UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
Departamento de Ingeniería informática

CARACTERIZACIÓN DE LA CLASIFICACIÓN DE LOS ESTABLECIMIENTOS PÚBLICOS DE SALUD EN CHILE SEGÚN COMPLEJIDAD MEDIANTE LA IDENTIFICACIÓN DE VARIABLES ASOCIADAS A CASUÍSTICA HOSPITALARIA

Angelo Jesús Carlier González

Profesor guía: Manuel Villalobos Cid

Tesis de grado presentada en conformidad a los requisitos para obtener el grado de Ingeniero de Ejecución en informática

Santiago – Chile
2019

RESUMEN

Los establecimientos de salud tienen como misión elevar el nivel de salud de la población entregando atención integral, oportuna, segura y de calidad, resguardando el uso eficiente de los recursos destinados a su funcionamiento. En la práctica, resulta difícil cuantificar el nivel óptimo de uso de los recursos dado los distintos escenarios que viven los establecimientos de salud. El Ministerio de Salud propuso una clasificación de los establecimientos de salud (alta, mediana y baja complejidad), basándose en distintas mediciones. No obstante, se desconoce como estas definiciones tienen relación con la casuística de los diferentes establecimientos. En este contexto, muchos datos asociados a la atención de los pacientes se encuentran disponibles en el Departamento de Estadísticas e Información de Salud, por lo que la casuística de los establecimientos de salud podría caracterizarse por medio de este conjunto de datos. Un problema no menor en el uso de estos, es la cantidad de variables y registros asociados al problema, en donde los algoritmos tradicionales de selección de características no son aplicables del todo. Se plantea una metaheurística basada en algoritmo genético como una buena estrategia para abordar el problema, en donde se buscará aquellas variables que caracterizan la clasificación definida por el MINSAL, con el fin de reducir la dimensión del problema e identificar aquellas variables fundamentales dentro de la clasificación. Los resultados demuestran que existe un conjunto de variables asociados a la casuística hospitalaria, las cuales se relacionan con la clasificación propuesta por el MINSAL. Sin embargo esta caracterización de la clasificación es muy acotada en relación al gran universo de variables relacionadas al problema y parte importante de ellas solo define perfiles específicos dentro de esta clasificación. Esta investigación permite identificar variables de importancia para establecer una categorización de establecimientos de salud o en caso contrario, sentar las bases para definir una nueva, que considere la casuística como parte importante a la hora de agrupar establecimientos de salud y realizar comparaciones justas entre ellos.

Palabras Claves: casuística hospitalaria, eficiencia técnica, algoritmo genético, metaheurística.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE ALGORITMOS

CAPÍTULO 1. INTRODUCCIÓN

1.1 ANTECEDENTES Y MOTIVACIÓN

Los establecimientos de salud tienen como misión elevar el nivel de salud de la población entregando atención integral, oportuna, segura y de calidad, resguardando el uso eficiente de los recursos destinados a su funcionamiento. Es de conocimiento popular en Chile que los presupuestos designados a las instituciones públicas son limitados, y el área de la salud no es una excepción. Si bien, el gasto público como porcentaje del Producto Interno Bruto (PIB) en el sector ha aumentado en los últimos años (2010 - 2017) alcanzando un 4,92% (OCDE, 2018) (Figura 1.1), se mantiene bajo en comparación a otros países pertenecientes a la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cuyo promedio en el año 2017 fue de 6,34% (OCDE, 2018).

