PREDICCIÓN HEPATITIS

Julia García Flores, Paula Yuan César Aguilar, Juan María Sojo López, Pablo Santos Alises, Cristian Antonio Cabello Arango

Escuela Técnica Superior de Ingeniería Informática

Introducción

La hepatitis es una inflamación del hígado que puede ser causada por diversos. De acuerdo con datos de la Organización Mundial de la Salud (OMS), alrededor de 1,2 millones de personas murieron a causa de hepatitis aguda o alguna de sus secuelas (cáncer o cirrosis) en 2019. En este proyecto, se realizará un análisis fundamental de un conjunto de datos, compuesto por una vista minable de 20 atributos (variables) y 155 ejemplos de cada atributo, que representa distintos hallazgos y pruebas clínicas obtenidas de un estudio sobre pacientes que padecen esta enfermedad.

Preprocesamiento

Paso de variables categóricas a discretas: **Observando el dataset, nos damos cuenta de que todas las variables numéricas son continuas y que todas las variables categóricas son de tipo caracteres. Por ello, decidimos transformar variables categóricas a discretas binarias, de forma que las variables se transformarán a unos y ceros.**

Valores faltantes: El conjunto de datos contiene 167 valores faltantes en total. Como el atributo protime tiene casi la mitad de los valores nulos y la media es de 61 segundos, fuera del rango de valores normales, se decide eliminar.

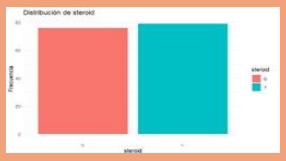
```
          age           sex       steroid     antivirals
            0             0             1              0
      fatigue      malaise     anorexia      liver_big
            1             1             1             10
  liver_firm spleen_palpable      spiders       ascites
           11             5             5             5
     varices     bilirubin alk_phosphate          sgot
            5             6            29             4
     albumin       protime     histology         class
           16            67             0             0
[1] 167
```

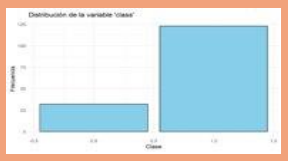
Para el resto de las variables, realizamos un estudio estadístico de los valores de cada atributo en función de su tipo, de la siguiente manera:

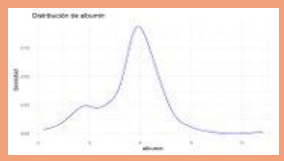
1. ***Para los atributos que han sido transformados de variables categóricas a valores para rellenar los valores faltantes con el valor que más se repita en dicho atributo.***
2. ***Para valores numéricos continuos, se realizó un análisis estadístico de la media y la mediana para reemplazar los valores nulos.***
    1. ***Para 'bilirubin' y 'albumin' se prefirió la media ya que los valores no nulos están en decimales.***
    2. ***Para 'alk_phosphate' y 'sgot' rellenamos con la mediana por estar más cercana a los valores normales.***

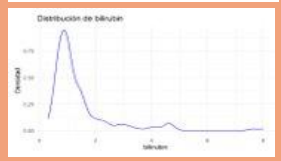
Visualización

En este apartado, hemos representado un análisis gráfico sobre la distribución de cada variable del conjunto de datos. Obtuvimos que hay más variables numéricas discretas y numéricas continuas desbalanceadas, al igual que la variable dependiente 'class' está desbalanceada lo que puede afectar el rendimiento y el aprendizaje de los modelos de aprendizaje automático. Además, las variables continuas presentan una gran diferencia entre sus escalas, por lo que se decidió utilizar una técnica de escalado para concentrar los valores del dataset para poder obtener un mayor número de reconocimiento de patrones.



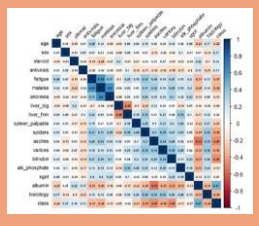






$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Selección de Atributos



Observamos que todas las variables no están muy correlacionadas entre sí y se decidió eliminar tres atributos que apenas presentaban correlación ni con la variable predictoria ni con el resto de variables.

