Clustering explicable de historiales médicos para el ámbito de una obra social

Adriano Lusso
Universidad Nacional del Comahue
adriano.lusso@est.fi.uncoma.edu.ar

Antonella Torres
Universidad Nacional del Comahue
bianca.torres@est.fi.uncoma.edu.ar

I. Introducción

El Aprendizaje Automático, o *Machine Learning*, es una herramienta fundamental para el análisis de la información. En combinación a la gran cantidad de datos disponibles en la actualidad, permite realizar categorizaciones automáticas, detectar elementos fraudulentos en tiempo real, encontrar y caracterizar patrones y otra variedad de tareas.

Entre sus subramas, se encuentra el análisis de segmentación o *clustering*, que permite segmentar en diferentes grupos a un gran conjunto de datos de interés en función de que tan parecidos son unos datos con otros. Es utilizado tanto entornos académicos como empresariales e industriales. Algunas de sus principales aplicaciones se encuentran en la mercadotecnia [23], el análisis financiero [7], la salud [20] y las ciencias sociales [12].

Aún así, en la actualidad existe debate sobre el uso ético del *Machine Learning* y otras subramas de la Inteligencia Artificial [3, 14]. En entornos críticos y socialmente sensible, como la salud o el derecho, los fallos y sesgos que pueden introducir estas técnicas del estado del arte repercuten directamente sobre la calidad de vida de las personas.

En lo referente a la salud, se ha realizado investigación de relevancia [5, 10, 19]. La gran cantidad de datos que existen de pacientes médicos brindan de un potencial enorme al *clustering* a la hora de caracterizar y segmentar a los mismos. Los resultados de estas investigaciones pueden ser adaptados para el uso en diferentes hospitales, instituciones y obras sociales.

Respecto al entorno local de la ciudad de Neuquén, Argentina, existe el Servicio de Obra Social de la Universidad Nacional del Comahue (SOSUNC). SOSUNC cuenta con un área de Auditoría Médica, encargada de la gestión y evaluación de solicitudes de cobertura médica. En base a la información disponible de los afiliados, se decide si una determinada solicitud cumple los requerimientos necesarios para que su costo sea cubierto por SOSUNC. Además, el área trabaja en la creación de diferentes planes de cobertura, que permiten agrupar y generalizar conductas médicas repetidas.

El objetivo de este trabajo es buscar oportunidades de aplicación de técnicas de *clustering* explicable para el área de Auditoría Médica de SOSUNC. Las mismas deben permitir subdivir planes de cobertura en otros planes más específicos y de menor tamaño, con una menor cantidad de medicamentos y prácticas médicas involucrados. En base estos nuevos subplanes, SOSUNC podrá negociar mejores porcentajes de cobertura con los prestados de servicios médicos. Además, los resultados deben facilitar el análisis de los afiliados a Auditoría Médica, permitiéndoles tomar mejores decisiones respecto a las coberturas aprobadas. Se diseñará un flujo de trabajo para *Machine Learning*, que involucre la elección y estudio de los conjuntos de datos necesarios, su preprocesamiento, la elección y aplicación del algoritmo de *clustering* y el análisis de su rendimiento. Las conclusiones hechas serán formalmente estructuradas y entregadas a SOSUNC para que puedan hacer uso de las mismas.

Además, se diseñará en implementará el prototipo de una aplicación de software que permita aplicar *clustering* sin necesidad de un conocimiento herramientas avanzadas de minería y ciencia de

datos (lenguajes de programación, librerías especializadas, etcétera). Esto ayudará a que la ejecución de flujos de *clustering* pueda repetirse en diferentes oportunidades por el propio equipo de Auditoría Médica.

En la Sección II, se define el marco teórico necesario para el desarrollo de este trabajo, lo que incluye *clustering*, el algoritmo *Optimal Matching* y el marco matemático de explicabilidad *SHAP*. En la Sección III, se introduce el dominio de SOSUNC, mientras que en la Sección IV se diseña el flujo de *Machine Learning* que se llevará a cabo para las pruebas. En la Sección V se mencionan las herramientas y métodos utilizados para la implementación de las pruebas. En la Sección VI se detallan los conjuntos de datos de estudio, entre los que se incluyen los consumos médicos, prácticas y monodrogas de interés. En la Sección VII, se detallan las consideraciones necesarias para la selección del número de *clusters* a utilizar, entre las que se incluyen el protocolo de selección diseñado y los resultados de su utilización. Luego, en la Sección VIII se presentan los resultados obtenidos en las pruebas de *clustering*, dando pie a las discusiones en la Sección IX. La Sección X explica las limitaciones de las pruebas experimentales, mientras que la Sección XI introduce el prototipo de software diseñado. Las Secciones XII y XIII hablan sobre los trabajos a futuro y las conclusiones. Finalmente, en la Sección XIV se brinda acceso al repositorio con las implementaciones realizadas.

II. Marco teórico

II-A. Clustering

El análisis de segmentación, o *clustering* [9] comprende una gran clase de métodos cuyo fin es encontrar un número limitado de grupos, segmentos o *clusters* significativos en un conjunto de datos. Pertenece a la categoría de aprendizaje no supervisado, lo que significa que no se tiene conocimiento previo de la división de los grupos, como si sucede en el aprendizaje supervisado.

Si bien existen una amplia variedad de algoritmos de *clustering*, con sus propias ventajas y desventajas, para este trabajo de laboratorio se hará uso del algoritmo conocido como K-medoides. Este pertenece a la categoría de *clustering* no jerárquico, la cual se caracteriza principalmente por:

- Requerir que se especifique el número exacto de *clusters* a generar.
- No requerir el cálculo completo de una matriz de disimilitud $n \times n$.

K-medoides, como su nombre indica, consiste en encontrar K *clusters*, donde cada uno de estos se identifica por un medoide. Se define al medoide como la unidad de dato de su respectivo *cluster* que minimiza la suma de las distancias cuadradas entre esta unidad y todo el resto de unidades del *cluster*. Esta sumatoria puede volver ineficiente a K-medoides, sobretodo al de compararlo con K-medias, otro conocido algoritmo de *clustering* no jerárquico [29].

Existen diferentes mecanismos, heurísticas y métricas para elegir el número óptimo de *clusters*. Una posible métrica es la silueta o *silhouette*. Esta evalúa, para cada dato, si este es más cercano a los elementos de su propio *cluster* o a los elementos de los *clusters* vecinos [4]. Toma valores en el intervalo $[-1, 1]$, donde a más cercano sea a 1, mayor es la similitud del dato con los demás datos de su *cluster*. Se puede calcular la *silhouette* de un *cluster* como el promedio de las *silhouettes* de todos los datos de ese *cluster*. A su vez, se puede calcular la *silhouette* de un *clustering* como el promedio de las *silhouettes* de todos sus *clusters*.

El algoritmo clásico de K-medoides también es conocido como el Particionamiento Alrededor de Medoides (PAM), y cuenta con su implementación en el paquete *cluster* del lenguaje de programación R [18].

II-B. Optimal Matching

Optimal Matching (OM) es un conjunto de métodos que permiten realizar análisis de secuencias. Originalmente, fueron desarrollados para el análisis eficiente de secuencias de ADN y de proteínas [28].

Se puede definir una secuencia de longitud L como una sucesión de estados $\{E_l\}_{l \leq L}$. Teniendo un conjunto de n secuencias, OM genera una matriz de disimilitud $n \times n$, donde el elemento (S_1, S_2) indica la disimilitud que hay entre las secuencias $0 \leq S_1 < n$ y $0 \leq S_2 < n$.

La disimilitud entre dos secuencias se calcula en base a un costo de sustitución y un costo de inserción/eliminación entre estados de la misma posición l. Se define al costo de sustitución como la cantidad de elementos de una misma posición que no coinciden entre las dos secuencias de análisis. El costo de inserción/eliminación refiere a la necesidad de eliminar o insertar estados en alguna de las secuencias a causa de tener longitudes diferentes [8]. Si todas las secuencias del conjunto de análisis tienen la misma longitud L, entonces OM jamás hará uso del costo de inserción/eliminación.

Para definir los costos de sustitución se puede optar por usar el método TRATE [8]. Este indica la probabilidad de transicionar de un estado S_i a un estado S_j. El costo disminuye a medida que la transición entre estados se vuelve más común, asumiendo que si hay muchas transiciones entre i y j, entonces son estados similares. Formalmente, TRATE se define como una matriz de sustitución tal que

$$\forall i,j \begin{cases} C(i,j) = C(j,i) = 2 - p(i|j) - p(j|i) \\ \quad ,si\ i \neq j \\ \\ C(i,j) = C(j,i) = 0 \quad ,si\ i = j \end{cases} \tag{1}$$

donde $C(i,j)$ es el costo de transición entre los estados y $p(i|j)$ es la razón de transición del estado j al i.

Finalmente, la matriz de disimilitud resultante de OM pueden servir como entrada a los métodos de *clustering*, tanto jerárquicos como no jerárquicos.

II-C. SHAP

Shapley Additive Explanations, o SHAP [16], es un marco de trabajo para la interpretación de modelos predictivos. En el mismo, se define una clase de *Métodos Aditivos de Atribución de Características*, entre los que se encuentran LIME, DeepLIFT y TreeExplainer [17].

La base conceptual de SHAP consiste en interpretar cualquier explicación de un modelo predictivo f como un modelo en sí mismo. De esta manera, cuando la complejidad del modelo predictivo f dificulta su entendimiento, se puede hacer uso de un modelo explicativo g que funcione como una aproximación interpretable y simplificada del modelo original f.

Haciendo uso de los métodos descritos, se pueden aproximar eficientemente los valores de SHAP. Estos son el elemento principal del marco de trabajo, ya que funcionan como una métrica de la importancia de los *features* en f. Dada una entrada x, los valores de SHAP le atribuyen a cada *feature* el cambio en la predicción esperada del modelo al verse condicionada por la *feature*, tal como muestra la figura Figura 1. Permite explicar como es que f llega desde $E(f(z))$, el valor base que seria predicho si no hubiera conocimiento de los *features*, hasta la predicción final $f(x)$.

El valor $E(f(z))$ es calculado como el promedio de todas las predicciones de f sobre el conjunto de entrenamiento. De esta forma, los valores de SHAP muestran como que es que cada *feature* genera variabilidad en la predicción para una determinada entrada respecto al promedio de las predicciones de todas las entradas de entrenamiento.

III. DESCRIPCIÓN DEL DOMINIO

SOSUNC cuenta con una base de datos en la que se almacenan sus afiliados, consumos médicos, planes de cobertura, medicamentos (formados por monodrogas), prácticas médicas, entre otros. Los consumos médicos de los afiliados pueden ser, principalmente, de medicamentos o de prácticas médicas. Al momento de almacenar esta información, también se especifica su fecha de consumo.

Los planes de cobertura médica en SOSUNC se realizan con el objetivo de segmentar diferentes tipos de consumos en un mismo tipo de plan, y brindarles a los afiliados de porcentajes personalizados de cobertura médica.

Tomando inspiración en el trabajo hecho en [4], se realizará *clustering* de secuencias. Bajo este dominio, una secuencia sera un historial médico $\{E_l\}_{l \leq L}$ de un afiliado, donde l indica uno de los L marcos temporales del historial. Un marco temporal se caracteriza por la unidad de tiempo que representa, pudiendo ser mensual, semestral o anual. Por otra parte, el código de estado E representa que consumos médicos ha realizado el afiliado en el respectivo marco temporal. Siguiendo esta

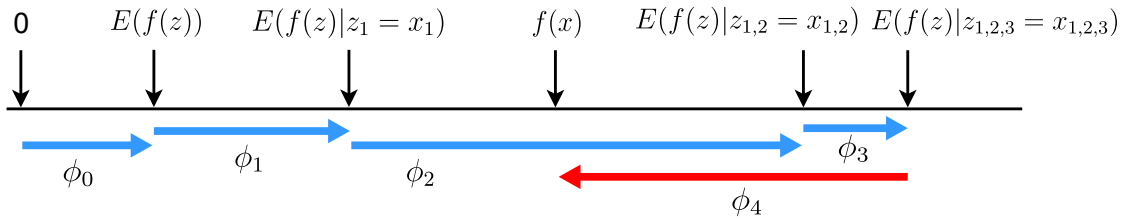


Figura 1: Valores de SHAP $\{\phi_i\}_{0<i<5}$ para un modelo predictivo f y una entrada x. Imagen inspirada en [16]. Los valores ϕ mencionados representan el impacto que tiene cada una de las 4 *features* de x en f. Se comienza en el valor inicial $E(f(z)) = \phi_0$, calculado como el promedio de las predicciones para todo el conjunto de entrenamiento. Este valor representa la predicción de f para una entrada de la que no se conocen sus *features*. A partir del valor inicial, se muestra como agregar una a una las *features* afecta a la predicción hecha, llegando finalmente a la predicción final $f(x)$ cuando las 4 *features* son observadas por f.

lógica, si se tiene el historial médico A_1, B_2, A_3, el marco temporal es semestral y $A = \{$*ibuprofeno*$\}$, $B = \{$*radiografía*$\}$, se puede interpretar que en el primer y tercer semestre de análisis el afiliado consumió ibuprofeno y en el segundo semestre se realizó una radiografía.

Además, para restringir la búsqueda de *clusters*, SOSUNC proporcionará un conjunto de prácticas médicas y monodrogas de interés relacionadas a una determinada patología. De esta manera, para generar los posibles estados S, solo se tendrán en cuenta consumos con estos elementos.

IV. Diseño del flujo de Aprendizaje Automático

En función al dominio descrito anteriormente, se diseño un flujo de trabajo de *Machine Learning*, el cual puede verse en la Figura 2. El mismo cuenta con las etapas típicas de preprocesamiento de lo datos, aplicación del algoritmo de *Machine Learning* y análisis de resultados. Además, se añadieron etapas posteriores a la aplicación del algoritmo y previas al análisis de datos, con el objetivo de aplicar SHAP. En estas, se entrena un modelo predictivo sobre el cual se puede generar un modelo explicativo para sus *features* [15]. Para este trabajo, se hará uso de un *Random Forest* [24]. De esta manera, el análisis de resultados no involucra unicamente a los *clusters* y sus métricas de rendimiento, sino también a los valores de SHAP calculados.

Si bien en la sección Sección II-A se explica que una de las características de K-medoides es que no necesita de una matriz de disimilitud completa, en el flujo de trabajo definido se observa el cálculo de la misma. Esto se debe a limitaciones de las librerías de programación utilizadas para la implementación de las pruebas, detalladas en la Sección V. Particularmente, la librería Kmedoids implementa diferentes métodos para K-medoides, y en todos ellos requiere como dato de entrada la respectiva matriz de disimilitud. Como enuncia el marco teórico, no hace falta la matriz completa, pero puede hacerse uso de ella para ejecutar el algoritmo. La desventaja de esto es que se están calculando más disimilitudes de las necesarias para el funcionamiento de K-medoides.

