Lima

# INTRODUCCIÓN

El presente documento trata del análisis de dos datasets, el primero está relacionado con trastornos del hígado (“UCI Machine Learning Repository: Lung Cancer Data Set,” 2015) y el segundo relacionado con el cáncer de pulmón (Forsyth, 1990). Para lograr obtener un modelo que clasifique los mencionados dataset de manera satisfactoria, se hará uso de la regresión logística; se hace uso de métodos de imputación para manejar los valores nulos, específicamente con el dataset de cáncer de pulmón. Los modelos resultantes serán validados usando una matriz de confusión y el método de validación cruzada; para mejorar los resultados de estas pruebas se emplearán métodos de normalización y de la Lasso.

# BASES TEÓRICAS

## Regresión logística

Según (Webb et al., 2011) la regresión logística proporciona un mecanismo para aplicar las técnicas de regresión lineal a problemas de clasificación; este modelo es aplicable a los dataset relacionados con trastornos del hígado (“UCI Machine Learning Repository: Lung Cancer Data Set,” 2015) y al cáncer de pulmón (Forsyth, 1990). Debido a que se espera clasificar esta data las variables dependientes deben ser categóricas, pero en el dataset de trastornos del hígado la variable dependiente propuesta en la información del dataset es la correspondiente a la columna “drinks”, esta variable resulta ser continua, es por esto que se necesita un método para convertir esta variable a categórica.

## Dicotomizar variables

Para convertir la variable “drinks” se usará la información propuesta por (Mcdermott, 2016), donde se indica que, tras un acertado análisis, lo “correcto” es usar esta columna como variable dependiente pero dicotomizado. Dicotomizar variables no es otra cosa que aplicar una condicional que agrupará un conjunto de datos continuos.

## Evaluación del modelo

Para evaluar el modelo de regresión logística se usará la matriz de confusión, esta según (Ting, 2017) resume el desempeño de un clasificador con respecto a algunos datos de prueba. Es una matriz bidimensional, indexada en una dimensión por la clase verdadera de un objeto y en la otra por la clase que asigna el clasificador. Para un resultado más preciso se usará el método cross-validation, que (Refaeilzadeh, Tang, & Liu, 2009) lo define como un método estadístico para evaluar y comparar algoritmos de aprendizaje dividiendo los datos en dos segmentos: uno que se usa para aprender o entrenar un modelo y el otro que se usa para validar el modelo.

## Métodos para mejorar la exactitud del modelo

Para mejorar la exactitud del modelo se hará uso del análisis de la aplicación de una matriz de correlación y del método Lasso. El método Lasso realiza una contracción L1 de modo que haya "esquinas" en la restricción, que en dos dimensiones corresponde a un diamante. Si la suma de los cuadrados "golpea" una de estas esquinas, entonces el coeficiente correspondiente al eje se reduce a cero. (“5.4 - The Lasso | STAT 508,” n.d.). El coeficiente de correlación de Pearson (r) es una medida de la asociación lineal de dos variables; los valores del coeficiente de correlación varían de –1 a +1, donde los valores positivos del coeficiente de correlación indican las dos variables tienden a aumentar o disminuir de igual forma. (“Pearson’s Correlation Coefficient,” 2008)

# METODOLOGÍA Y EXPERIMENTACIÓN

Para la experimentación se hizo uso de la librería scikit learn de Python.

## Liver Disorders Data Set

Siguiendo la información proporcionada por el dataset, las variables independientes usadas son “mcv”, “alkphos”, “sgpt”, “sgot“y “gammagt”. La variable dependiente será “drinks”, pero siguiendo la información proporcionada por (Mcdermott, 2016), se usará dicotomizado como > 3, entonces este informe considera que los valores > 3 serán tipo 2 y los demás tipo 1.



Fig. 1. Conteo de datos para la variable “drinks”.

Debido que la información proporcionada por el dataset indica que estos datos no contienen valores nulos se realizará la primera prueba implementado el modelo de regresión logística. Sin aplicar el método de cross-validation el modelo consigue un accuracy de 0.64

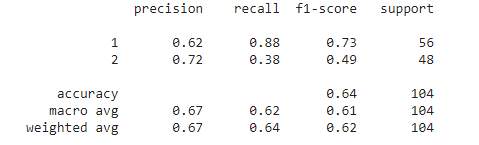


Fig. 2. Evaluación del primer model

Aplicando el método de cross-validation se obtiene una media de 0.643.