PREDICCIÓN HEPATITIS

Julia García Flores, Paula Yuan César Aguilar, Juan María Sojo López, Pablo Santos Alises, Cristian Antonio Cabello Arango

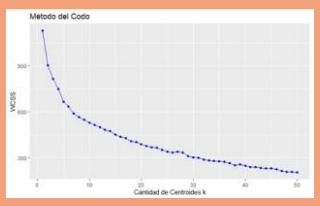
Escuela Técnica Superior de Ingeniería Informática

Clustering

Sabemos que hay dos clases, vive o muere. A continuación usaremos el algoritmo K-Means para agrupar el conjunto de datos. Como podemos observar, si se agrupan en dos clusters, 98 irian al cluster 1 y 57 al cluster 2. En nuestro dataset hay 32 ejemplos correspondientes a "muere" y 123 correspondientes a "vive" y podríamos pensar que pueden estar medianamente bien diferenciados, en concreto, ha "acertado" 118 ejemplos de 155. Sin embargo, si observamos el valor de Between_SS / total_SS es de 19.99%, es un porcentaje muy bajo lo que indica que no hay una buena separación entre clusters.

```
K-means clustering with 2 clusters of sizes 57, 98

Cluster means:
         age        sex   steroid antivirals    fatigue   malaise  anorexia
1  0.3638369 0.12280702 0.4035088 0.05263158 0.8947368 0.6315789 0.3157895
2 -0.2116194 0.09183673 0.8714286 0.21428571 0.5102041 0.2551020 0.1428571
  liver_big liver_firm spleen_palpable   spiders    ascites   varices
1 0.7894737  0.5964912       0.36842105 0.5964912 0.31578947 0.28070175
2 0.8673469  0.2653061       0.09183673 0.1734694 0.02040816 0.02040816
   bilirubin alk_phosphate      sgot   albumin histology
1  0.6656623    0.7319112 0.4218266 -0.8670305 0.7543860
2 -0.3871593   -0.4257035 -0.2453481  0.5042933 0.2758102
```



Regresión

En este punto evaluaremos la capacidad de diferentes modelos de precedir valores a partir de datos previos, que serán tomados dividiendo el conjunto de entrenamiento mediante validación cruzada con k = 5,los modelos evaluados serán los siguientes: - Regresión lineal - Random Forest - KNN regresivo Para cada uno de ellos calcularemos las siguientes métricas: - El error cuadrático medio. - La raíz del error cuadrático medio. - El error absoluto medio. - R2.

Modelo	MSE	RMSE	MAE	R2
Regresión lineal	.	.	.	0.50236747 0213745
Random Forest	0.15821546 9747058	0.39776308 2433574	0.28173002 1506215	0.84075715 7072766
KNN regresivo	0.90667741 9354839	0.95377010 8230029	0.67096774 1935484	0.82300606 0739955

Modelo	MSE	RMSE	MAE	R2
Regresión lineal	.	.	.	0.4842984 24027902
Random Forest	0.1750743 64314286	0.4184188 05182577	0.2971458 91527188	0.8237887 8914421
KNN regresivo	0.9161290 32258065	0.9571462 05162747	0.8903225 80645161	. 6.0006361 32252

Mediante estas métricas observamos que el modelo de regresión que presenta mejores resultados es el Random Forest tanto para la validación cruzada como para el dataset reducido.

Clasificación

Los modelos de clasificación son algoritmos que se utilizan para predecir a qué categoría o clase pertenecerá una instancia basándose en sus características o atributos. Estos modelos se entrenan con datos históricos con ejemplos de instancias y sus respectivas etiquetas de clase. En este caso, aplicamos las redes neuronales, regresión logística, máquinas de soporte vectorial, random forest y knn. En datasets desbalanceados como en el que nos encontramos, el accuracy puede ser engañoso. Por eso hemos añadido la precisión que mide la exactitud de las predicciones positivas y la especifidad que indica qué proporción de las instancias que realmente son negativas fueron correctamente identificadas como negativas por el modelo.

	Accuracy	Precision	Error	Especificiad
Red neuronal	0.9483871	0.96875	0.05161290	0.9430894
Regresión Logística	0.8838710	0.65625	0.11612903	0.9430894
Máquinas de Vector de Soporte (SVM)	0.8129032	0.09375	0.18709677	1.0000000
Random Forest	0.9870968	0.93750	0.01290323	1.0000000
	AUC	Duracion		
Red neuronal	0.9559197	0.02191687		
Regresión Logística	0.7996697	0.01876521		
Máquinas de Vector de Soporte (SVM)	0.5468750	0.02514005		
Random Forest	0.9687500	0.03812599		

Con el dataset reducido tras la selección de atributos, los resultados son muy parecidos, sin ninguna diferencia significativa con las pruebas llevadas a cabo que con el dataset completo, el modelo Random Forest sigue siendo el modelo que mejor resultados nos proporciona y que por tanto deberíamos elegir para llevar a cabo clasificación en nuestro problema.
Por lo tanto, podemos decir que no merece la pena hacer la selección de atributos y utilizar un dataset reducido para este problema en concreto, aunque es una técnica que sí que puede ser de gran utilidad en otros problemas.