La inclusión de SHAP tiene dos objetivos principales en este trabajo. El primero de ellos es brindar a los auditores médicos de SOSUNC una mayor cercanía al funcionamiento del *clustering*. Esto permitirá generar confianza en el método, gracias a la explicabilidad que otorgan los valores de SHAP. Esta explicabilidad, a su vez, ayudará a los auditores a entender los factores que generaron el respectivo *clustering*, aumentando el entendimiento que tienen del dominio de SOSUNC. Finalmente, la ética profesional juega un papel importante en la

decisión de implementar SHAP. Al estar tratando con un area sensible como la salud, es importante tomar decisiones justificadas en base a un modelo de segmentación que funcione correctamente y sin sesgos de aprendizaje. Los valores de SHAP ayudan a comprender el razonamiento llevado a cabo por K-medoides, y por lo tanto, ayuda a detectar estos posibles sesgos que pueden surgir durante el procedimiento.

El segundo objetivo de SHAP es facilitar la interpretación de resultados y confirmar las conclusiones tomadas en las Secciones IX-A y IX-B. Los patrones encontrado en los *clusters* pueden verse sesgados por los métodos de visualización elegidos. Entre algunos factores se encuentran el tipo de gráfico, selección de colores y el tamaño del gráfico. Además, en las Secciones A y VI-C se explica que la cantidad posibles de estados E en un historial médico crece exponencialmente, lo que dificulta encontrar una selección de colores disponibles que los vuelva identificables entre sí para la vista humana. Se busca mitigar este inconveniente limitando la cantidad de estados analizados, lo cual a su vez introduce un gran faltante de información que puede sesgar las conclusiones tomadas. Estos y otros problemas esperan ser solventados con el uso de SHAP como un marco que brinde mayor explicabilidad al *clustering*.

V. Métodos y herramientas utilizadas

Para el preprocesamiento de los datos, la aplicación de K-medoides y generación de valores de SHAP se utilizó el lenguaje de programación Python junto a algunas librerías como Numpy, Pandas, sklearn, SHAP y Kmedoids [11, 13, 16, 21, 22, 25, 30]. Por otra parte, la librería Traminer [8] de R brindó la implementación para aplicar el método OM.

Para generar la matriz de disimilitud con Traminer, se utilizó la función `seqdist(sequences, method=OM, indel=1, sm=TRATE)`, indicando los historiales médicos, un costo de 1 para inserción/eliminación y haciendo uso del método TRATE para el costo de sustitución . En Kmedoids, se utilizó la clase `KMedoids`, indicando el número de *clusters* a utilizar y el método *fasterpam* que implementa una versión optimizada del método tradicional PAM [26, 27].

Para crear y entrenar un modelo predictivo basado en el *clustering*, se utilizó la clase `RandomForestClassifier` de sklearn. Esta fue posteriormente utilizada en la clase `TreeExplainer` de SHAP para crear un modelo explicativo del *RandomForest*. La función de SHAP `beeswarm` será usada para graficar los resultados obtenidos.

Para las pruebas experimentales del flujo de *Machine Learning* se utilizó Jupyter Notebook como entorno de programación. Por otra parte, el desarrollo del prototipo de software fue también trabajado en Python haciendo uso de la librería Streamlit, con la que se desarrollaron las interfaces de usuario.

VI. Conjuntos de datos de estudio

En esta sección, se explicarán las estructuras de los conjuntos de datos utilizados en este trabajo. Entre estos, se encuentras los brindados por SOSUNC como datos de entrada y los resultantes de aplicar el preprocesamiento correspondiente. Todos los conjuntos de datos fueron trabajados en formato CSV.

VI-A. Consumos médicos

Los consumos médicos son los consumos de monodrogas y de prácticas médicas. Cada tipo de consumo pertenece a un archivo CSV diferente. Ambos tienen una estructura similar. Cada consumo es una tupla compuesta del identificador del afiliado consumidor, el identificador de la monodroga o práctica consumida, el nombre legible de la misma y la fecha de consumo. Se tiene registro de consumos de prácticas desde el año 2019 al 2024 y de consumos de monodrogas desde el año 2008 al 2024.

Es importante destacar que, por privacidad de los afiliados, los identificadores utilizados son de generación aleatoria y no representan ningún dato sensible de los mismos.

VI-B. Monodrogas y prácticas de interés

Como se mencionó en la sección III, SOSUNC brinda las práctica y monodrogas de interes para

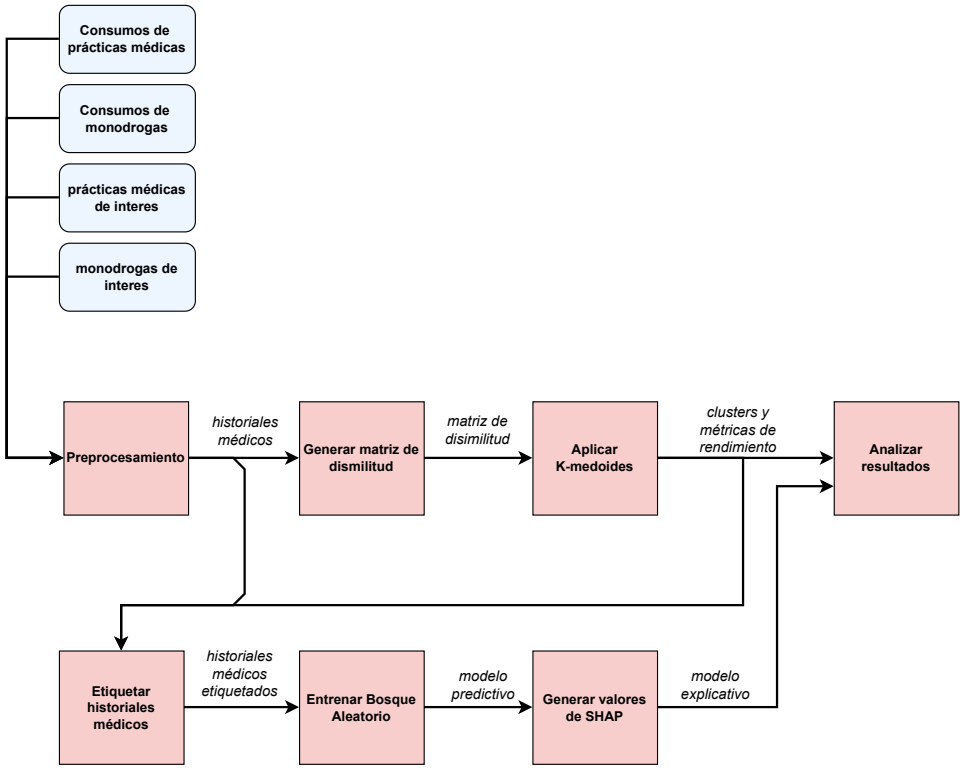


Figura 2: Diseño del flujo de *Machine Learning*. En el mismo, se realiza un preprocesamiento de los datos de entrada para obtener los historiales médicos. Luego, se genera la matriz de disimilitud utilizando el método de *Optimal Matching* (OM). Se aplica el algoritmo de K-medoides haciendo uso de la matriz de disimilitud. Se etiqueta el conjunto de historiales médicos con los *clusters* obtenidos para poder entrenar un árbol de decisión. El modelo predictivo resultante es usado para generar un modelo explicativo con el marco de trabajo SHAP. Finalmente, los valores de SHAP, los *clusters* y otras métricas de rendimiento son analizados para encontrar patrones en los datos.

el análisis. Cada tipo de monodroga o práctica de interés consta de un archivo CSV diferente. Particularmente, SOSUNC brindo tres archivos CSV, que corresponden respectivamente a las prácticas relacionadas a diabetes, prácticas relacionadas a problemas cardiovasculares y monodrogas relacionadas a diabetes. Cada conjunto de datos consta de tuplas formadas por el identificador de la práctica o monodroga y su nombre legible.

En este trabajo, se limitará a utilizar unicamente los conjuntos de datos de prácticas médicas y monodrogas de diabetes. Estos incluyen 17 prácticas médicas y 38 monodrogas respectivamente. Como se detallará en la sección VI-C, hay un crecimiento exponencial de almacenamiento de datos en función de la cantidad de elementos de interés. Debido a esto, en este trabajo se decidió establecer un limite máximo de hasta 20 elementos de interés (monodrogas y prácticas médicas) a analizar. Para filtrar del total de 55 elementos a un número final 19 se llevo a cabo el preprocesamiento correspondiente.

Primero, se filtraron aquellos elementos de in-

terés que no formaban parte de ningún consumo médico en el respectivo conjunto de datos. Luego, se decidió agrupar algunas monodrogas en subconjuntos excluyentes, en función del parecido de las mismas. Esto resultó en 5 grupos de monodrogas. Cada grupo se interpretará como un elemento de interés único. Finalmente, se ordenan de manera independiente las prácticas de interés y las monodrogas (y grupos) de interés en función de su frecuencia en consumos médicos. Para el conjunto de elementos de interés final se tomarán las 10 prácticas médicas con mayor frecuencia y las 10 monodrogas y/o grupos de monodrogas con mayor frecuencia. Debido a que tras el agrupamiento de monodrogas solo quedaron 9 de ellas desagrupadas, el número final fue de 10 prácticas médica y 9 monodrogas y/o grupos. En la Tabla I, se observan los 20 elementos de interés finales, indicando si son una práctica, una monodroga o un grupo de monodrogas.

VI-C. Posibles estados en un historial médico

Él conjunto de los posibles estados que pueden formar un historial médico es generado haciendo uso del conjunto Δ de elementos de interés detallados de la Tabla I. Dado el cardinal $|\Delta| = 19$, se tiene un total de $2^{|\Delta|} = 524288$ posibles estados para un marco temporal en un historial médico. Este valor surge de interpretar a cada uno de los elementos en Δ como un valor binario que indica si el elemento fue consumido por el afiliado. Para simplificar el *clustering*, cada estado tendrá asignado un código hexadecimal que corresponderá al respectivo número binario que indica que elementos de interés fueron consumidos por el afiliado.

VI-D. Historiales médicos y matriz de disimilitud

Haciendo uso de los conjuntos de datos de las secciones VI-A, VI-B y VI-E se generan los historiales médicos. Para este trabajo, se trabajará con dos conjuntos de historiales médicos. En el primero de ellos, los historiales médicos se conformaran de 7 marcos temporales semestrales o semestres, iniciando en el primer semestre del año 2021 y terminando en el primer semestre del año 2024. En el segundo, se conformarán de 3 años, iniciando a mediados de 2021 y terminando a mediados de 2024.

Además, como se explicó en la sección II-B, se utiliza el algoritmo OM para genera una matriz de disimilitud entre los diferentes historiales médicos. Esta matriz es la principal entrada al método de K-medoides.

VI-E. Adaptación de historiales médicos para SHAP

En la Sección III se define a un historial médico como una secuencia $\{E_l\}_{l \leq L}$ de estados E en L marcos temporales diferentes. A su vez, como se indico en la Sección VI-C, un estado es uno de los $2^{|\Delta|}$ valores hexadecimales que representan que elementos de interés han sido consumidos en el marco temporal. Al momento de analizar en detalle un historial médico, sera necesario descomponer cada estado E_l en su máximo de hasta $|\Delta|$ posibles consumos, con el objetivo de poder entender que representa el código hexadecimal. Esto, si bien es factible para analizar los *clusters* directamente, dificulta el análisis de los valores de SHAP.

La librería de SHAP espera que cada *feature* del modelo sea una variable continua o binaria para que los datos se adapten correctamente a la función del gráfico `beeswarm`, introducida en la Sección V. A causa de esto, será necesario adaptar las *features* categóricas de los historiales médicos a *features* binarias para entrenar el modelo predictivo y generar el modelo explicativo.

El conjunto de historiales médicos adaptado cuenta con $L \cdot |\Delta|$ *features*, donde cada una indica si en un marco temporal de análisis $l \leq L$ se ha consumido un elemento de interés $\delta \in \Delta$.

VII. Selección de número de clusters

Como se mencionó en la sección II-A, es necesario seleccionar el número de *clusters* sobre el que trabajar. Por lo tanto, en la Sección VII-A se presenta un protocolo que permite encontrar, en función a ciertos criterios, un número adecuado de *clusters* a utilizar. Luego, en la Sección VII-B se hace uso del protocolo, se analizan y discuten sus resultados con el objetivo de concluir la cantidad óptima de grupos seleccionados. En esta misma sección, se descarta

Elemento de interés	Categoría	Elementos del grupo
Control de (C.d.) glucemia	Práctica	
C.d. Colesterol hdl	Práctica	
C.d. hemoglobina glicosilada	Práctica	
C.d. microalbuminuria	Práctica	
C.d. colesterol total	Práctica	
Consulta a especialista en diabetes	Práctica	
Consulta inicial con plan nutricional	Práctica	
Consulta oftalmológica	Práctica	
Creatinina sérica o urinaria	Práctica	
C.d. colesterol ldl	Práctica	
Gliclazida	Monodroga	
Glipizida	Monodroga	
pioglitazona	Monodroga	
dapagliflozina	Monodroga	
Insulina	Grupo de 13	Insulina humana modifica, insulina bovina, insulina humana, insulina porcina, insulina glargina, insulina aspártica, insulina detemir, dispositivo para aplicar insulina, insulina glulisina, insulina lispro, insulina defludec, insulina aspártica bifásica, insulina lispro + insulina lispro protamina
empagliflozina	Grupo de 2	empagliflozina, empagliflozina + metformina clorhidrato
metformina	Grupo de 8	metformina, metformina + glibenclamida, rosiglitazona + metformina, glimepiride + metformina, sitagliptina + metformina clorhidrato, vildagliptin + metformina, linagliptina + metformina, empagliflozina + metformina clorhidrato
sitagliptina	Grupo de 2	sitagliptina, sitagliptina + metformina clorhidrato
vildagliptin	Grupo de 2	vildagliptin, vildagliptin + metformina

Cuadro I: Tabla con los 19 elementos de interés para formar los 2^{19} posibles estados de los marcos temporales de un historial médico. Se indica si es una práctica médica, una monodroga o un grupo de monodrogas. Si es un grupo, se indica la cantidad y nombre de las monodrogas que lo conforman.

el uso de los historiales médicos formados por 7 semestres, de manera que las pruebas principales serán llevadas a cabo unicamente con el conjunto de historiales formados por 3 años.

