# Desarrollo y análisis de sesgos de un modelo de desidentificación de historias clínicas electrónicas en español

Sabrina L. López*[1], Mariela Rajngewerc*[2,3,4], Luciano Silvi[5], Laura Acion[3,1], Laura Alonso Alemany[4]

[1] Instituto de Cálculo, UBA, CONICET, Buenos Aires, Argentina
sabrina.lopez.ds@gmail.com
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
marielaraj@gmail.com
[3] MetaDocencia, Argentina
laura.acion@metadocencia.org
[4] Sección de Computación, FAMAF, UNC, Córdoba, Argentina
lauraalonsoalemany@gmail.com.com
[5] supervecina.com, España
luciano.silvi@gmail.com

**Resumen** Los registros de salud provenientes de historias clínicas electrónicas (HCE) son una fuente valiosa de información para múltiples usos secundarios de investigación, planeamiento, etc. Sin embargo, son datos sensibles y se encuentran legalmente protegidos por su potencial impacto en derechos fundamentales, como el derecho a la intimidad o a la no discriminación (por ej. en el acceso al mercado de trabajo).
Por ello, para adecuar los registros de salud a un uso secundario, resulta imprescindible aplicar procesos que eliminen aquella información que permita identificar al titular de los datos (desidentificación).
En este trabajo presentamos una prueba de factibilidad de la desidentificación automática de texto libre en HCE.
Se analizó una aproximación a la tarea, con especial atención a los potenciales sesgos en su funcionamiento, que pueden resultar en comportamientos discriminatorios. Teniendo en cuenta las particularidades de los datos: homogeneidad, heterogeneidad y proporción de entidades identificatorias, se aplicaron las métricas: *Treatment Equality*, *Equal Opportunity*, *Equalized Odds* y *Conditional Use Accuracy Equality*, para analizar la existencia de sesgos producidos por el modelo.

**Keywords:** Historia clínica electrónica · anonimización · sesgos

## 1. Introducción

Los registros de atención de salud contienen información muy útil para múltiples propósitos de investigación y políticas públicas. El proyecto “Gestión epide-

* Estas autoras contribuyeron por igual a este trabajo

miológica basada en inteligencia artificial y ciencia de datos" (ARPHAI)[6] avanza en la sistematización y uso de esta información.

En la línea de Uso Responsable de Datos de ARPHAI se trabaja especialmente en cuestiones críticas, como por ejemplo el tratamiento de datos de salud protegidos legalmente [3,4]. Los campos de texto libre de las historias clínicas electrónicas (HCE), donde se anotan cuestiones relativas a la consulta (síntomas, diagnósticos, etc.), presentan información personal protegida (IPP): expresiones identificatorias como documento de identidad, nombres y apellidos, y también expresiones que, sin señalar a una persona específica facilitarían su reidentificación, como datos sobre su familia, pertenencia a instituciones, enfermedades poco frecuentes, etc.

Un relevamiento inicial sobre una muestra de registros provenientes de la HCE de La Rioja encontró que entre un 6 % y un 12 % tenían información que permitiría la identificación de la persona atendida. Esta alta prevalencia de IPP hace que sea ineludible aplicar mecanismos de desidentificación. Sin embargo, aplicar estos mecanismos presenta varias dificultades: la casuística es muy variable; las personas expertas no alcanzan un acuerdo total sobre qué información de los textos permite efectivamente reidentificar a un usuario y, por su sensibilidad, existen muy pocos datos disponibles.

Aún con estas limitaciones, resulta valioso incorporar procesos de desidentificación al tratamiento de registros de salud electrónicos para tratar de reducir los riesgos que supone la presencia de IPP en esos registros, siempre con conciencia de que ninguno de estos procesos de desidentificación ofrece una garantía de protección total. Para ello, evaluar con profundidad el funcionamiento de estos procesos resulta crítico. Las métricas clásicas que se utilizan para la evaluación de modelos supervisados suelen ser agregadas y no aportan información acerca del comportamiento de los modelos en distintos subgrupos de la población.

