Rev Esp Cardiol. 2024;77 (Supl 1): 33

Revista Española de Cardiología

6101-2. RENDIMIENTO DE LAS EXPRESIONES REGULARES EN EL ANÁLISIS DE INFORMES DE ALTA PRESENTES EN LA HISTORIA CLÍNICA ELECTRÓNICA: EXPRIMIENDO LOS DATOS SECUNDARIOS

Alain García Olea[1], Ane García Domingo-Aldama[2], Marcos Merino Prado[2], Koldo Gojenola Galletebeitia[2], Aitziber Atutxa Salazar[2], Mikel Maeztu Rada[1], Iván García Díaz[1], Adrián Costa[1], Iván Cano[1], Fernando Díaz[1], Irene Hernández[1], Uxue Millet[1], Ainhoa Etxenike[1] y José Miguel Ormaetxe Merodio[1]

[1]Hospital Universitario de Basurto, Bilbao (Vizcaya), España y [2]Facultad de Ingeniería, Departamento de Lenguajes y Sistemas Informáticos. Universidad del País Vasco, Bilbao (Vizcaya), España.

Resumen

Introducción y objetivos: Este estudio tiene como objetivo investigar cómo la aplicación de expresiones regulares (regex) identifica en los informes de alta de la historia clínica electrónica (HCE) variables no codificadas o mal codificadas en el sistema de informado. Se pretende demostrar cómo esta optimización del proceso de imputación puede mejorar la precisión de los algoritmos predictivos de recurrencia en una población con debut de fibrilación auricular (FA).

Métodos: Se realizó un análisis retrospectivo de los informes de alta en formato de texto libre (.txt) de pacientes con debut de FA entre 2015 y 2018. Se entrenó un modelo sobre el que se implementaron expresiones regulares para identificar y extraer información relevante de los informes, especialmente aquella relacionada con las variables codificadas, el debut de FA y su recurrencia. Se compararon los datos obtenidos mediante este método con los datos codificados y se utilizó la comprobación manual de la HCE como *gold standard* comparativo.

Resultados: Sobre un *dataset* de 2453 instancias, la aplicación de expresiones regulares sobre los informes de alta resultó en una reducción significativa (58,1%) del número de *missing values* en las variables analíticas codificadas (figura A). Además, gracias a la identificación de valores no codificados como los datos ecocardiográficos (figura B) se observó una mejora sustancial en la integridad y la completud de los datos. La herramienta identificó el debut de FA en el 88,23% de las instancias e identificó en el 61% de los informes datos relativos a la recurrencia o ausencia de recurrencia de FA.

Porcentaje de valores faltantes en el análisis con regex frente a la codificación hospitalaria (A, %). Identificación de datos ecocardiográficos en informes de alta (B, %).

Conclusiones: Los resultados de este estudio respaldan la eficacia de utilizar expresiones regulares sobre los informes de alta de la historia clínica electrónica para disminuir los *missing values* en las variables analizadas para estudios a partir de datos secundarios. Además, el análisis de texto libre permite la optimización del

proceso de imputación de datos al identificar variables no codificadas de forma sistemática.