Este primer resultado indica que el modelo no tiene una buena precisión, haciendo un análisis a la correlación de sus variables se puede determinar que existe una correlación considerable en las variables “sgpt” y “sgot” como se muestra en la figura 3.

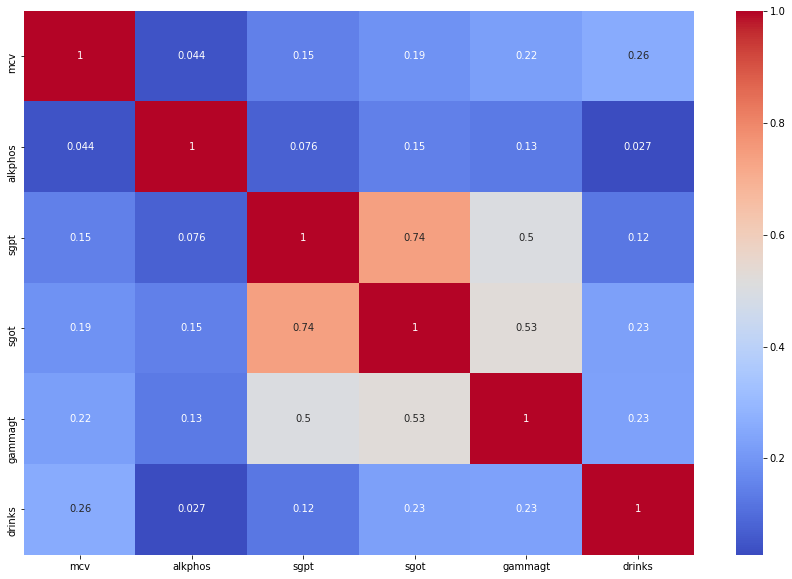


Fig. 3. Mapa de calor de la correlacionando entre variables

Analizando la figura 4 se puede comprobar que existe una considerable correlación.

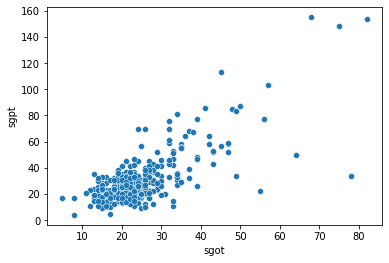


Fig. 4. Scatterplot de las variables “sgpt” y “sgot”

Eliminando las variables "sgot" y "sgpt" el resultado del cross-validation tras aplicar el modelo es de 0.641, eliminando únicamente la variable "sgot" se obtiene 0.638 y eliminando únicamente la variable "sgpt" se obtiene 0.632

Analizando los resultados se puede concluir que esté método no ayudo a mejorar la exactitud del modelo.

Aplicando el método Lasso se puede obtener la figura 5, se puede considerar eliminar las variables que se aproximan a 0 y las variables negativas.