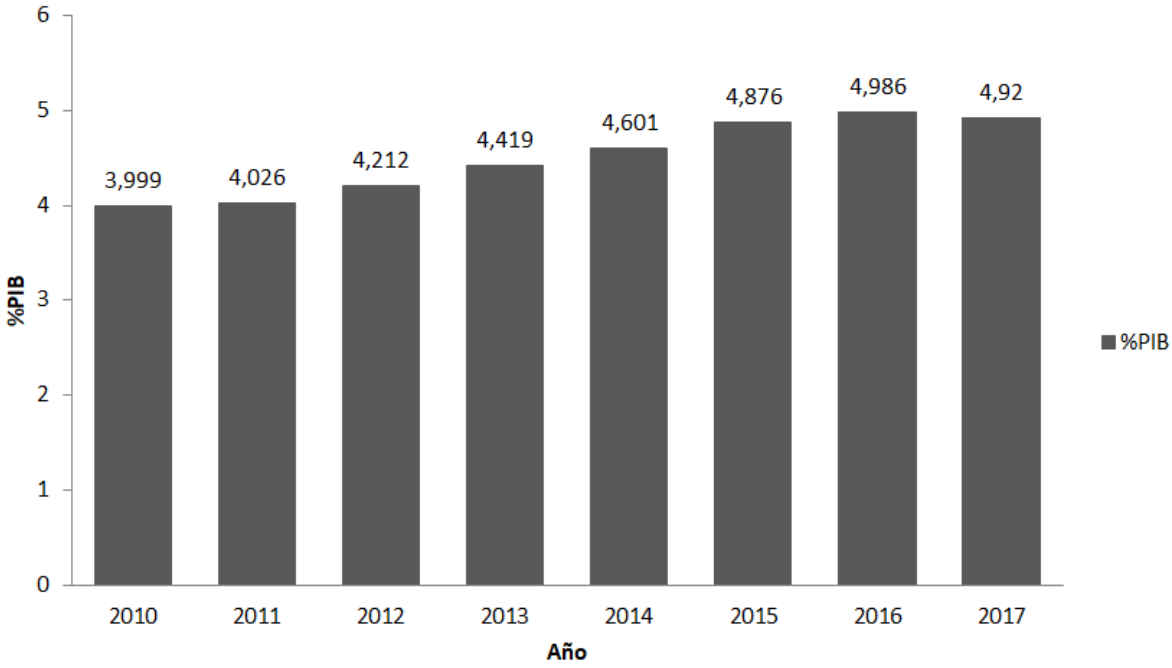

Figura 1.1: PIB destinado a salud en Chile 2010-2017.
Fuente: OCDE Elaboración propia, 2018.

En la teoría, un aumento en los recursos destinados a salud pública debiese mejorar los servicios prestados por las instituciones de salud, siempre y cuando sean usados eficientemente. A raíz de ello, es imprescindible que las autoridades y direcciones de los establecimientos cuenten con herramientas, indicadores y datos, que les permitan tomar decisiones que apunten a mejorar su gestión y eficiencia. En este aspecto se han realizado

diferentes estudios que plantean medir cuantitativamente la eficiencia de los establecimientos, específicamente empleando el concepto de eficiencia técnica. Esta se refiere a la habilidad de obtener el máximo producto posible dados una canasta de factores de producción y un nivel de tecnología determinados (Coll & Blasco, 2006).

Si se consideran los establecimientos hospitalarios como prestadores de servicios (Cortes-Martínez, 2010), sus productos pueden ser pacientes egresados, intervenciones quirúrgicas, días camas ocupados, indicadores de calidad, entre otros. En tanto, los factores de producción pueden ser uso de camas hospitalarias, personal, gasto financiero y uso de otros recursos (Kohl et al., 2018). En la práctica, resulta difícil cuantificar el nivel óptimo de uso de los recursos, por lo que las técnicas diseñadas para evaluar eficiencia técnica lo hacen de manera relativa, definiendo como establecimiento óptimo aquel que tenga el mayor valor de eficiencia técnica y a partir de éste, evaluar a los demás hospitales. Sin embargo, esto presenta un inconveniente, ya que los establecimientos poseen características diferentes que imposibilitan su directa comparación. Por ejemplo, no es recomendable contrastar un establecimiento como el Hospital Dr. Exequiel González Cortez, de complejidad alta con enfoque en la atención pediátrica, con el Hospital El Peral, cuyo foco es la atención psiquiátrica con internación de pacientes. Los productos hospitalarios son diversos con una casuística totalmente diferente, volviendo injusta su comparación.