En este artículo presentamos trabajo en curso sobre la evaluación de una aproximación automática para la desidentificación de texto libre de HCE de La Rioja, con especial atención a los patrones de error que pueden constituirse en sesgos perjudiciales para parte de la población.

## 2. Materiales

En el contexto del proyecto ARPHAI, se obtuvieron 2.394.499 registros de 214.308 pacientes dentro del periodo 04/10/2016 - 28/01/2021 del sistema Acuario de la provincia de La Rioja, utilizado para atención ambulatoria.

A partir de estos registros, se construyó una muestra de 2500 textos para el análisis de factibilidad de desidentificación automática. Se seleccionaron 1000 registros aleatoriamente y 1500 mediante muestreo dirigido para conseguir una buena representatividad de IPP y de esta manera poder evaluar mejor el modelo propuesto, así como inferir modelos con métodos de aprendizaje automático.

Para el proceso de anotación se contó con dos anotadores del ámbito de la salud que mediante la herramienta Label Studio [6], reconocieron 22 categorías

[6] http://www.ciecti.org.ar/arphai/

de entidades de IPP a partir de una guía de anotación adaptada de Marimon et al. (2019) [5]. A cada anotador se le asignaron 6 lotes de registros de igual tamaño con un 10 % de superposición para el cálculo de métricas de acuerdo.

## 3. Métodos

### 3.1. Algoritmo de desidentificación

Se desarrolló un modelo basado en reglas a partir de la exploración de los textos de la HCE. Las expresiones regulares son funciones que se aplican sobre el texto y, si se dan las condiciones de aplicación, reconocen una subsecuencia del texto como alguna de las categorías de IPP. Las reglas se basan en correspondencia de patrones a partir de palabras clave y expresiones regulares. Por ejemplo, si después de la palabra clave *Dra* figura una palabra incluída en la lista de apellidos esa secuencia es enmascarada como una cierta categoría de IPP: "*Dra. Pérez*" ⇒ "*Dra. <Personal de Salud>*".

### 3.2. Métricas de evaluación

Para evaluar el rendimiento del algoritmo de desidentificación y su comparación con otras aproximaciones se aplicaron métricas clásicas: *accuracy* balanceado, *f1 score* y *recall* [2].

Para complementar esta información agregada, las métricas de *fairness* están diseñadas para visibilizar el comportamiento de los modelos con atención a subgrupos específicos de población [7]. En este trabajo queremos visibilizar especialmente los errores de tipo falsos negativos (FN), en los que no se enmascara una entidad IPP, exponiendo al usuario. Para ello necesitamos recurrir también a la cantidad de entidades (TP) y no-entidades (TN) que fueron correctamente reconocidas por el algoritmo y la cantidad de no-entidades incorrectamente reconocidas como entidades (FP). Las siguientes métricas nos permiten comparar el desempeño en la desidentificación a través de subgrupos de población, representados en las fórmulas como $i$ y $j$:

*Treatment Equality* $\max_{i,j}\left\{\left|\frac{FN_i}{FP_i} - \frac{FN_j}{FP_j}\right|\right\}$.
Esta métrica visibiliza si el tipo de error que consideramos más grave, los falsos negativos (FN), es más prevalente en alguno de nuestros grupos de interés.

*Equal Opportunity* $\max_{i,j}\left\{\left|\frac{TP_i}{TP_i+FN_i} - \frac{TP_j}{TP_j+FN_j}\right|\right\}$.
Esta métrica muestra la proporción de entidades correctamente reconocidas (TP) respecto a la cantidad de entidades totales (TP+FN). De esta forma podemos visibilizar si en algún grupo se da una mayor exposición de datos personales porque no se reconocen correctamente las entidades.