VII-A. Protocolo presentado

Se diseñó un protocolo que, dado un rango $i \in [2, n]$ de números de *clusters* a evaluar, realiza un muestreo de 100 *clusterings* para cada i. De esta manera, para cada número de *clusters* establecido i se realiza un promedio de las 100 *silhouettes* obtenidas.

Teniendo el historial de *silhouettes* promedio generado, el criterio para seleccionar el número óptimo puede variar. Dependiendo las circunstancias, puede ser conveniente elegir aquel i con mayor *silhouette* promedio. Aún así, utilizar esta heurística sin criterio puede llevar a una elección sesgada de i. A partir de cierto i lo suficientemente grande, a medida que aumenta el número de *clusters*, la *silhouette* promedio también lo hará uniformemente. Esto, lejos de necesariamente representar un mejor *clustering*, puede suceder debido a que la cantidad de secuencias en cada *cluster* disminuye, indicando que se esta segmentando en exceso y perdiendo capacidad de generalización de patrones en el conjunto de datos. Lo ideal es encontrar un i lo suficientemente pequeño con un *silhouette* promedio lo suficientemente alto para que la métrica no esté sesgada. Esto da pie a plantear un *trade-off* entre la *silhouette* promedio y la selección final de i.

VII-B. Uso del protocolo

El protocolo presentado sera utilizado para el conjunto de historiales de 7 semestres y para el de historiales de 3 años. Los historiales resultantes pueden ser vistos en la figura 3.

Para este trabajo, se asumen un conjunto de supuestos que no necesariamente pueden ser generali-

zados por todo tipo de pruebas. Se considerará que el *silhouette* de un *clustering* es mínimamente aceptable si se encuentra en el intervalo de $[0{,}2, 0{,}5]$. En casos superiores, se esta hablando de un *clustering* de muy buena calidad, y en casos inferiores, de uno de mala calidad.

Otro factor importante al hacer el siguiente análisis es tener en cuenta el sesgo introducido por un número de clusters excesivamente alto. No hay un valor predeterminado para esto mismo, pero en este trabajo se lo asume dentro del rango de $[10, 20]$ *clusters*. Como en la sección anterior, este sesgo se debe a que a mayor el número de *clusters*, menor la cantidad de secuencias en cada uno de estos.

Al analizar el historial de *silhouettes* superior, correspondiente a los historiales médicos de 7 semestres, se observa una calidad general de *silhouette* muy baja. Si bien para 2 *clusters* se alcanza un valor que podría considerarse de alta calidad, el mismo decae en las iteraciones posteriores y no logra recuperarse. Incluso a partir de la iteración 13, que es donde puede observarse que el sesgo descrito anteriormente empieza a aparecer, no se logra superar una *silhouette* de $0{,}2$.

Por otra parte, los resultados de los historiales médicos de 3 años son más prometedores. Se empieza con una *silhouette* aceptable con 2 *clusters*, la cual decae posteriormente, pero logra recuperarse a partir de la iteración 8. Además, se detecta que en la iteración 12 se empieza a introducir el sesgo. Considerando las iteraciones previas a la aparición del sesgo, existe un pico claro en la iteración 10.

En base al análisis previo, se concluye como conveniente avanzar unicamente con los historiales médicos de 3 años. No encontrar ningún número de *clusters* superior a 2 que llegue a un nivel promedio de *silhouette* aceptable da indicios de que los historiales médicos de 7 semestres no están preparados para ser procesados exitosamente en un *clustering*. Es importante recordar que la *silhouette* es una métrica que mide directamente que tanta cohesión hay entre elementos de un mismo *cluster* y que tanta distancia entre elementos de diferentes *clusters*.

Para continuar el trabajo con los historiales médicos de 3 años, se decidirán dos valores diferentes para el número de *clusters*. El primer valor será 2 y el segundo 10. Esta decisión es tomada teniendo el cuenta el *tradeoff* entre la *silhouette* y un número bajo de *clusters*.

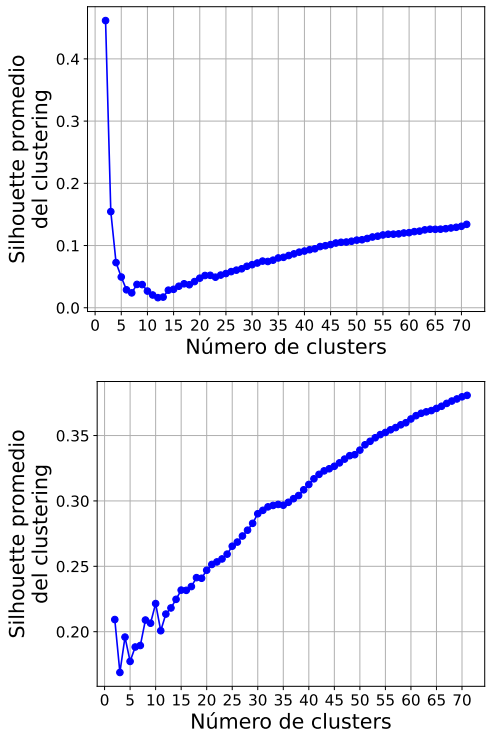


Figura 3: Historiales de *silhouettes* promedio obtenidos por el protocolo presentado en la Sección VII-A. En el eje X, se itera sobre el número de *clusters*. En el eje Y, se visualiza el *silhouette* promedio obtenida. El historial superior fue calculado para el conjunto de datos de historiales médicos de 7 semestres. El inferior, para el de 3 años.

VIII. Resultados experimentales

Los resultados de las pruebas realizadas se conforman de dos conjuntos de *clusters* con sus respectivos valores de SHAP. En la Sección A, se observan los *clusters* encontrados para el 2-*clustering* y el 10-*clustering*, describiendo la notación necesaria para la correcta lectura e interpretación de los gráficos. Además, en la Sección B se presentan los valores

de SHAP asociados a cada *cluster* haciendo uso de gráficos *beeswarm*.

IX. Discusiones

En esta sección, se discutirán sobre los resultados encontrados. Se busca analizar la viabilidad del algoritmo de *clustering*, del marco de trabajo SHAP y la calidad del conocimiento descubierto a través de los métodos utilizados. Durante el transcurso de esta sección, se utilizara la notación $CreColGlu$ para referenciar a la realización de controles médicos de creatinina sérica o urinaria, alguno de los tipos de colesterol (total, ldl y hdl) y glucemia. Esta combinación de prácticas será observada de forma recurrente durante las discusiones y utilizar su notación facilitará su la lectura de las mismas.

En la Sección IX-C se mencionan, de forma general y resumida, las conclusiones respecto a las discusiones de los *clusterings*. No se busca hacer un análisis detallado de los datos crudos obtenidos del experimento, sino presentar los resultados principales con un estilo amigable a la lectura.

Por otro parte, las Secciones IX-A y IX-B sí presentan las discusiones en detalle. Cada sección profundiza sobre uno de los respectivos *clusterings*. La linea de pensamiento consiste en primero analizar los *clusters* de manera directa y superficial para plantear un conjunto de hipotesis. Luego, se observarán los valores de SHAP de cada *cluster* para verificar, extender o negar el conjunto de hipotesis originalmente presentado.

Se recomienda la lectura de las Secciones IX-A a IX-C en caso de que se busque seguir el análisis paso a paso llevado a cabo en este trabajo, en donde se plantean ciertas hipótesis que pueden terminar siendo falsas en el transcurso de las comparaciones. Por otro lado, se puede leer directamente la Sección IX-C si solo son de interés las conclusiones finales de las pruebas.

IX-A. 2-clustering

En la Figura 7, se observan los dos grupos encontrados en el primer *clustering*. Estos tienen 3744 y 1284 historiales médicos respectivamente.

Lo primero a notar es que el primer grupo tiene una alta densidad del estado 0. Esto corresponde a un año de nulo consumo de alguna de las práctica o monodrogas analizadas. Algunos de los afiliados pasaron los 3 años de análisis sin ningún consumo de interés, aunque también pueden verse ciertos subconjuntos caracterizados por tener en alguno de los 3 años el estado 44400 o 400. Ambos estados marcan un control de creatinina, y el primer estado también corresponde a control de glucemia y colesterol total.

El segundo grupo tiene mayor frecuencia del estado 400, principalmente en el primer y segundo año de análisis, además de mantener cierta aparición del estado 44400. Esto indica que los controles de $CreColGlu$ se realizan de forma mucho mas recurrente que en el grupo 1. Además, los estados 44400 y 400 se encuentran secuenciados durante dos o tres años en mayor cantidad de afiliados, lo que indica aún una mayor recurrencia del control de creatinina.

En principio, se puede interpretar que la principal diferencia entre grupos corresponde a una alta o una baja tasa de realización de ciertas practicas médicas en el marco temporal desde 2021 a 2024. El primer grupo con menor actividad médica es el que más afiliados tiene, en contraposición al segundo grupo de mayor actividad médica. Es probable que el algoritmo haya agrupado en el segundo segmento, en términos generales, a los afiliados con un historial médico propio de la patología diabetes. Y, por otro lado, en el primer segmento de mayor densidad de afiliados, a aquellos sin una actividad recurrente en la realización de estas prácticas. Si bien muchos de los afiliados del primer grupo tienen registro de haber realizado las prácticas médicas analizadas, la cantidad de casos de dos o tres años de consumos consecutivos es menor, pudiendo ocasionarse por un control esporádico debido a causas ajenas a la diabetes.

Los valores de SHAP de los *clusters*, detallados en la Figura 9, dan una explicación de la importancia de las *features* que en su mayoría concuerda con el análisis hecho. En los mismos, se puede observar que la probabilidad de ser segmentado en el grupo 1 se ve positivamente afectada cuando, en ninguno de los años de análisis, el historial médico tiene consumos de control de $CreColGlu$.

Es decir, cuando las *features* binarias son negativas. La mayor población de estas observaciones se encuentran cercana a un valor de SHAP de $0{,}05$. Como ninguna de las *features* negativas domina en términos de impacto en el modelo a las demás, indica que el modelo predictivo espera que las mayoría de estas sean efectivamente negativas para que un historial médico sea exitosamente segmentado en este grupo. Dicho de otra forma, se busca que los estados de los años de un historial médico se asemejen lo máximo posible al estado 0, donde el consumo es nulo. Existen dos subconjuntos de *features* cuya interpretación difiere de lo explicado. El control de hemoglobina, a diferencia del de $CreColGlu$, afecta positivamente a la probabilidad de segmentación. Aún así, la distribución de puntos es constante a lo largo de los valores positivos de SHAP, indicando menor fuerza en este patrón. El efecto del control de hemoglobina no permite generalizar que el primer grupo se caracterice por una baja actividad médica general en todo tipo de medicamentos y prácticas, sino particularmente en los controles de $CreColGlu$. Luego, se encuentran las consultas oftalmológicas, sin un patrón claro a detallar. La puntuación negativa de estas *features* no tienen grandes impactos en la probabilidad de segmentado, pero a su vez, la puntuación positiva tiene tanto impacto positivo como negativo en la probabilidad. Esto indica que el atender a consultas oftalmológicas no es un diferenciador al momento de pertenecer o no a este grupo

Un comportamiento complementario se observa los valores de SHAP del segundo *cluster*. Los controles de creatinina, algún tipo de colesterol y glucemia afectan positivamente a la probabilidad de formar parte del *cluster*. Se observa mayor variabilidad hacia valores de SHAP altos en la distribución de puntos en los diferentes *features*. Esto puede indicar que puntuar positivamente y de manera independiente en alguna de las *features* tendrá un gran impacto positivo en la probabilidad de pertenecer a este *cluster*. Esto contrasta con la menor variabilidad hacia valores de SHAP altos en el primer *cluster*. Como se mencionó, este ultimo requiere una cooperación sinérgica en puntuar negativamente en las diferentes *features*, de forma de que el estado de los años se asemeje lo máximo posible al valor 0. Una pequeña cantidad de *features* que puntúen positivamente puede causar que la probabilidad de pertenecer al segundo *cluster* superé a la del primero, incluso aunque la cantidad total de *features* negativas sea mayor. El comportamiento poco diferenciador para las consultas oftálmologicas se mantiene, al igual que la contribución complementaria al primer *cluster* para el control de hemoglobina. Con esto se termina, de confirmar que el segundo *cluster*, si bien se caracteriza de un alto grado de controles de $CreColGlu$, también lo hace por baja recurrencia en el consumo de controles de hemoglobina.

Los valores de SHAP también proporcionan información adicional que un análisis directo y superficial de los *clusters* no brinda. Si bien se había logrado detectar la importancia del control de colesterol total a la hora de diferenciar entre los dos grupos, SHAP permite generalizar este comportamiento para otros tipos de colesterol como el hdl y el ldl. Este detalle, difícilmente observable en los *clusters* de forma directa, permite un mejor entendimiento de las diferencias entre historiales médicos e incluso simplificar futuros estudios sobre el conjunto de datos. Se podría agrupar los diferentes tipos de control de colesterol en un único grupo de análisis sin afectar la calidad del segmentado. Otra conclusión que pudo ser realizada gracias a SHAP fue el comportamiento complementario y dificilmente observable para las prácticas de hemoglobina. El *cluster* 1 se caracteriza por un bajo consumo de prácticas, a diferencia del *cluster* 2. Pero, aún así, existen excepciones como la del control de hemoglobina, cuyo comportamiento es opuesto al descrito para el general de prácticas.

IX-B. 10-*clustering*

En la Figura 8, se observan los diez grupos encontrados en el segundo *clustering*. Estos tienen 299, 252, 334, 2013, 277, 262, 441, 438, 477 y 235 historiales médicos respectivamente.

Lo primero a notar es que los *clusters* 2, 3, 7 y 9 se caracterizan por una alta aparición del estado 400 (correspondiente al control de creatinina) en forma de diferentes patrones. Se los puede

interpretar como una descomposición del segundo *cluster* del 2-*clustering* de la Sección IX-A. Se diferencian entre ellos en función de en que año de análisis aparece el estado 400. Los *clusters* 3, 7 y 9 incluyen historiales con control de creatinina en el tercer, segundo y primer año de análisis respectivamente, mientras que el segundo *cluster* tiene un consumo de la práctica uniforme a lo largo de los tres años. También se visualiza una aparición de menor frecuencia de los estados 44400, 74600 y 24200. En los estados 44400 y 74600 se mantiene la aparición del control de creatinina y se le suman al primer estado el control de glucemia y colesterol total, y al segundo estado el control de glucemia, los tipos de colesterol, hemoglobina y creatinina. El estado 24200 referencia a un control de los 3 tipos de colesterol. Con esto último, si bien no se puede afirmar aún que el colesterol, glucemia o hemoglobina sean prácticas características de estos *clusters*, si se puede brindar una intuición de que su aporte como *features* se ve sobrepasado por el aporte del control de creatinina.