El concepto de casuística hospitalaria hace referencia a los distintos tipos de pacientes que consultan un establecimiento de salud, y está definida por distintos factores como el diagnóstico, el pronóstico de los pacientes, la dificultad del tratamiento, el nivel de cuidado médico, los recursos utilizados, entre otros (Hornbrook, 1982).

Desde el año 2011, con el fin de determinar la casuística de los hospitales, se comenzó a implementar el Sistema de Grupos Relacionados por el Diagnóstico (GRD) en los hospitales públicos y privados de Chile. Esta herramienta usada a nivel mundial permite la categorización de pacientes en cuanto a sus diagnósticos y consumo de recursos empleando los datos de los pacientes egresados (Vega M., 2015), definiendo un perfil para cada hospital. Sin embargo, hasta la fecha, sólo 61 de los 190 hospitales públicos de Chile cuentan total o parcialmente con esta herramienta (Villalobos-Cid et al., 2016), limitando el uso de indicadores de eficiencia técnica (Santelices et al., 2013)(FONASA, 2015).

Para resolver esta dificultad, la literatura ha propuesto categorizar los hospitales en base a la clasificación de complejidad definido por el Ministerio de Salud (MINSAL), para luego evaluar su eficiencia técnica (Castro, 2007). El año 2013, el Ministerio de Salud propuso una clasificación de los establecimientos de salud (alta, mediana y baja complejidad), basándose en criterios de disponibilidad de recursos, complejidad y niveles de atención (MINSAL, 2013). No obstante, se desconoce como estas definiciones -establecidas administrativamente- tienen

relación con la casuística de los diferentes establecimientos.

En este contexto, muchos datos asociados a la atención de los pacientes se encuentran disponibles en el Departamento de Estadísticas e Información de Salud (DEIS), por lo que la casuística de los establecimientos de salud podría caracterizarse por medio de este conjunto de datos.

En este estudio se busca determinar si el uso de los conjuntos de datos dispuestos por el DEIS consiguen describir desde el punto de vista de la casuística de los establecimientos de salud, la clasificación propuesta por el Ministerio de Salud. En caso de no identificar aquellas variables pertenecientes a la casuística hospitalaria, se plantea como un punto a considerar en la reformulación de esta clasificación. Adicionalmente, se espera poder apoyar el desarrollo de futuras investigaciones relacionadas a medición de eficiencia técnica.

1.2 DESCRIPCIÓN DEL PROBLEMA

En la actualidad el MINSAL clasifica los establecimientos de salud en base a un conjunto de variables que combinan características asociadas a la casuística de los pacientes atendidos, tipo de infraestructura e indicadores de gestión. Esta clasificación es usada con fines administrativos, políticos, presupuestarios y de gestión. Sin embargo, no se ha estudiado en forma empírica si esta clasificación definida administrativamente se relaciona con la casuística hospitalaria. En este aspecto, la interrogante a resolver es: ¿Existen variables en los datos recolectados por el DEIS, referentes a egresos hospitalarios, estadísticos hospitalarios y categorización que definen los aspectos de la casuística hospitalaria que permitan explicar la clasificación definida por el MINSAL para la complejidad de los establecimientos de salud?

1.3 SOLUCIÓN PROPUESTA

La solución propuesta busca identificar las variables que fundamentan la clasificación de los establecimientos de salud, en base a la propuesta del MINSAL. Para ello se utilizarán técnicas de minería de datos junto a una metodología de generación de conocimiento. En base a este planteamiento, la solución propuesta se divide en dos partes.