*Equalized Odds* $\max_{i,j}\left\{\left|\frac{TP_i}{TP_i+FN_i}-\frac{TP_j}{TP_j+FN_j}\right|,\left|\frac{FP_i}{FP_i+TN_i}-\frac{FP_j}{FP_j+TN_j}\right|\right\}$.
Esta métrica muestra otra perspectiva sobre *Equal Opportunity*, incorporando la proporción de no-entidades que fueron correctamente reconocidas (TN) respecto al total de no-entidades totales (TN+FP).

*Conditional Use Accuracy Equality* $\max_{i,j}\left\{\left|\frac{TP_i}{TP_i+FP_i}-\frac{TP_j}{TP_j+FP_j}\right|,\left|\frac{TN_i}{TN_i+FN_i}-\frac{TN_j}{TN_j+FN_j}\right|\right\}$.
Esta métrica muestra la proporción de entidades y no-entidades reconocidas automáticamente que efectivamente pertenecen al tipo de entidad asignado.

## 4. Análisis de Resultados

Para la evaluación del algoritmo de desidentificación se utilizó un subconjunto de 1409 registros de los 2500 anotados, ya que se descartaron aquellos que no contaban con IPP, obteniéndose un total de 9758 entidades y 78398 no-entidades.

Para comparar modelos en aprendizaje automático suelen reportarse las métricas *f1 score*, que considera tanto el *precision* y el *recall*, y *accuracy*. El valor de *f1 score* macro obtenido por el modelo es de 0,78 y el de *accuracy* balanceado de 0,76. Para este problema, donde nos interesa poder reconocer aquellas entidades identificatorias, nos parece relevente reportar también el valor de *recall* ya que representa la proporción de entidades que fueron correctamente identificadas. Para el modelo propuesto el valor de *recall* es de 0,56, indicando que aproximadamente el 50 % de las entidades fueron correctamente identificadas.

Seguidamente, realizamos un análisis desagregado de rendimiento con respecto a género y grupo etario.

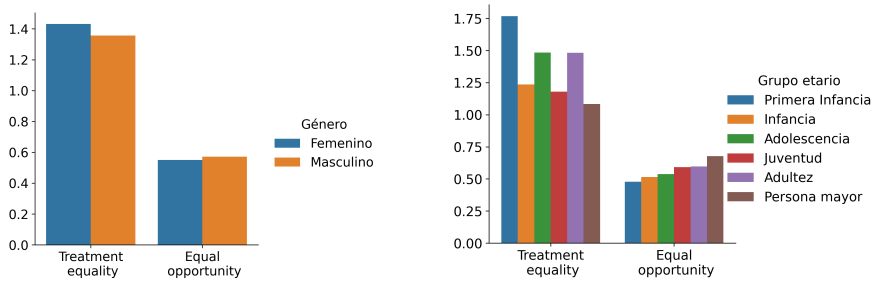


Figura 1: Valores de Treatment Equality y Equal Opportunity para desidentificación a través de subgrupos de género (izq.) y grupo etario (der.).

**Análisis de sesgos por género** Los registros utilizados cuentan con dos categorías de género: femenino (859 registros) y masculino (550 registros).

Los valores de las 4 métricas para evaluar *fairness* arrojaron diferencias menores a $0,023$ entre los dos grupos, con lo cual no apreciamos sesgos del algoritmo con respecto al género. La Figura 1 (izq.) muestra los valores obtenidos por las métricas *Treatment Equality* y *Equal Opportunity* para cada género.

**Análisis de sesgos por edad** Para el análisis de sesgos respecto a grupos etarios se consideraron las siguientes categorías: primera infancia (menor a 6 años), infancia (entre 6 y 11 años), adolescencia (entre 12 y 18 años), juventud (entre 19 y 26 años), adultez (entre 27 y 59 años) y personas mayores (mayores a 60 años). La proporción de entidades en cada uno de los grupos está entre 0,09 y 0,13, siendo la mayor la de primera infancia.