Un análisis similar se puede realizar para los grupos 4, 6, 8 y 10, pensándolos como una descomposición del primer grupo del 2-*clustering*. En este caso, el estado que domina es el 0, es decir, un estado con nulo consumo. Aún así, existen diferencias importantes a destacar entre estos grupos. El grupo 4 incluye un subconjunto de historiales médicos que en alguno de sus 3 años de análisis tienen el estado 44400 o 400. Si bien no es una condición dominante del *cluster*, es un patrón lo suficientemente frecuente como para destacarlo. Además, el grupo 4 cuenta con una notable cantidad de historiales médicos con dos o más años en un estado diferente a 0. Estas observaciones disminuyen en gran medida para el grupo 6. La aparición de los estados 44400 y 400 es baja, así como de historiales con dos o más años de estado diferente a 0. Finalmente, en el grupo 8 estas observaciones son casi inexistentes. Si bien los tres grupos se caracterizan por una baja actividad médica, se pueden sub-categorizar en función de cuan baja es la misma. Finalmente, el grupo 10 agrupa los historiales con un nivel de baja actividad médica similar al del grupo 4, diferenciándose respecto a que estados tienen mayor aparición. Destaca el estado 24200 y un subconjunto de estados de menor relevancia caracterizados por el color negro.

El *cluster* 1 destaca, al igual que el *cluster* 10, por el consumo de los tres tipos de colesterol, pero con un mayor grado de actividad médica distribuido durante los 3 años de análisis. Algo similar sucede con el *cluster* 5 y los controles de $CreColGlu$. En conjunto, Estos dos *clusters* junto al 2, 3, 7 y 9 son los categorizados como de alta actividad médica, diferenciadose entre ellos por el estado que domina.

Los valores de SHAP, al igual que en la sección anterior, proporcionan información de gran relevancia para profundizar sobre el análisis del 10-*clustering*. Brindan ayuda para confirmar o rechazar conclusiones obtenidas de un análisis superficial de los *clusters*, así como para generar nuevas conclusiones de mayor profundidad.

Para los grupos 2, 3, 7 y 9, SHAP permite caracterizar las diferencias previamente encontradas entre estos. Por ejemplo, la conclusión previa de que el grupo 2 tiene valores de SHAP positivos cuando el historial médico incluye controles de $CreColGlu$ en cualquiera de los años de análisis. En este grupo, la distribución de valores de SHAP para las *features* que incluyen estas prácticas tienen relativamente poca variabilidad y promedian a un valor cercano a $0{,}05$. Este comportamiento puede indicar una aporte equitativo entre todas las *features*, para que un historial médico tenga alta probabilidad de pertenecer a este grupo mayormente cuando en todos sus años de análisis se realicen los respectivos controles médicos. Los valores SHAP también permiten confirmar las conclusiones previas de que los grupos 3, 7 y 9 priorizan historiales que incluyen los controles médicos discutidos unicamente en el tercer, segundo y primer año de análisis respectivamente.

Como se mencionó en esta misma sección, los *clusters* 2, 3, 7 y 9 del 10-*clustering* pueden interpretarse como una descomposición del segundo *cluster* del 2-*clustering*. En este ultimo, ya explicado en la Sección IX-A, el control de microalbuminuria y la consulta oftalmológica no tenían un patrón claro detectado. Ahora, en su descomposición en estos 4 *clusters*, se les puede encontrar ciertos patrones debilmente marcados. Los historiales médicos sin registro de control de hemoglobi-

na, microalbuminuria y de consulta oftalmológica tienen mayor probabilidad de pertenecer a estos *clusters*. Esto podría indicar que la separabilidad entre *clusters* es relativamente mejor que en el 2-*clustering*.

Por otro lado, SHAP introdujo cuestionamientos sobre al análisis hecho respecto a los grupos 4, 6, 8 y 10. Siguiendo la linea de pensamiento presentada, los historiales sin registro de controles de $CreColGlu$ aumentan la probabilidad de pertenecer a estos grupos. Pero estas *features* distan de ser las que mayor impacto tienen en el modelo predictivo. Esto quiere decir que la conclusión inicial de que estos grupos se caracterizan principalmente por una baja actividad médica (o baja actividad en $CreColGlu$) es incompleta. El sexto *clusters* se ve principalmente representado por historiales médicos con consultas oftalmológicas en el primer y segundo año de análisis, mientra que el octavo por aquellos con consultas oftalmológicas en el tercer año. El *cluster* 10 se caracteriza principalmente por tener altos valores de SHAP ante registros de consumo de la monodroga metformina. Esto lo vuelve el primer *cluster* analizado cuyo principal elemento diferenciador se rige por una monodroga y no por una práctica médica. El grupo 4 mantiene los patrones obtenidos desde un análisis directo. Este *cluster*, efectivamente, corresponde a historiales médicos con la menor cantidad de actividad médica.

Las conclusiones de los *clusters* 1 y 5 también se encuentran incompletas sin incluir un análisis de SHAP. Para el primero de ellos, los valores de SHAP indican que el principal elemento de impacto son los controles de hemoglobina en cualquiera de los 3 años de análisis. En segundo grado de impacto, los controles de $CreColGlu$ en el segundo año de análisis también aumentan la probabilidad de inclusión en el *cluster*. En tercer y ultimo grado de impacto, estas ultimas prácticas hechas en el primer y tercer año, con valores de SHAP promediando a 0,02, aumentan con menor consistencia la probabilidad de inclusión. El *cluster* 1 puede ser interpretado como una intersección de los *clusters* del 2-*clustering*, en la que se agrupan aquellos historiales con actividad tanto en control de hemoglobina como en control de $CreColGlu$.

Para el *cluster* 5, si bien se cumple la conclusión inicial de un impacto positivo por controles de creatinina y glucemia en cualquiera de los años, no se mantiene para el colesterol. En ese caso, la realización de un control de alguno de los tipos de colesterol es de los principales factores en impactar negativamente en la probabilidad de inclusión en el *cluster*.

IX-C. Resumen de las discusiones

En las Figuras 4 y 5 se observan unos esquemas que resumen las conclusiones para el 2-*clustering* y el 10-*clustering*. En estas figuras, se representa a los *clusters* como conjuntos que puede ser interpretados como descomposición o intersección de otros *clusters*. Estas relaciones no buscan ser matemáticamente exactas y rigurosas. Su objetivo es brindar una intuición sencilla e informal a lo que se puede interpretar de los *clusterings*.

En el 2-*clustering*, sus grupos se diferencian principalmente en como los controles de $CreColGlu$ y hemoglobina estan presentes en los historiales médicos. Los afiliados del grupo 1 tienen historiales médicos con mayor presencia de control de hemoglobina, a diferencia de los afiliados del grupo 2 con mayor presencia de control de $CreColGlu$.

En el 10-*clustering*, se generan dos sub-agrupaciones de *clusters*, formadas por los numerados como 4, 6, 8 y 10 por un lado y 2, 3, 7 y 9 por el otro. La primer sub-agrupación se puede interpretar como una descomposición del primer grupo del 2-*clustering*. Todos estos grupos se caracterizan por baja aparición de controles de $CreColGlu$. Además, los grupos 6 y 8 también lo hacen por la aparición de consultas oftalmológicas en el primer y segundo año de análisis y en el tercer año respectivamente, mientras que el grupo 10 se caracteriza por el consumo de metformina.

Por otro lado, la segunda sub-agrupación se puede interpretar como una descomposición del segundo grupo del 2-*clustering*. Todos estos grupos se caracterizan por una baja aparición de control de hemoglobina. El grupo 2 tiene, además, controles de $CreColGlu$ en los 3 años de análisis. Los grupos 3, 7 y 9 tienen control de $CreColGlu$

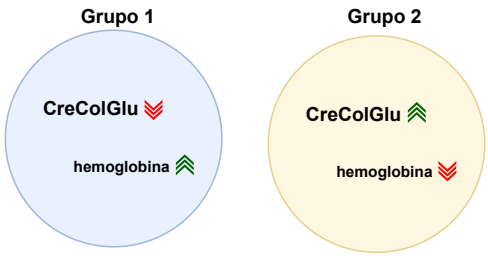


Figura 4: Resumen gráfico de las discuciones planteadas sobre el 2-*clustering*.

unicamente en el tercer año, segundo año y primer año respectivamente.

El grupo 5 se puede interpretar como la intersección entre los grupos del 2-*clustering*. Es decir, historiales médicos con alta aparición de control de hemoglobina (propio del primer *cluster* del 2-*clustering*) y alta aparición de control de $CreColGlu$ (propio del segundo *cluster* del 2-*clustering*).

Finalmente, el grupo 1 es independiente y descompone $CreColGlu$, caracterizándose por controles de creatinina y glucemia y por poca aparición de controles de colesterol.

X. Limitaciones en pruebas experimentales

La complejidad espacial exponencial de la cantidad posible de estados para los marcos temporales introdujo limitaciones importantes. En primer lugar, limita la capacidad de escalabilidad del flujo de *Machine Learning* a conjuntos de datos de mayor complejidad. Al no poder superar la barrera de los 20 elementos de interés dado el *hardware* disponible, en este trabajo se tuvo que omitir el uso de las prácticas médicas relacionadas a condiciones cardiovasculares. Además, fue la causa del agrupamiento de monodrogas realizado en la Sección VI-E. Este agrupamiento, si bien resulto certero y permitió encontrar patrones de interés, puede verse afectado por sesgos introducidos por los autores. Este problema también afecto a la capacidad de visualización de los mapas de calor de los *clusters*. Como se detalla en la Sección A, no existe la cantidad necesaria de colores distinguibles al ojo humano para caracterizar a cada uno de los estados que hacen aparición en los *clusters*. Para mitigar esto, se se le asigno colores unicamente a los estados de mayor aparición y se generalizó al resto con un único color negro. Esto puede estar sesgando el análisis y generando perdida de información relevante para el mismo.

XI. Prototipo de software

En esta sección, se introduce el prototipo de software diseñado e implementado. El objetivo es establecer un marco de trabajo para el desarrollo futuro de una plataforma de inteligencia médica. Esta asienta las bases para que, en futuras mejoras del prototipo, el equipo de Auditoría Médica de SOSUNC pueda utilizar técnicas de estadística avanzada y *Machine Learning* para el estudio de los datos disponibles en la obra social. Por definición conceptual, el prototipo actual no esta preparado para un entorno de producción real, ni garantiza su funcionamiento ante todo tipo de casos de uso.

XI-A. Diseño

El prototipo tiene un diseño inspirado en la arquitectura cliente-servidor, el cual puede ser visto en la figura 6. La modularidad, escalabilidad y separación de aspectos son principios claves que rigen a este diseño, con el fin de facilitar su desarrollo incremental en trabajos futuros. Además, esto facilitará un despliegue y mantenimiento a través de plataformas Web.

Del lado del servidor, se conceptualiza la lógica algorítmica encargada de llevar a cabo el flujo de *Machine Learning*. Para mantener un diseño limpio de lógica repetida, se decidió separar este flujo en dos etapas independientes entre sí: los Dominios y los Algoritmos. Un Dominio es una abstracción que modela un determinado área de conocimiento, o conjunto de datos, que puede ser utilizado por un Algoritmo de *Machine Learning*. De esta manera, un Algoritmo se abstrae del conjunto de datos con el que sera utilizado, permitiendo su reuso para diferentes Dominios.

Siguiendo el marco de este trabajo, un posible Dominio incluye a los historiales médicos y su respectiva matriz de similitud. Un posible Algoritmo

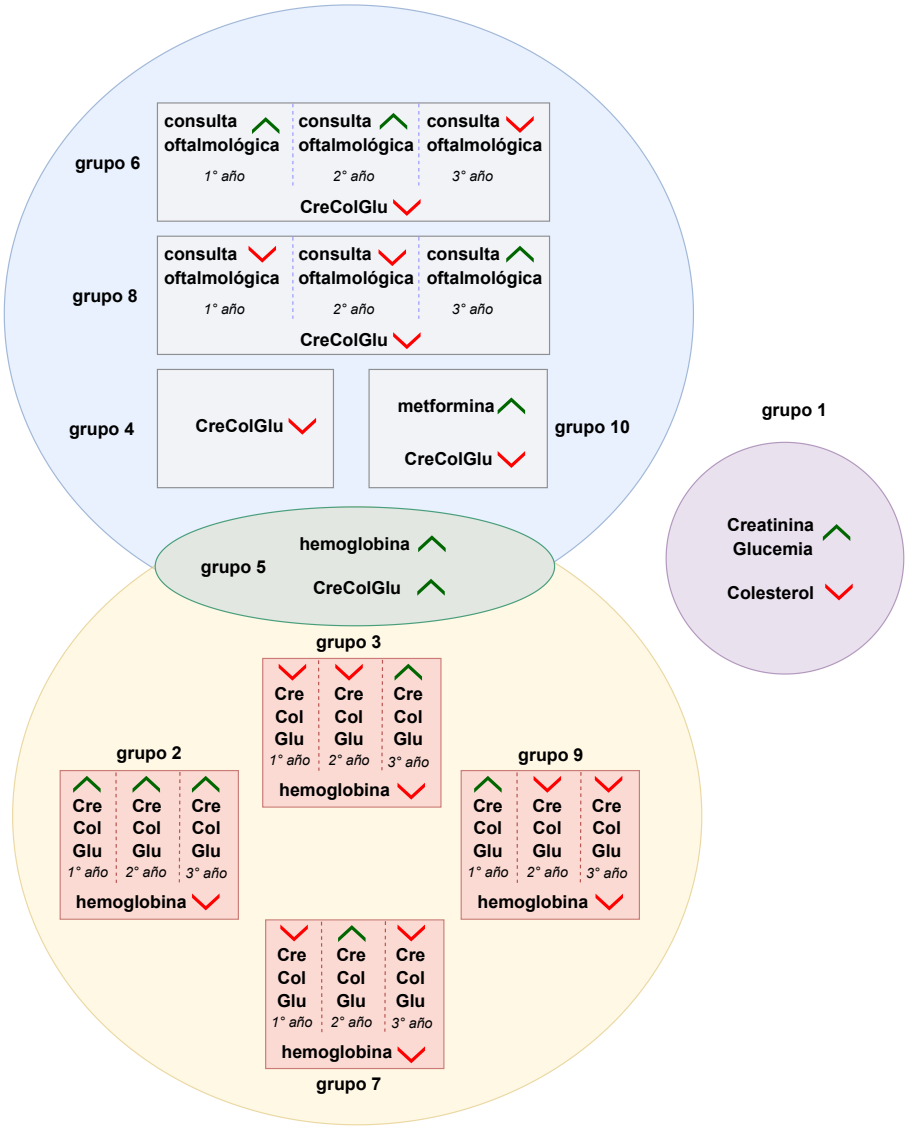


Figura 5: Resumen gráfico de las discuciones planteadas sobre el 10-*clustering*. Algunos *clusters* se representan como conjuntos que pueden ser interpretados como descomposición o intersección de otros *clusters*. Estas relaciones no buscan ser matemáticamente exactas y rigurosas. Su objetivo es brindar una intuición sencilla e informal a lo que se puede interpretar de los *clusterings*.

es el *clustering* de secuencias. De esta manera, los historiales médicos podrían ser utilizados en otras tareas de aprendizaje (por ejemplo, clasificación o regresión) y el Algoritmo de *clustering* de secuencias podría utilizar otros conjuntos de secuencias diferentes.