- Previas al análisis de datos
  - Orden de datos: se efectuará una limpieza y almacenamiento de los datos desarrollando un modelo ad-hoc, esto considera un modelo de datos para almacenar de forma

correcta lo datos en un motor de base de datos, que además permita una correcta extracción posterior, que alimente las técnicas de minería de datos. Es importante considerar que la información propuesta por el MINSAL no se encuentran alojada en base de datos ni dispuesta a usuarios mediante alguna API.

  - Limpieza: una vez construída la base de datos, se procederá a los procesos de limpieza preliminares sobre los datos, que buscan generar la estructura de datos inicial para los procesos de minería de datos.

- Durante y posterior al análisis de datos

  - Aplicación de minería de datos: implementación de un algoritmo, inicialmente definido como algoritmo genético, que permita abordar el problema de clasificación. Para ello se debe proponer la estructura básica del algoritmo, sus operadores genéticos, estrategias de parametrización e identificación de métricas evaluación.
  - Resultados: evaluación de los datos obtenidos. Es importante destacar que los análisis sobre datos son procedimientos recursivos, que en base a las conclusiones obtenidas deben realizarse correcciones y en ocasiones volver a aplicar. Adicionalmente, es necesario comparar los resultados obtenidos con otras técnicas de análisis de datos. Esta comparación debe realizarse con el fin de evaluar rendimiento desde el punto de vista computacional y de calidad de la solución hallada.

El propósito de la solución es implementar una estrategia basada en algoritmo genético, sobre los datos de producción hospitalaria dispuestos por el MINSAL, que permita identificar las variables que año a año, fundamentan la categorización que plantea el MINSAL (MINSAL, 2013), en base a la complejidad de sus establecimientos de salud. El identificar estas variables permitiría analizar indicadores de gestión orientados a las variables de mayor importancia, estableciendo parámetros de eficiencia más certeros y con foco en aquellas variables. Adicionalmente, considerando que la casuística hospitalaria es un factor fundamental en este análisis, permitiría atender las necesidades de cada establecimiento de forma personalizada en cada uno de ellos, considerando con mayor importancia aquellas variables pertenecientes a cada grupo.

1.4 OBJETIVOS Y ALCANCE DEL PROYECTO

1.4.1 Objetivo general

Proponer una estrategia para identificar aquellas variables de la casuística hospitalaria que permitan caracterizar la clasificación hospitalaria definida por el MINSAL (Ministerio de Salud) para los hospitales públicos de Chile, considerando los datos disponibles de casuísticas disponibles por el DEIS (Departamento de Estadísticas e Información de Salud).

1.4.2 Objetivos específicos

1. **OE1**: establecer el modelo de datos para el conjunto de datos de producción hospitalaria actuales disponibles en el DEIS.
2. **OE2**: evaluar la calidad de los datos disponibles según su completitud, conformidad y consistencia. Depurar en casos que corresponda.
3. **OE3**: diseñar y construir un algoritmo basado en metaheurística para identificar las variables que fundamentan la clasificación establecida por el MINSAL, que permita evaluar combinaciones con el número de variables disponibles en DEIS.
4. **OE4**: comparar resultados finales obtenidos con otras estrategias de la literatura, evaluando el parentesco de los agrupamientos obtenidos con el propuesto por el MINSAL.
5. **OE5**: evaluar y analizar la relación entre la clasificación de establecimientos en base a su complejidad definida por el MINSAL, extrayendo conclusiones desde el punto de vista computacional.

1.4.3 Alcances

Para efectuar un análisis dinámico, se considerará datos correspondientes a los años 2014, 2015, 2016, 2017, y 2018. Estos análisis se realizarán en forma independiente año a año, debido a que algunas variables no son comparables en el periodo completo de análisis. Adicionalmente, se realizará un análisis comparativo de los agrupamientos finales, de modo de identificar que elemento de la casuística hospitalaria se repiten año a año.