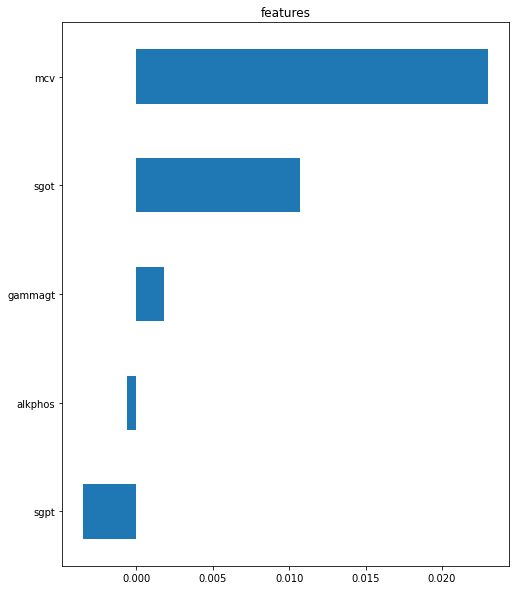
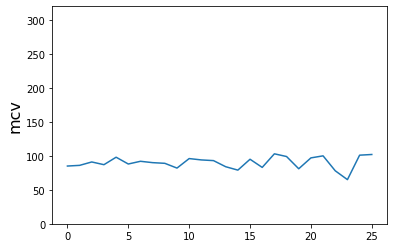
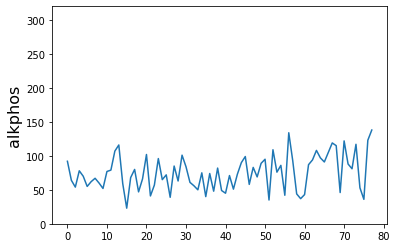
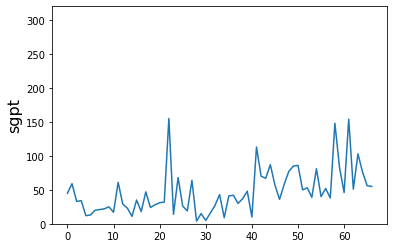
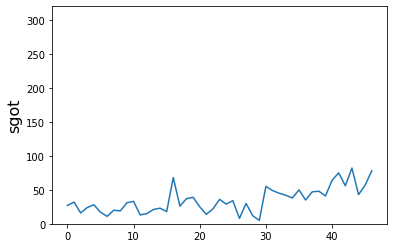


Fig. 5. Gráfico donde se muestran los coeficientes obtenidos de la regresión Lasso

Eliminando las variables con resultado negativo "alkphos" y "sgpt" el resultado del cross-validation tras aplicar el modelo es de 0.649

Para analizar los picos en las variables se usó un gráfico de líneas para todas las variables independientes, con el fin de analizar los datos únicos.

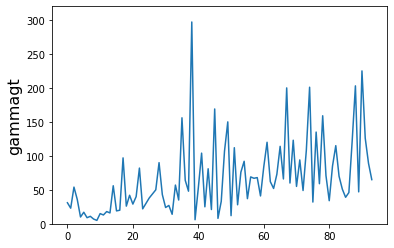


Fig. 6. Gráfico de líneas para los datos únicos de las variables independientes

Analizando la figura 6 se puede observar que las variables “alkphos” , “sgpt”, “sgot” y “gammagt” tienen picos altos, considerando un rango de 0 a 300.

Tras aplicar el método de normalización en la variable “alkphos” y “gammat” se obtuvo una mejora, siendo el resultado del cross-validation tras aplicar el modelo es de 0.652.

## Lung Cancer Data Set

Siguiendo la información del dataset, la variable dependiente será la columna 0 y todas las demás columnas serán las independientes.

La información del dataset indica que existen valores nulos representados por “?” en el dataset.



Fig. 7. Mapa de calor de los datos nulos en las variables independientes

Como se puede observar en la figura 7, las líneas amarillas representan los datos nulos, con esto se puede determinar que la columna 4 y las columnas 38 contienen datos nulos.

Para una primera prueba del modelo de regresión logística se eliminarán las columnas que contengan valores nulos, el resultado del cross-validation tras aplicar el modelo es de 0.408.

Buscando mejorar el modelo, se propone rescatar la columna 38 aplicando el método de imputación, siguiendo la estrategia de la más frecuente, tras esto el resultado sigue siendo el mismo.

Aplicando el método Lasso para seleccionar variables, se rescatan aquellas que no den 0 como resultado de la regresión, siendo estas las columnas 2, 6, 10, 13, 14, 19, 20, 23, 27, 33, 34 y 37, aplicando el modelo se obtiene como resultado del cross-validation 0.633

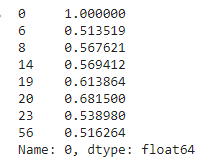


Fig. 8. Variables independientes que no obtienen 0 como resultado de la regresión Lasso

Analizando la correlación de las variables independientes con la variable dependiente se propone seleccionar las variables que tengan una correlación superior a 0.5 con la variable dependiente.