Los módulo de Dominios y Algoritmos pueden ser manipulados directamente desde el cliente para una configuración especifica de los mismos. Esto, si bien permite una personalización a medida de las necesidades de cada caso de uso, requiere un conocimiento de alto nivel de la arquitectura del servidor. En algunos casos, puede no requerirse de configuraciones avanzadas para la creación de Dominios y uso de los Algoritmos, siendo suficiente con una inicialización genérica y sencilla de los mismos que facilite el proceso de desarrollo. Motivado por esta posible situación, se diseñan las Aplicaciones, que funcionan como interfaces de alto nivel para estas configuraciones complejas.

Desde el lado cliente, se mantienen las interfaces de usuarios, compuestas por pantallas, componentes lógico-gráficos y controladores. Las pantallas definen componentes visuales básicos con nula o poca lógica de procesamiento de datos. Son la capa más externa del cliente, ya que es con la que el usuario interactúa directamente. En el momento que se requiera definir un componente visual con una mayor complejidad en su lógica y funcionamiento, se definen los componentes lógico-gráficos. Estos componentes son declarados en las pantallas, abstrayendo a las mismas de la lógica que definen. Otro utilidad de los componentes lógico-gráficos es la reutilización de los mismo en diferentes pantallas con casos de uso similares.

Por ultimo, los controladores funcionan como intermediarios entre las pantallas y el servidor. Realizan un preprocesamiento de los datos ingresados por el usuario y recibidos por el servidor, preparando su estructura para ser correctamente procesadas o visualizadas por los respectivos componentes.

XI-B. Limitaciones del prototipo actual

Tanto el diseño como la implementación del prototipo cuentan con limitaciones que son importantes considerar. En esta primera versión, no se ha diseñado ni implementado nada referido al marco de trabajo SHAP. Se intuye que el mismo puede ser incorporado como un tipo de Algoritmo en la lógica algorítmica del servidor. Esto, si bien permite no realizar grandes modificaciones a la estructura general, puede llevar a otro tipo de problemas. Si se optara por considerar a SHAP un Algoritmo, el mismo tendría que servir de soporte explicativo a otros Algoritmos de *Machine Learning*, y no como un flujo de procesamiento para cierto Dominio. Esto genera contradicciones semánticas en la definición original del concepto Algoritmo en el diseño del servidor.

Otra opción, que requiere mayor modificación del servidor pero podría solucionar el problema semántico respecto a la propuesta anterior, es crear un nuevo tipo de entidad conceptual. Suponiendo que sea interés explorar, diseñar e implementar nuevas técnicas explicativas para *Machine Learning*, es factible incorporar la entidad Explicador/Razonador.

Otra importante limitación es la forma de acceder a los conjuntos de datos iniciales y de almacenar los *clustering* y otros conjuntos de datos resultantes. Para un entorno de producción ideal, se debería de crear un flujo de datos interconectado a través de tecnologías tales como bases de datos, *pipelines* de datos y almacenamiento en la nube. Para el desarrollo de este prototipo no se tuvo acceso al entorno de producción de SOSUNC, por lo que los datos de entrada deben ser subidos a la interfaz en forma de archivos CSV. De la misma forma, los *clusters* y otros resultados deben descargarse en el mismo formato.

La implementación, como se mencionó en la sección V, fue realizada en lenguaje Python y con la librería de interfaces de usuario Streamlit. Python y Streamlit se caracterizan permitir una implementación veloz y sencilla gracias a su sintaxis altamente legible y de alto nivel. Por un lado, los vuelve un conjunto de herramientas ideales para la realización de prototipos, pero pueden verse limitados en aplicaciones de gran escala. En caso de de haber interés en seguir trabajando este prototipo para llevarlo a escalas de alto nivel productivo, es factible que el mismo no logre el rendimiento esperado. Una

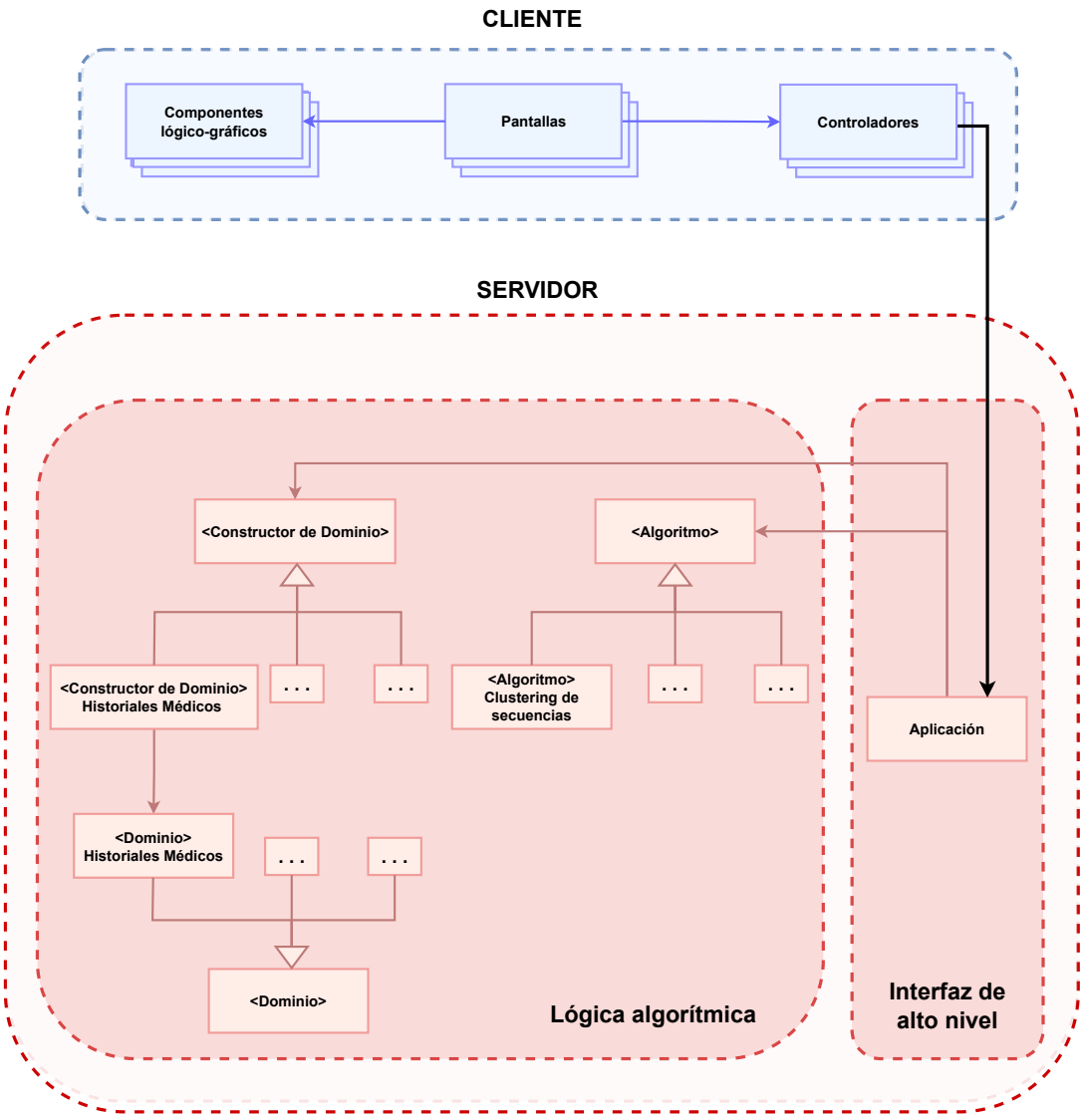


Figura 6: Diseño del prototipo de software. Se inspira en la arquitectura cliente-servidor, donde el cliente define las interfaces de usuario y el servidor la lógica algorítmica del flujo de *Machine Learning*.

opción posible es considerar migrar la implementación del prototipo a otras tecnologías de mayor rendimiento para el desarrollo de software.

Finalmente, existen un conjunto *bugs* visuales, de control de interfaz y de experiencia de usuario que no fueron solucionados ya que se alejan del objetivo general de este trabajo.

XI-C. Instrucciones de uso

El prototipo se encuentra disponible para su uso en la Nube y funcional al dia del 21 de Noviembre de 2024. Se puede acceder al mismo a través de [1]. Particularmente, se utilizó el servicio *Streamlit Cloud*, que permite desplegar aplicaciones de Streamlit de manera gratuita.

Dentro de la aplicación, se debe acceder a la sección `Principal` del módulo de `Agrupamiento de historiales médicos`, que puede ser vista en la Figura 15.

Se deben cargar, en los datos de entrada, el conjunto de datos de consumo de prácticas médicas por afiliado y las prácticas de interes, es decir, las de diabetes. Estos archivos CSV se encuentran en el repositorio del proyecto [2], en el directorio `dataset/raw`. El nombre de los mismos son `cons_pract_medicas.csv` y `practicas_diabetes.csv`. Luego, en los datos de entrada, tambien debe configurarse para que los historiales médicos esten formados por 3 años. Finalmente, en los hyperparámetros se debe configurar a una cantidad de 2 o 10 grupos para el *clustering*.

Al seleccionar el botón *Ejecutar*, comenzará el procesamiento del flujo de *Machine Learning*. El mismo puede llevar varios minutos. El avance de su estado sera visualizado por pantalla. Al finalizar, se observará el estado final de la sección como se ve en la Figura 16.

Se puede seleccionar cada una de las solapas de *clusters* para visualizar el respectivo mapa de calor. Ademas, se puede interactuar con los bloques de un mapa de calor. Al mantener el puntero sobre estos, se visualiza a que afiliado y marco temporal de análisis corresponde ese bloque. También muestra a que estado corresponde el color del mismo. Al seleccionarlo con el puntero, esta información será visualizada por debajo del mapa de calor. Si se selecciona el botón etiquetado como `Más info.`, se accede a la segunda pantalla desarrollada en este prototipo.

Esta segunda pantalla permite ver los detalles de consumo de cada código hexadecimal de los estados de marcos temporales. En la Figura 17, tras seleccionar el botón `Más info.`, la aplicación redirige automáticamente al usuario a la segunda pantalla. Aquí, se mostrarán que consumos de práctica médicas y monodrogas corresponden al estado que fue seleccionado previo a la activación del botón. El usuario también puede buscar de manera directa otros estados haciendo uso de la barra de busqueda incorporada en la pantalla.

Volviendo a la pantalla principal, el usuario puede usar el deslizador ubicado arriba de las solapas de *clusters* para configurar la *silhouette* mínima. Todo *cluster* cuya *silhouette* no sea igual o mayor a este valor seleccionado sera invisibilizado para el usuario.

Si se quiere repetir la ejecución de un *clustering* para los **mismos datos cargados en la sección de Datos de entrada**, se recomienda descargar la matriz de disimilitud e historiales médicos generados la primera vez. Estos datos son los que involucrán mayor tiempo de cómputo y pueden ser reutilizados si unicamente se cambiarán los hyperparámetros. Para la descarga, se hace uso de los botones ubicados en la sección inferior de la pantalla. Para cargar estos nuevos archivos debe hacerse en la sección de Datos de entrada.

XII. TRABAJOS FUTUROS

Dada la longitud del trabajo realizado, se abren diferentes lineas para la implementación de mejoras y la realización de nuevas investigaciones.

La limitación planteada en la Sección X respecto a la cantidad total de datos utilizada da pie a repetir el flujo de *Machine Learning* diseñado con otros conjuntos de datos. Se podrían realizar otras combinaciones de elementos de interés, tales como las prácticas médicas de diabetes con las de enfermedades cardiovasculares.

También sería conveniente profundiza en el ya mencionado problema de la cantidad exponencial

de estados de un marco temporal en un historial médico. Estudiar otro tipo de representaciones para los historiales podría ayudar a solucionar esta condición inherente al actual diseño de los mismos.

En simultaneo, sería de utilidad estudiar otro tipo de métodos de visualización para los historiales médicos diseñados. Bajo el supuesto de que no se logre encontrar una mejor representación para los historiales, el problema de la cantidad de estados podría mitigarse con métodos de visualización diferentes al uso de mapas de calor.

Respecto al marco de trabajo SHAP, se propone investigar y evaluar la viabilidad de implementar POSHAP [6] para el dominio de SOSUNC. POSHAP es una extensión de SHAP para modelos predictivos entrenados con secuencias. Esto podría permitir un mejor análisis sobre los resultados, ya que se adecuaría de mejor manera a la estructura secuencial en la que se basan los historiales médicos. Por limitaciones de tiempo, el uso de POSHAP no fue contemplado dentro de los objetivos principales de este trabajo.

Finalmente, se propone implementar las correcciones y y nuevas características para el prototipo planteadas en la Sección XI-B.

XIII. Conclusiones

En este trabajo se diseñó e implementó un flujo de *Machine Learning* para aplicar técnicas de *clustering* explicable sobre el dominio de la obra social SOSUNC. Se trabajo con historiales médicos estructurados como secuencias, el algoritmo de k-medoides para el *clustering* y el marco de trabajo $SHAP$ para brindar explicabilidad sobre la segmentación realizada. Los resultados obtenidos permitieron caracterizar el consumo de prácticas médicas y monodrogas de los afiliados, encontrando diferentes patrones entre estos. Se demostró viable la utilización de *clustering* y de SHAP para la el descubrimiento de conocimiento que pueda ser relevante para las actividades diarias del area de Auditoría Médica de SOSUNC. Además, se diseñó e implementó un prototipo de software para la aplicación de herramientas de *Machine Learning* por parte de personal no experimentado en esta área. Este prototipo introduce un diseño modular que facilita futuras mejoras e integraciones al mismo.

XIV. Disponibilidad del código

El código implementando para este trabajo se encuentra en un repositorio de GitHub [2]. En el directorio `notebooks_experimentation` se encuentran las pruebas de los algoritmos y en `ml_application_suite` el prototipo de software.