1.5 METODOLOGÍA Y HERRAMIENTAS UTILIZADAS

1.5.1 Metodología

La metodología a utilizar será KDD (por sus siglas en inglés, Knowledge Discovery and Data Mining). Esta es ampliamente utilizada en el análisis de información y se adecua al ciclo de vida de un proyecto.

La metodología KDD hace referencia a todo el proceso de descubrimiento de conocimiento, a diferencia del proceso de minería de datos, que en este caso solo forma parte de una etapa dentro de la aplicación de esta metodología (Fayyad et al., 1996).

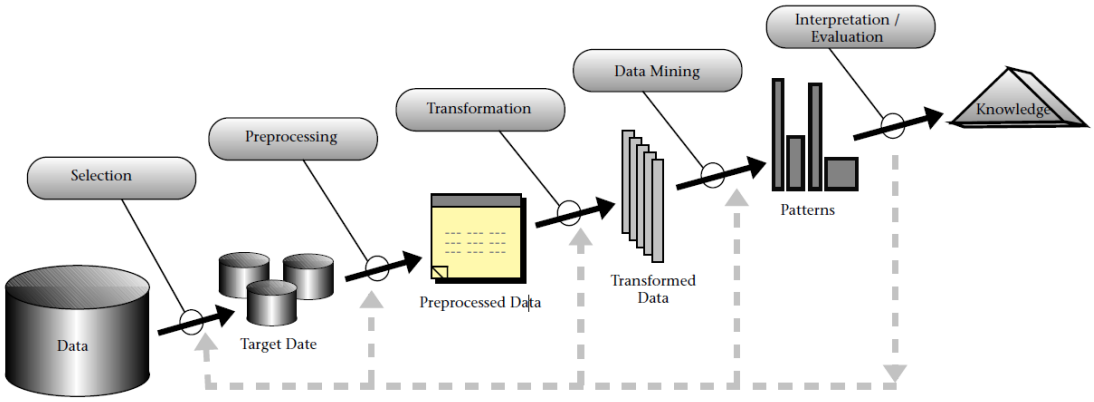

Figura 1.2: Overview metodología KDD.
Fuente: (Fayyad et al., 1996)

Las fases de la metodología a utilizar se definen de la siguiente forma (Figura 1.2):

- Selección (OE1): etapa que consiste en la selección de los datos a utilizar en base a la gran cantidad de datos disponibles para el estudio y en este contexto se debe seleccionar los datos a utilizar, dentro de todos los que se encuentran disponibles en la plataforma del DEIS.

- Preprocesamiento (OE2): etapa que consiste en la generación de los datos preliminares de estudio. Su generación es fundamental para las transformaciones iniciales que soportan los elementos de entrada a las técnicas de minería de datos. Particularmente en este estudio, se debe generar un preprocesamiento de información, en donde se eliminan variables que se encuentren mal registradas y se acota el universo de variables a considerar a futuro.

- Transformación (OE2): etapa que consiste en la transformación de los datos en base a lo necesario para la aplicación de las técnicas de minería de datos. Esta transformación de datos puede ser en la generación de nuevas variables (tomando como entrada las ya existentes) o una nueva distribución (cambios de estructura). En este aspecto, se hace fundamental la etapa de transformación en el estudio, ya que es la encargada de generar la estructura de datos inicial a utilizar por las técnicas de minería de datos.

- Minería de datos (OE3, OE4): etapa que consiste en la aplicación de las técnicas de minería de datos que, para este trabajo en particulará, se enfoca principalmente en la aplicación de computación evolutiva. Como no existe en la literatura una estrategia ad-hoc propuesta para resolver este problema, se propone el diseño de una estrategia propia. En este contexto, se aplicará una metaheurística basada en algoritmo genético utilizando sus fases principales (Yang & Koziel, 2011), en donde se generará una población inicial (cada individuo corresponde a una solución) que irá mejorando conforme a las generaciones (iteraciones). Cada generación busca la mejora de las soluciones, aplicando operaciones genéticas:

  - *Crossover*: etapa en donde se seleccionan dos individuos y se genera un nuevo individuo combinando información de las dos iniciales.
  - Mutación: para cada nuevo individuo generado en la etapa de entrecruzamiento, se modifica de forma aleatoria una o más características, obteniendo un nuevo individuo, con características modificadas.
  - Selección: etapa que consiste en la evaluación, según una métrica definida, de los individuos iniciales y finales de cada generación para finalmente seleccionar a los mejores.