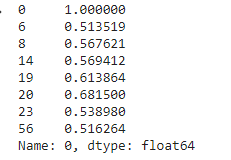


Fig. 9. Variables independientes que tienen una correlación mayor a 0.5 con la variable dependiente

Usando únicamente estas variables, al aplicar el modelo se obtiene como resultado del cross-validation 0.658

Aplicando una unión de estas variables, es decir usando las columnas 2, 6, 8, 10, 13, 14, 19, 20, 23, 27, 33, 34, 37 y 56, el resultado del cross-validation es 0.633.

Analizando los datos de las columnas, se puede determinar que la gran mayoría tiene 3 valores, existen algunos que tienen 2 y otros con 4.

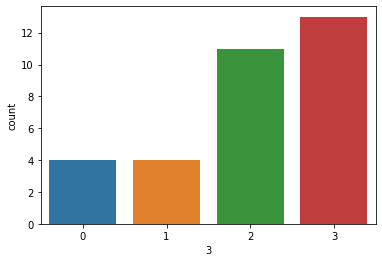
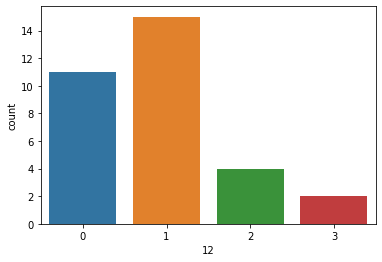


Fig. 10. Variables independientes que tienen 4 valores únicos

Existen 2 casos (col 3 y col 12) donde se tienen 4 tipos, para estas variables se aplicará el método de One Hot Encoding, con esto el resultado de aplicar cross-validation al modelo resultante es 0.5.

Aplicando las variables con One Hot Encoding junto con la una unión de columnas que obtenidas del análisis del método Lasso y de Correlación se obtiene un modelo que tiene como resultado del cross-validation 0.675, este es el mejor resultado hasta ahora.

Aplicando One Hot Encoding a todas las variables, método no recomendado debido a que se van a generar muchas variables independientes, se obtiene un cross-validation de 0.366, este es el resultado obtenido más bajo.

# CONCLUSIONES

Después de analizar varios métodos que buscan mejorar la exactitud del modelo de regresión logística, el mejor resultado obtenido para el primer dataset (Liver Disorders) es de un cross-validation de 0.652. En el segundo data set el mejor resultado es 0.675. Se puede concluir que los modelos obtenidos no son precios. Si bien los resultados no son mejores, considerando que los datasets son pequeños, se puede afirmar que con una mayor cantidad de datos los modelos obtenidos van a mejorar, también se puede probar aplicando otros modelos de clasificación como KNN o Random Forest, modelos mencionados en algunos papers consultados para la elaboración de este informe.

**Link del repositorio:**

https://github.com/alvaroenrique/ML\_course-project1

**REFERENCIAS**

5.4 - The Lasso | STAT 508. (n.d.). Retrieved October 16, 2020, from https://online.stat.psu.edu/stat508/lesson/5/5.4

Forsyth, R. S. (1990). UCI Machine Learning Repository: Liver Disorders Data Set. Retrieved October 16, 2020, from BUPA Medical Research Ltd. website: https://archive.ics.uci.edu/ml/datasets/liver+disorders

Mcdermott, J. (2016). *Diagnosing a disorder in a classification benchmark*. https://doi.org/10.1016/j.patrec.2016.01.004

Pearson’s Correlation Coefficient. (2008). In *Encyclopedia of Public Health* (pp. 1090–1091). https://doi.org/10.1007/978-1-4020-5614-7\_2569

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In *Encyclopedia of Database Systems* (pp. 532–538). https://doi.org/10.1007/978-0-387-39940-9\_565

Ting, K. M. (2017). Confusion Matrix. In *Encyclopedia of Machine Learning and Data Mining* (pp. 260–260). https://doi.org/10.1007/978-1-4899-7687-1\_50

UCI Machine Learning Repository: Lung Cancer Data Set. (2015). Retrieved October 16, 2020, from University of California, Irvine website: http://archive.ics.uci.edu/ml/datasets/Lung+Cancer

Webb, G. I., Sammut, C., Perlich, C., Horváth, T., Wrobel, S., Korb, K. B., … Raedt, L. De. (2011). Logistic Regression. In *Encyclopedia of Machine Learning* (pp. 631–631). https://doi.org/10.1007/978-0-387-30164-8\_493