Referencias

[1] Antonella Torres Adriano Lusso, 2024. URL https://suite-aplicaciones-aprendizaje-automatico.streamlit.app/create_Screen_SequencesClustering. Accessed: Nov. 21, 2024.

[2] Antonella Torres Adriano Lusso, 2024. URL https://github.com/AdrianoLusso/Clustering_healthcare_sequences. Accessed: Nov. 21, 2024.

[3] Zeynep Akata, Dan Balliet, Maarten De Rijke, Frank Dignum, Virginia Dignum, Guszti Eiben, Antske Fokkens, Davide Grossi, Koen Hindriks, Holger Hoos, et al. A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[4] Roméo Baulain, Jérémy Jové, Dunia Sakr, Marine Gross-Goupil, Magali Rouyer, Marius Puel, Patrick Blin, Cécile Droz-Perroteau, Régis Lassalle, and Nicolas H Thurin. Clustering of prostate cancer healthcare pathways in the french national healthcare database. *Cancer Innovation*, 2(1):52–64, 2023.

[5] Pavlos Delias, Michael Doumpos, Evangelos Grigoroudis, Panagiotis Manolitzas, and Nikolaos Matsatsinis. Supporting healthcare management decisions via robust clustering of event logs. *Knowledge-Based Systems*, 84:203–213, 2015.

[6] Quinn Dickinson and Jesse G Meyer. Positional shap (poshap) for interpretation of machine learning models trained from biological

sequences. *PLOS Computational Biology*, 18 (1):e1009736, 2022.

[7] Pierpaolo D'Urso, Carmela Cappelli, Dario Di Lallo, and Riccardo Massari. Clustering of financial time series. *Physica A: Statistical Mechanics and its Applications*, 392(9):2114–2129, 2013.

[8] Alexis Gabadinho, Gilbert Ritschard, Nicolas S Müller, and Matthias Studer. Analyzing and visualizing state sequences in r with traminer. *Journal of statistical software*, 40:1–37, 2011.

[9] Paolo Giordani, Maria Brigida Ferraro, Francesca Martella, Paolo Giordani, Maria Brigida Ferraro, and Francesca Martella. *Introduction to clustering*. Springer, 2020.

[10] Ramzi A Haraty, Mohamad Dimishkieh, and Mehedi Masud. An enhanced k-means clustering algorithm for pattern discovery in healthcare data. *International Journal of distributed sensor networks*, 11(6):615740, 2015.

[11] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825): 357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

[12] David B Henry, Patrick H Tolan, and Deborah Gorman-Smith. Cluster analysis in family psychology research. *Journal of Family Psychology*, 19(1):121, 2005.

[13] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[14] Takashi Izumo and Yueh-Hsuan Weng. Coarse ethics: how to ethically assess explainable artificial intelligence. *AI and Ethics*, 2(3):449–461, 2022.

[15] Mouad Louhichi, Redwane Nesmaoui, Marwan Mbarek, and Mohamed Lazaar. Shapley values for explaining the black box nature of machine learning model clustering. *Procedia Computer Science*, 220:806–811, 2023.

[16] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[17] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

[18] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2023. URL https://CRAN.R-project.org/package=cluster. R package version 2.1.6 — For new features, see the 'NEWS' and the 'Changelog' file in the package source).

[19] Neerja Negi and Geetika Chawla. Clustering algorithms in healthcare. In *Intelligent healthcare: Applications of ai in ehealth*, pages 211–224. Springer, 2021.

[20] Godwin Ogbuabor and FN Ugwoke. Clustering algorithm for a healthcare dataset using silhouette score value. *Int. J. Comput. Sci. Inf. Technol*, 10(2):27–37, 2018.

[21] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830,

2011.

[23] Girish Punj and David W Stewart. Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, 20(2):134–148, 1983.

[24] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[25] Erich Schubert and Lars Lenssen. Fast k-medoids clustering in rust and python. *Journal of Open Source Software*, 7(75):4183, 2022.

[26] Erich Schubert and Peter J Rousseeuw. Faster k-medoids clustering: Improving the pam. *CLARA, and CLARANS Algorithms. arXiv e-prints, page*, 2018.

[27] Erich Schubert and Peter J Rousseeuw. Fast and eager k-medoids clustering: O (k) runtime improvement of the pam, clara, and clarans algorithms. *Information Systems*, 101:101804, 2021.

[28] Angela Tsay. Sequence analysis and optimal matching in sociology: Review and prospect. *Sociological Methods Res*, 29(1):3–33, 2000.

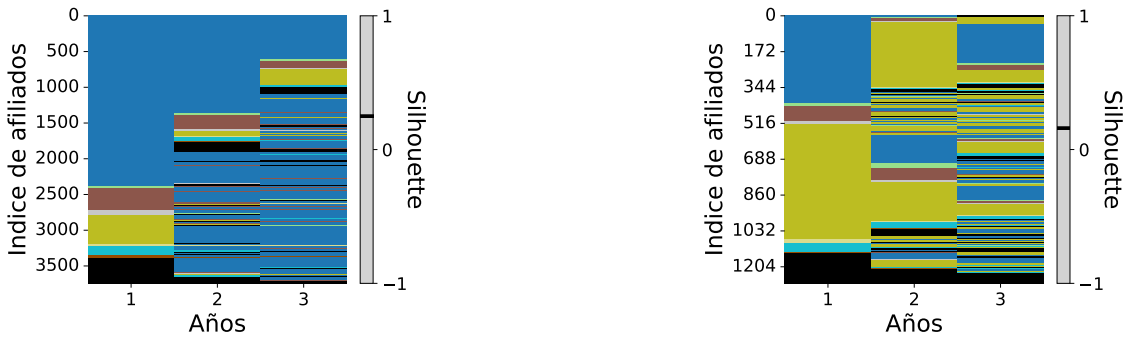
[29] Mark Van der Laan, Katherine Pollard, and Jennifer Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

[30] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

APÉNDICE A
CLUSTERS

Se muestran los gráficos correspondientes a los *clusters* del 2-*clustering* y el 10-*clustering* en las Figuras 7 y 8 respectivamente. Estos serán visualizados en mapas de calor que permiten describir una matriz de datos identificando a cada elemento con un color irrepetible. De esta manera, un elemento (i, j, k) en el mapa de calor indica que en el año i el afiliado j tiene un estado k, siendo k el color asignado a uno de los $2^{|\Delta|}$ posibles estados de año. Como se mencionó en la Sección VI-E, los identificadores de afiliados no corresponden a un dato sensible de los mismos. Por lo tanto, en los mapas de color se utilizan indices genéricos para visualizar los historiales médicos.

A causa del número exponencial de posibles estados, se vuelve poco práctico representar todos estos valores con un color irrepetible y reconocible al ojo humano. Por lo tanto, para facilitar la lectura de los *clusters*, se le asignará un color a unicamente a los 20 estados de mayor frecuencia en los historiales médicos. El resto de estados serán generalizados con el color negro. Los mismos pueden ser vistos en la Tabla II, junto a la descripción de que consumos de prácticas y monodrogas de interés representan.



Grupo 1 Grupo 2

Figura 7: Mapas de calor de 2-*clustering* para historiales médicos de 3 años. En el eje X se muestran los años de análisis, mientras que en el eje Y se muestran unos índices genéricos de afiliados. El mapa izquierdo corresponde al grupo 1, y el derecho al grupo 2.

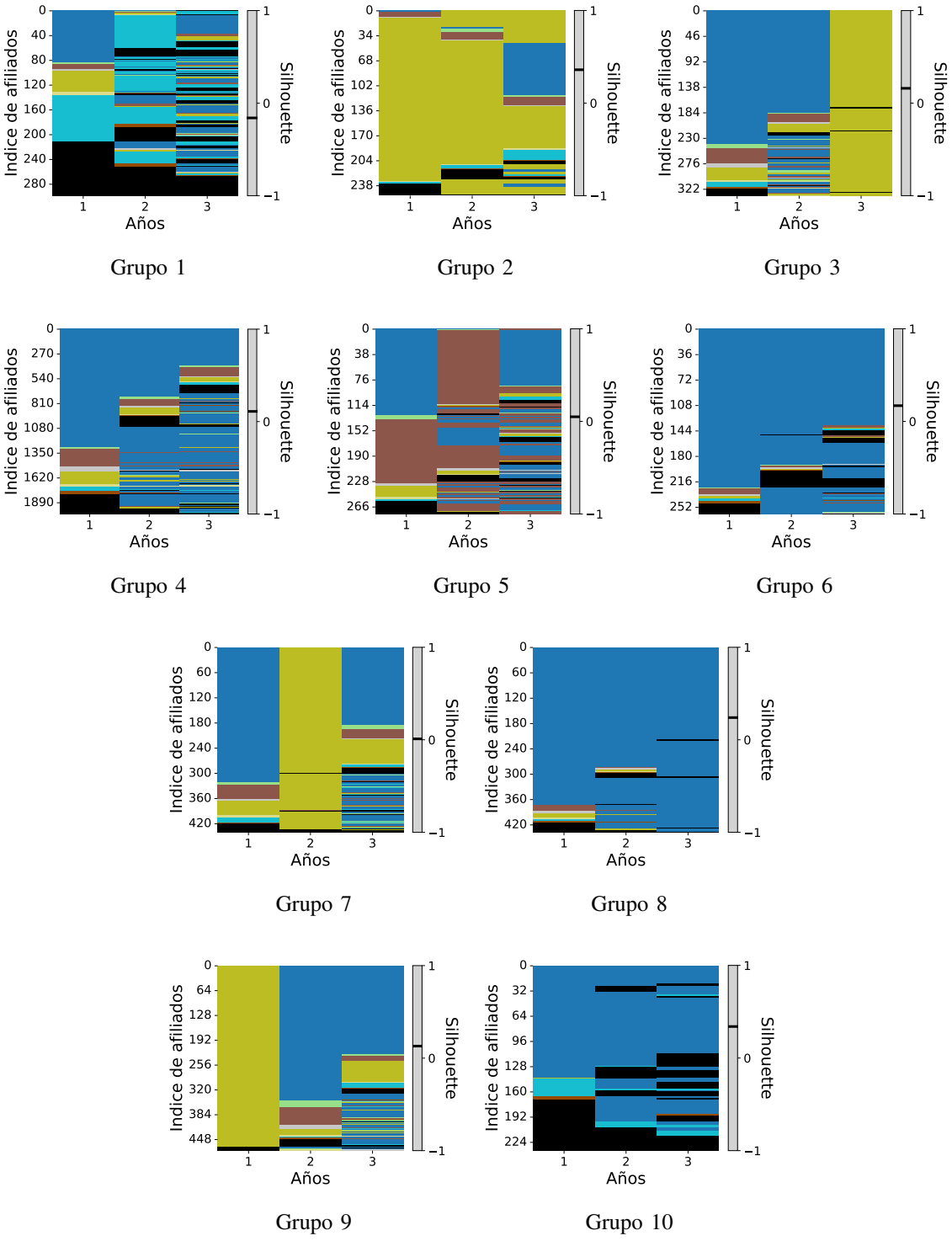
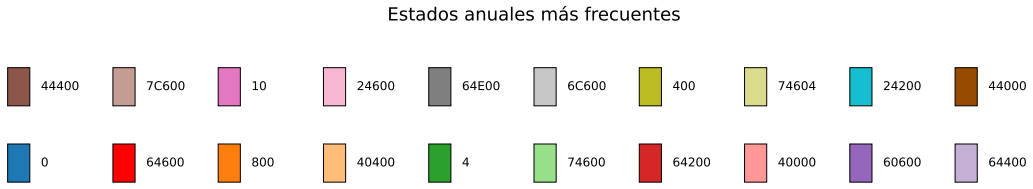


Figura 8: Mapas de calor de 10-*clustering* para historiales médicos de 3 años. En el eje X se muestran los años de análisis, mientras que en el eje Y se muestran unos índices genéricos de afiliados. Cada mapa esta etiquetado con el grupo que representa.



Código	Prácticas	Monodrogas
44400	glucemia, colesterol total, creatinina serica o urinaria	
7C600	glucemia, colesterol hdl, hemoglobina glicosilada, microalbuminuria, colesterol total, creatinina serica o urinaria, ldl colesterol	
10	insulina	
24600	colesterol hdl, colesterol total, creatinina serica o urinaria, ldl colesterol	
64E00	glucemia, colesterol hdl, colesterol total, consulta oftalmológica, creatinina serica o urinaria, ldl colesterol	
6C600	glucemia, colesterol hdl, microalbuminuria, colesterol total, creatinina serica o urinaria, ldl colesterol	
400	creatinina serica o urinaria	
74604	glucemia, colesterol hdl, hemoglobina glicosilada, colesterol total, creatinina serica o urinaria, ldl colesterol	metformina
24200	colesterol hdl, colesterol total, ldl colesterol	
44000	glucemia, colesterol total	
0		
64600	glucemia, colesterol hdl, colesterol total, creatinina serica o urinaria, ldl colesterol	
800	consulta oftalmológica	
40400	creatinina serica o urica	
4		metformina
74600	glucemia, colesterol hdl, hemoglobina glicosilada, colesterol total, creatinina sérica o urica, ldl colesterol	
64200	glucemia, colesterol hdl, colesterol total, ldl colesterol	
40000	glucemia	
60600	glucemia, colesterol hdl, creatinina serica o urinaria, ldl colesterol	
64400	glucemia, colesterol hdl, colesterol total, creatinina serica o urinaria	

Cuadro II: Estados más frecuentes en los *clusterings*, sus respectivas prácticas y monodrogas de interés, y el color que los representa en en los mapas de calor.

APÉNDICE B
VALORES DE SHAP

Se muestran los gráficos correspondientes a los valores de SHAP del 2-*clustering* y el 10-*clustering*, en la Figura 9 y Figuras 10 a 14 respectivamente. Estos serán visualizados en Enjambres de Abejas, o *beeswarms*, un estilo de gráfico implementado en librería de SHAP.

En el eje Y de cada gráfico, se visualizan unicamente las 20 *features* que mayor impacto tienen en la probabilidad de inclusión de un historial médico en el respectivo *cluster*. El eje X corresponde a los posibles valores de SHAP que un historial médico puede puntuar para cada *feature*. Para identificar entre prácticas y monodrogas, se etiqueta con *pr_* y *md_* respectivamente. Cada fila de n puntos en un gráfico corresponde a los valores de SHAP que cada uno de los n historiales médicos, pertenecientes al respectivo *cluster*, puntuó en la respectiva *feature*. Las *features* en esta codificación son binarias. De esta manera, si un punto es color rosa para una *feature*, indica que su valor es verdadero. Es decir, que el afiliado hizo un consumo del elemento de interés y año indicado en la *feature*. Para un punto azul, se describe la situación contraria.