  Para la construcción de este algoritmo genético, se consideran los siguientes pasos:

  - Caracterización de un individuo: se debe definir un individuo como representación de una solución, estableciendo sus características asociadas.
  - Inicialización: consiste en la generación de la población inicial en base a distintos parámetros de entrada, considerando de esta forma distintas estrategias en su elaboración.
  - Operaciones genéticas: consiste en el diseño de las operaciones genéticas, considerando distintos parámetros de entrada que establezcan la configuración y estrategia de aplicación de cada una de ellas.
  - Implementar función fitness: consiste en la aplicación de la función fitness a los individuos resultantes de cada iteración, en donde se seleccionará los individuos que obtengan un mejor resultado en esta función, para ser considerados en el siguiente

ciclo. La función fitness aplicada, corresponde a una medida de similaridad entre los *clusters* identificados, en comparación a los agrupamientos establecidos por el MINSAL. Estas medidas de similaridad serán cuantificadas por 3 distintos indicadores (índice Jaccard, índice RAND e índice Fowlkes-Mallows. Ver sección 3.5).

- Interpretación / Evaluación (OE5): etapa que consiste en la evaluación de los resultados obtenidos en la etapa anterior. Es importante destacar que la mayoría de los estudios en este plano, son de carácter recursivo, y requieren de constante evaluación y aplicación de técnicas de minería de datos, realizando correcciones entre cada una de ellas. La calidad de cada resultado obtenido, será en términos de semejanza al agrupamiento definido por el MINSAL, y será medida por el índices de similitud (índice Jaccard, índice Rand, entre otros). Particularmente en el caso de este estudio, se pretende generar correcciones al algoritmo genético, que serán implementadas y nuevamente evaluadas. Adicionalmente, se generará una comparación con otras estrategias identificadas en la literatura (Random forest, KNN, entre otros).
- Generación de conocimiento: escritura de tesis.

1.5.2 Herramientas de desarrollo

Las herramientas de desarrollo son las siguientes:

- **Python**: Lenguaje de programación orientado a objetos. Su selección se hace de gran importancia a la hora de realizar las 3 primeras etapas de la metodología KDD. Existen dos variables que permiten seleccionar este lenguaje de programación en desmedro de otros para las primeras etapas:
  - Conocimiento: existe conocimiento sobre este lenguaje de programación por sobre los demás.
  - Módulos diseñados: existen diversos módulos ya diseñados sobre este lenguaje para el tratamiento de archivos y distribución de los datos sobre motores de base de datos.
- **R**: lenguaje de programación con enfoque estadístico. Proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupación), y es altamente extensible. Una de las fortalezas de R es la facilidad con que pueden implementarse técnicas estadísticas y análisis de datos en general. Dentro de R existen diversos módulos de interés, que simplifican la codificación de una aplicación (foundation, 2019). Su selección se hace de mayor importancia en la

etapa de minería de datos. Existen diversos módulos diseñados que permiten una correcta implementación de algoritmos genéticos.

1.6 ORGANIZACIÓN DEL DOCUMENTO

En el Capítulo 2 se revisan los conceptos generales de técnicas de agrupamientos, tipos de métricas de distancia y métricas de calidad entre agrupamientos. La metodología utilizada en la investigación es descrito en el Capítulo 3, en donde se detalla cada etapa. En particular, la sección 3.4 se detalla la implementación de la metaheurística basada en algoritmo genético para abordar el problema. Los resultados se exponen en la sección 3.5. Finalmente se presentan las conclusiones obtenidas mediante la investigación en el capítulo 4, incorporando además trabajos futuros.