Las comparaciones entre los valores obtenidos para las diferentes métricas fueron: *Treatment Equality* dió 0,685, *Equal Opportunity* y *Equalized Odds* dieron 0,2 y *Conditional Use Accuracy Equity* dió 0,128. La Figura 1 (der.) muestra los valores obtenidos para las métricas *Treatment Equality* y *Equal Opportunity* para cada grupo etario. Se puede observar que primera infancia tiene un valor marcadamente más alto para *Treatment Equality*, lo cual indica que los pacientes de primera infancia sufrirían de una mayor exposición de IPP que el resto. Este tipo de comportamiento es un sesgo perjudicial para este grupo etario.

## 5. Conclusiones y trabajo futuro

En este trabajo hemos presentado un análisis de factibilidad de una aproximación a la detección automática de IPP en texto libre de HCE, con especial atención a los patrones de error que pueden constituirse en sesgos perjudiciales para parte de la población. No hemos observado inequidades en el funcionamiento del sistema para los distintos géneros, pero sí los habría con respecto a segmentos etarios. Es decir, los pacientes de primera infancia sufren de una mayor exposición de IPP que otros segmentos etarios. Este análisis nos va a permitir tomar medidas específicas para revertir este comportamiento perjudicial para el grupo de primera infancia desarrollando más reglas que traten la casuística específica de este segmento etario. También podremos incluir en el modelo técnicas de mitigación como por ejemplo en vez de suprimir los nombres hallados en el texto libre reemplazarlos por otros (esta técnica es conocida como *Hidden in Plain Sight* [1]). Esta técnica de mitigación tiene como beneficio que si algún dato de IPP no fue identificado por el modelo, el lector no podrá distinguir cuales de los IPP son reales y cuales son ficticios, de esta manera se incorporá una nueva dificultad para el proceso de reidentificación de los datos.

En lo que respecta a las anotaciones utilizadas en este estudio para poder identificar la IPP, presentar un análisis de acuerdo entre los anotadorares permitiría describir la calidad de las etiquetas consideradas y la complejidad del problema aún cuando se realiza de manera no automática. En futuros trabajos se incluirá un análisis detallado que describa los acuerdos y desacuerdos entre anotadoradores.

Además, sería beneficioso realizar análisis que consideren una mayor variedad de segmentos de la población para poder describir con mayor detalle las debi-

lidades y potencialidades del modelo presentado. Sin embargo, dichos análisis requieren de una mayor cantidad de datos anotados por expertos para lo cual serían necesarios nuevos fondos de financiamiento.

En la actualidad, existen modelos basados en inteligencia artificial que se utilizan para extraer entidades específicas de los textos libres (por ejemplo: nombres, localidades, etc.). Algunos de ellos pueden ser entrenados para extraer categorías que se consideran IPP. Los futuros pasos de este trabajo incluyen comparar el modelo propuesto con los modelos de desidentificación del estado del arte basados en grandes modelos de lenguaje.

## Referencias

1. Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., Hirschman, L.: Hiding in plain sight: Use of realistic surrogates to reduce exposure of protected health information in clinical text. Journal of the American Medical Informatics Association **20**(2), 342–348 (2013)
2. Flach, P., Kull, M.: Precision-recall-gain curves: Pr analysis done right. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)
3. Ley 25.326: de Proteccion de los datos personales. Buenos Aires, Argentina (2000)
4. Ley 26.529: de Derechos del Paciente en su Relación con los Profesionales e Instituciones de la Salud. Buenos Aires, Argentina (2009)
5. Marimon, M., Gonzalez-Agirre, A., Intxaurrondo, A., Rodriguez, H., Martin, J.L., Villegas, M., Krallinger, M.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: IberLEF@ SEPLN. pp. 618–638 (2019)
6. Tkachenko, M., Malyuk, M., Holmanyuk, A., Liubimov, N.: Label Studio: Data labeling software (2020-2022), https://github.com/heartexlabs/label-studio
7. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. p. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018)