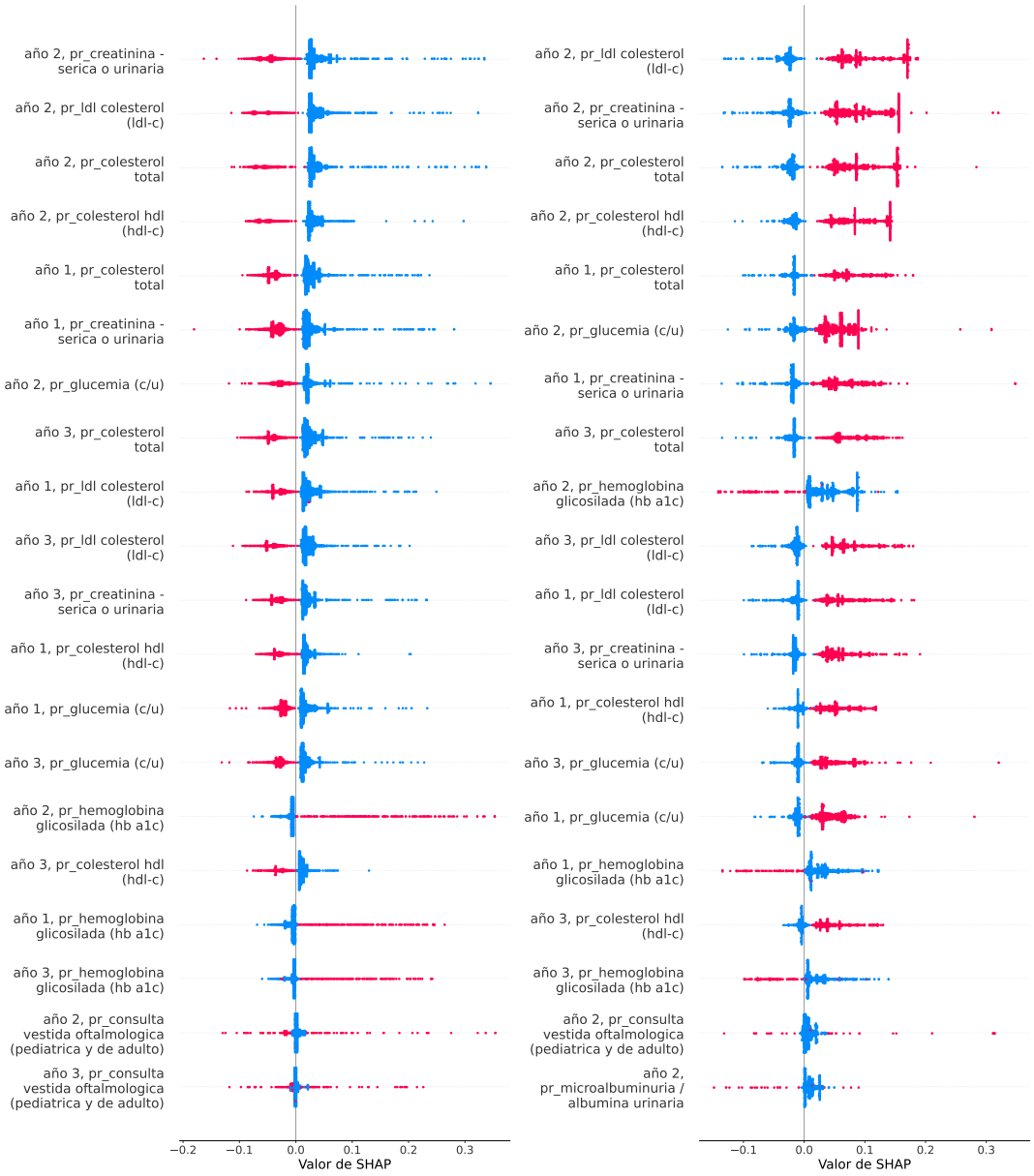


Figura 9: Valores de SHAP para el 2-*clustering*. El gráfico de la izquierda corresponde al primer *cluster*, mientras que el de la derecha corresponde al segundo. Cada fila del gráfico izquierdo, correspondiente a una *feature*, tendrá 3794 puntos, mientras que para el derecho cada una tendrá 1284 puntos.

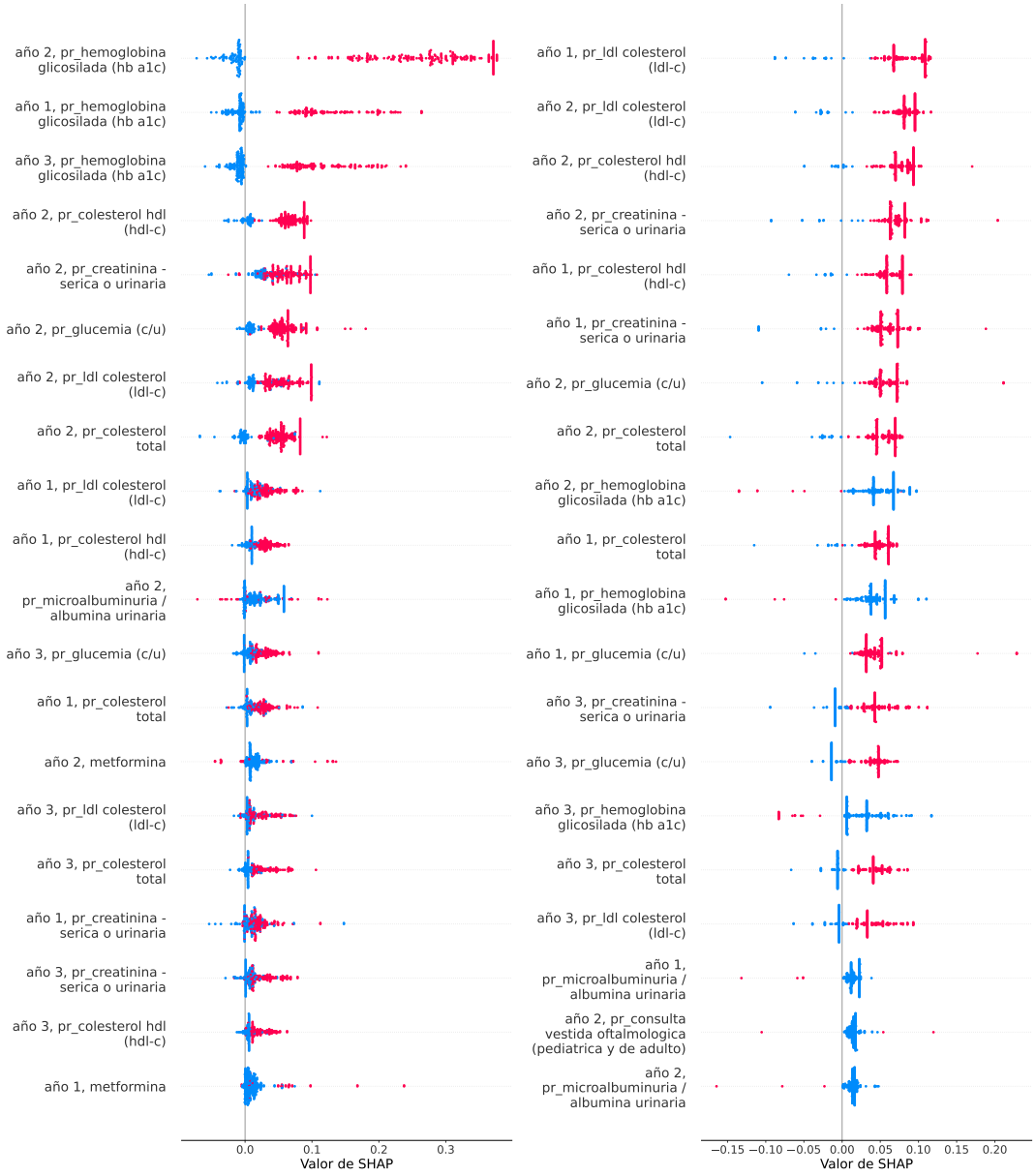


Figura 10: Valores de SHAP para el 10-*clustering*. El gráfico de la izquierda corresponde al primer *cluster*, mientras que el de la derecha corresponde al segundo. Cada fila del gráfico izquierdo, correspondiente a una *feature*, tendrá 299 puntos, mientras que para el derecho cada una tendrá 252 puntos.

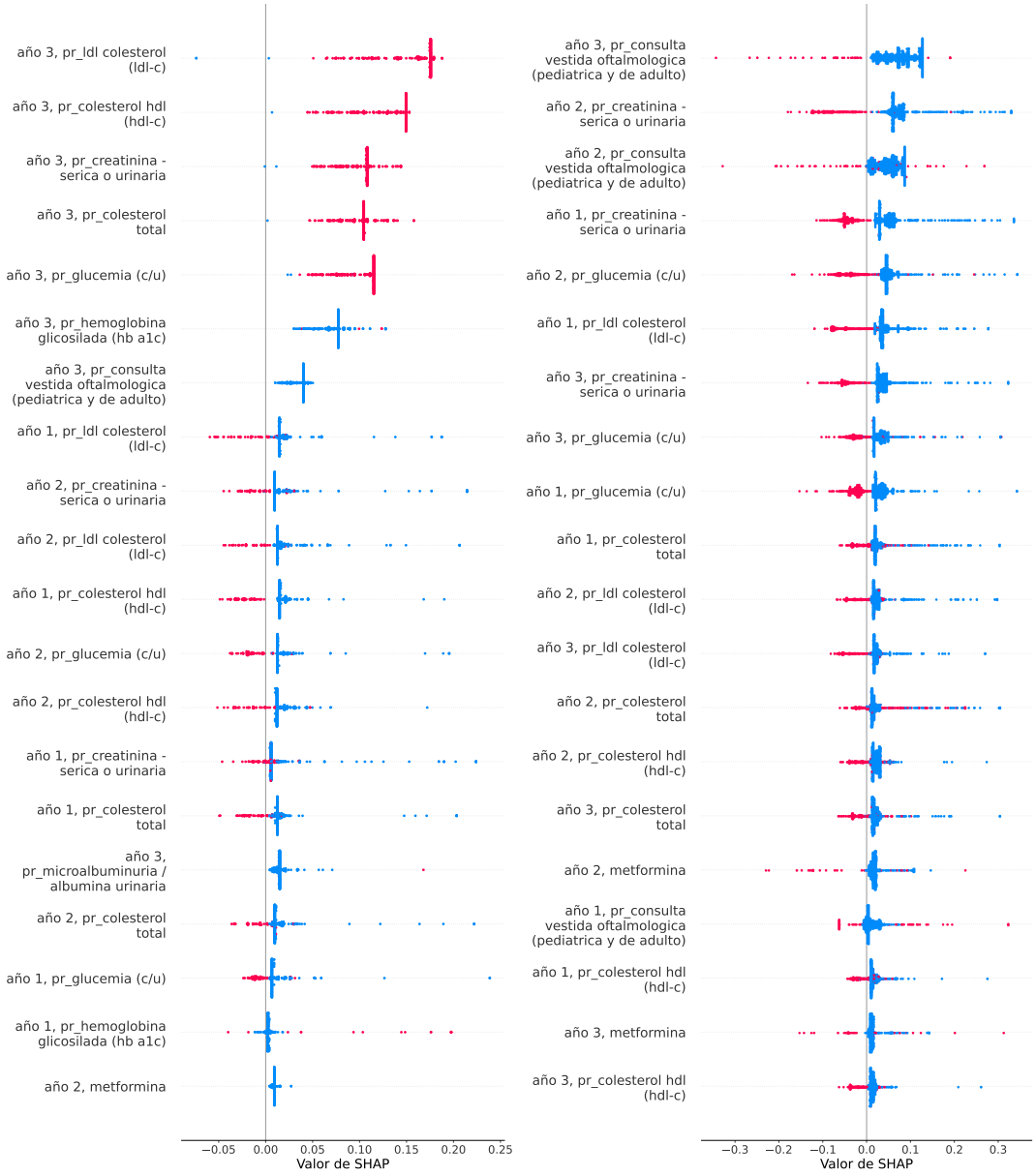


Figura 11: Valores de SHAP para el 10-*clustering*. El gráfico de la izquierda corresponde al tercer *cluster*, mientras que el de la derecha corresponde al cuarto. Cada fila del gráfico izquierdo, correspondiente a una *feature*, tendrá 334 puntos, mientras que para el derecho cada una tendrá 2013 puntos.

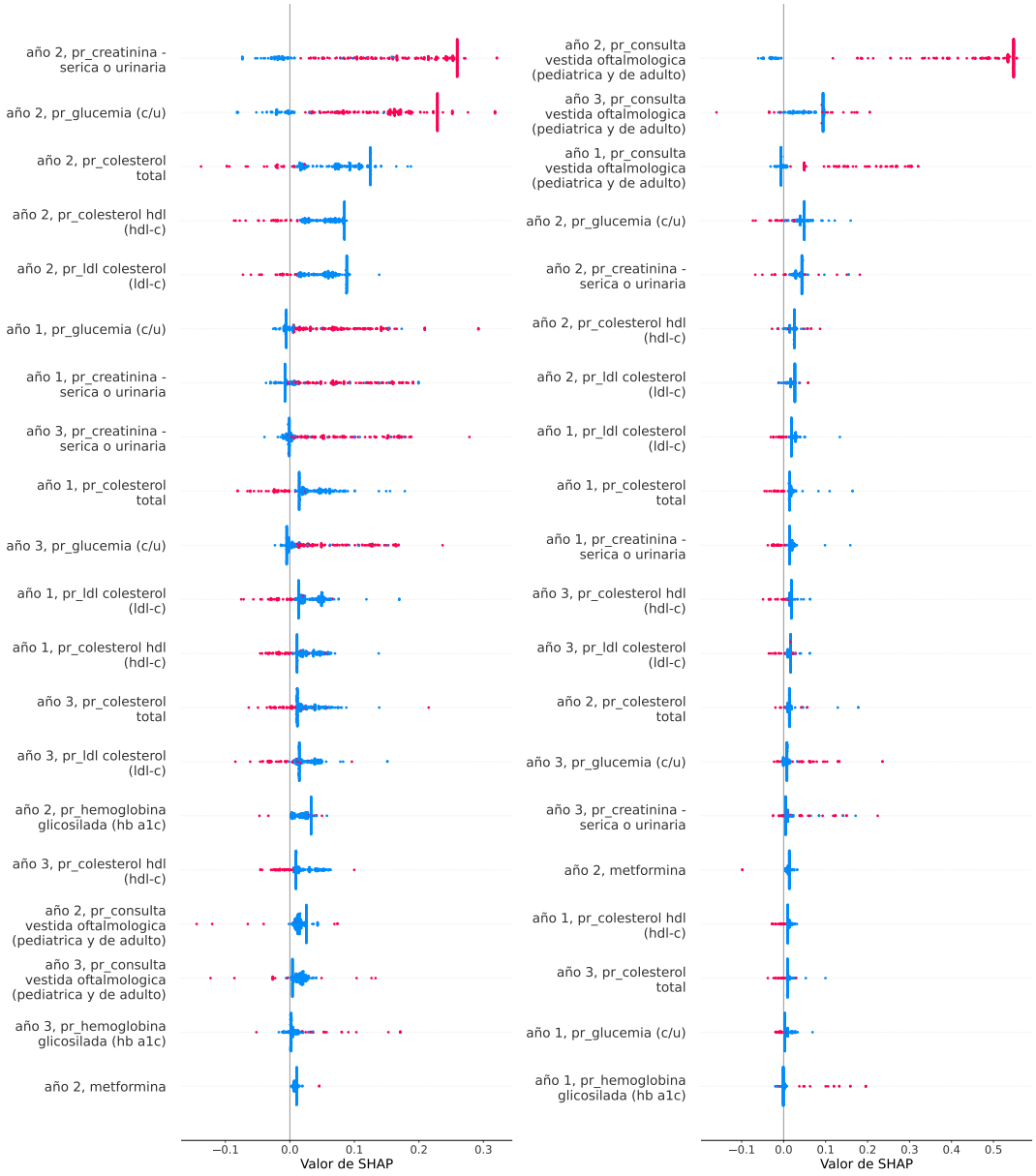


Figura 12: Valores de SHAP para el 10-*clustering*. El gráfico de la izquierda corresponde al quinto *cluster*, mientras que el de la derecha corresponde al sexto. Cada fila del gráfico izquierdo, correspondiente a una *feature*, tendrá 277 puntos, mientras que para el derecho cada una tendrá 262 puntos.