CAPÍTULO 2. ANTECEDENTES

El objetivo de este capitulo es establecer un marco referencial de estudios realizados en el ámbito de la eficiencia hospitalaria en Chile. En el podrá conocer algunos planteamientos realizados por distintos estudios. Adicionalmente se establecen las bases teóricas que ayudan a comprender el estudio realizado.

2.1 MARCO TEÓRICO

2.1.1 Métricas de distancia

Distintas métricas de distancias son utilizadas en el contexto de métodos de agrupamientos para la comparación entre elementos, su uso fundamental es cuantificar la distancia o similitud que existe entre los elementos agrupados. En esta sección es posible ver las más utilizadas en la literatura y el fundamento de las seleccionadas para la investigación. A continuación se presentan tres métricas de interés: Manhattan, Euclideana y Pearson.

Distancia Manhattan

La distancia Manhattan, o también conocida como distancia rectilínea, es aquella que mide la distancia de la ruta inspirada en el diseño de una cuadrícula (Krig, 2016). Se define de la siguiente forma:

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2.1}$$

Donde X e Y son vectores, x_i e y_i representan al elemento i de los vectores X, Y respectivamente.

Distancia Euclideana

La distancia Euclideana es la distancia más utilizada, considera el concepto de distancia entre dos puntos en un plano y está definida de la siguiente forma (Krig, 2016):

$$d(X,Y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2} \tag{2.2}$$

Donde X e Y son vectores, x_i e y_i representan al elemento i de los vectores X, Y respectivamente.

Distancia basada en correlación de Pearson

Esta distancia basa su cálculo en el índice de correlación de Pearson entre dos vectores, se define de la siguiente forma (Krig, 2016):

$$d(X,Y) = 1 - \frac{\sum_{i=1}^{n}(x_i + \overline{x})(y_i + \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i + \overline{x})^2 \sum_{i=1}^{n}(y_i + \overline{y})^2}} \tag{2.3}$$

Donde X e Y son vectores, x_i e y_i representan al elemento i de los vectores X, Y respectivamente y $\overline{x}$, $\overline{y}$ sus medias.

Para la investigación se usarán dos medidas de distancia: (1) la primera de ella permite diferencia los establecimiento según la magnitud de las variables, **distancia Euclidiana**, y (2) la segunda comparar su comportamiento usando la **correlación de Pearson**.

2.1.2 Selección de características

Los métodos de selección de características son utilizados en conjuntos de datos de alta dimensionalidad. Los conjuntos de datos cuya dimensionalidad es alta presentan dificultades al momento de extraer conocimiento de ellos, en donde la consideración de demasiadas características genera factores de ruido en soluciones encontradas, los tiempos de procesamiento de los datos es muy elevado y/o la asignación de recursos para el procesamiento de ellas es innecesario. Estos métodos se emplean para obtener un subconjunto de características , dicho subconjunto será aquel que represente al conjunto inicial, sin perder información relevante al considerar el total de elementos. El problema de selección de características consiste

básicamente en encontrar un subconjunto de f características de un conjunto inicial más numeroso con F características, $f < F$. Para ello, se deberá utilizar una función criterio que discrimine entre características y seleccione mejores subconjuntos de características y un algoritmo de búsqueda que permita encontrarlo. En este aspecto, a cntinuación se describen dos algoritmos: *random forest* y *k-nearest neighbor*

Random forest

Para entender el funcionamiento de *random forest*, es necesario comprender el comportamiento de un árbol de decisión. Un árbol de decisión es la representación de posibles soluciones utilizando condiciones dadas por características identificadas en los datos. Usualmente las personas utilizan los árboles de decisión en decisiones de carácter diario tal como el ejemplo (Figura 2.1)