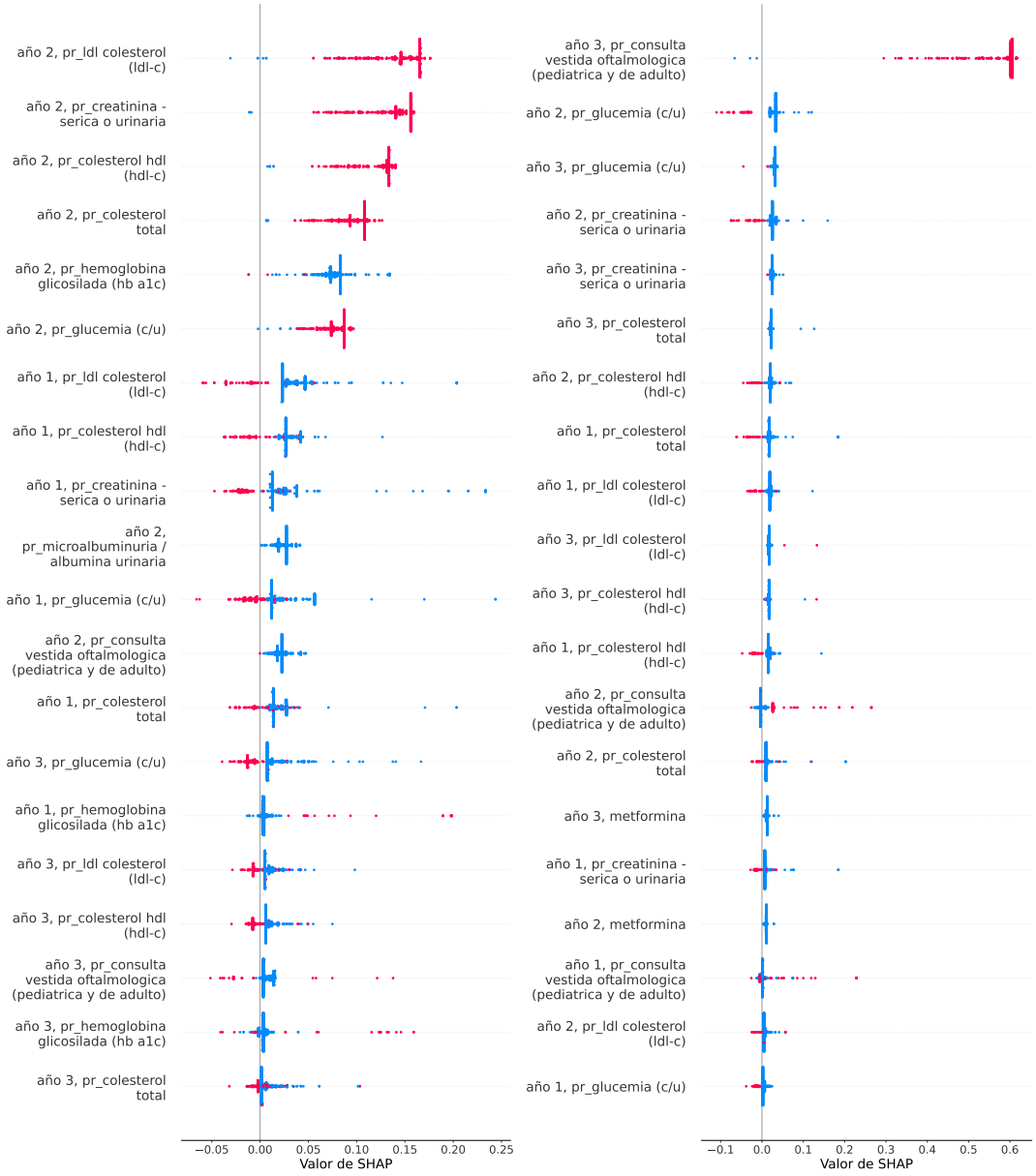


Figura 13: Valores de SHAP para el 10-*clustering*. El gráfico de la izquierda corresponde al septimo *cluster*, mientras que el de la derecha corresponde al octavo. Cada fila del gráfico izquierdo, correspondiente a una *feature*, tendrá 441 puntos, mientras que para el derecho cada una tendrá 438 puntos.

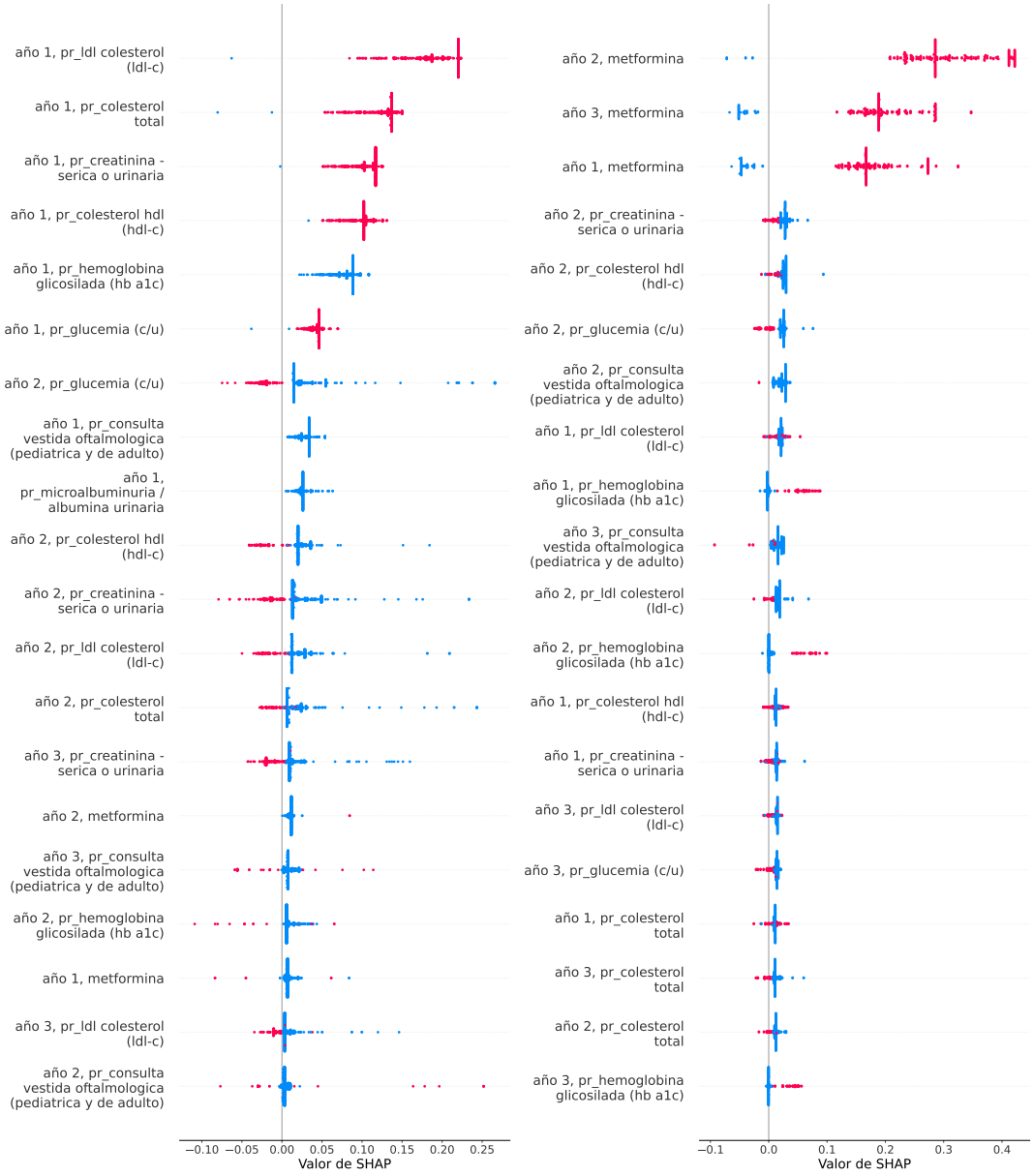


Figura 14: Valores de SHAP para el 10-*clustering*. El gráfico de la izquierda corresponde al noveno *cluster*, mientras que el de la derecha corresponde al décimo. Cada fila del gráfico izquierdo, correspondiente a una *feature*, tendrá 477 puntos, mientras que para el derecho cada una tendrá 235 puntos.

APÉNDICE C
PANTALLAS DEL PROTOTIPO DE SOFTWARE

Figura 15: Pantalla principal del prototipo de software. A la izquierda, se ven los módulos de datos de entrada e hyperparámetros. Estos deben ser configurados para una ejecución personalizada del *clustering*.

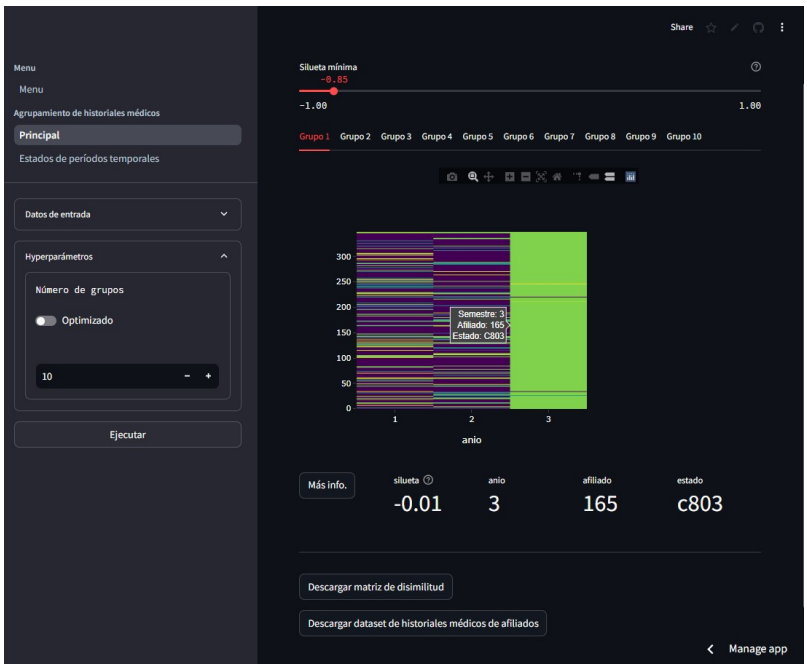


Figura 16: Pantalla principal del prototipo de software tras probar un caso de uso básico. Se observan un total de 10 solapas correspondientes a los *clusters*. El primer *cluster* se visualiza como un mapa de calor. El puntero marca uno de los estados para un afiliado y año de análisis, mostrando su información detallada.

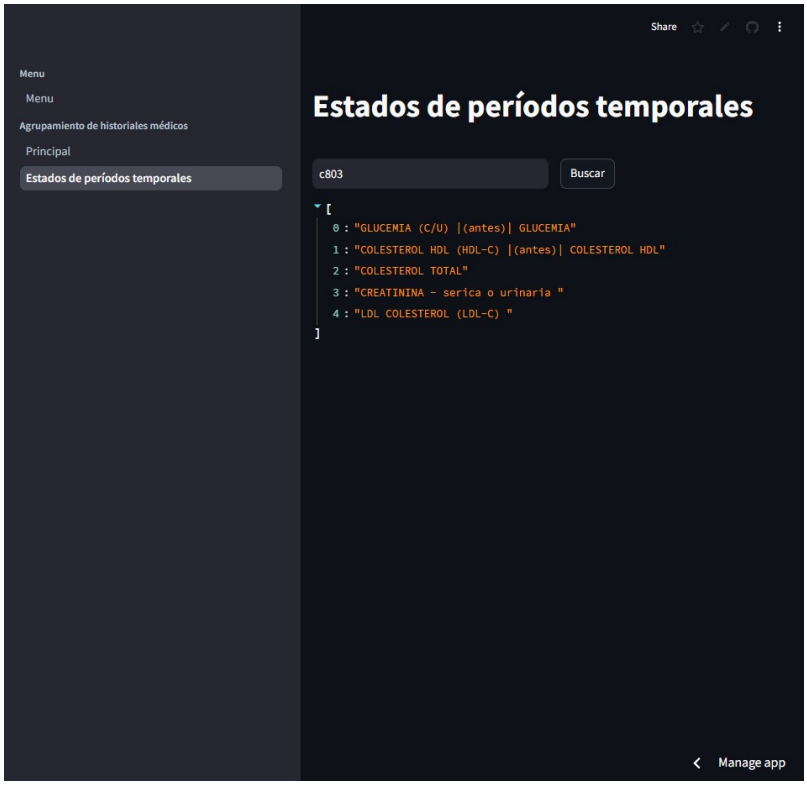


Figura 17: Pantalla desde la que se pueden ver los detalles de consumo de los estados.