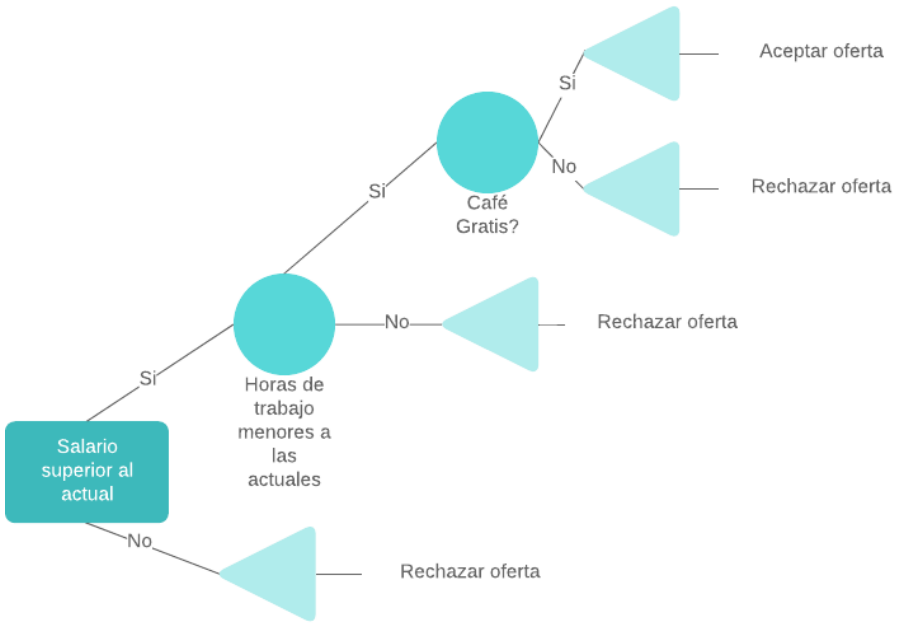

Figura 2.1: Ejemplo de árbol de decisión, en la consideración de oferta laboral
Fuente: Elaboración propia, 2019

Dentro de la generación de un árbol de decisión, existen consideraciones como el orden en que las variables son consideradas dentro del grafo, como la cantidad de variables utilizadas, *Random forest* está basado en la implementación de múltiples arboles de decisión (Breiman, 2001). Utilizando de manera aleatoria un subconjunto de variables en cada uno de los

árboles generados. En él se considera el concepto de bagging (Bootstrap Aggregation) que hace referencia al muestreo aleatorio de las soluciones, que ayuda a identificar mejores soluciones dentro del conjunto total de soluciones. En conjuntos de soluciones con baja varianza, este método no es tan eficiente, aún asi permite disminuir el sesgo que existe al considerar una (o pocas) solución. Como un clasificador, *Random forest* opera como una estrategia de selección de características de forma implícita, utilizando un subconjunto de variables denominadas de mayor impacto en el conjunto global de variables. El indicador que entrega este valor relativo a la importancia de cada característica está asociado a la impureza de Gini, que mide con qué frecuencia un registro elegido aleatoriamente del conjunto de datos utilizado para entrenar el modelo se etiquetará incorrectamente si se etiquetó aleatoriamente. La importancia de Gini se puede aprovechar para calcular la disminución media en Gini, que es una medida de importancia variable para estimar una variable objetivo.

K-nearest neighbor

La utilización de un clasificador como k-nearest neighbor (KNN) para la selección de características tiene su motivación en el uso de *random forest*, que considera múltiples aplicaciones de árboles de decisión. En este aspecto y dada la implementación de KNN, se debe aplicar una cantidad n de veces el algoritmo sobre distintos subconjunto de variables del conjunto total (Figura 2.2). Para seleccionar un subconjunto de variables que tienen la capacidad de clasificar al conjunto inicial, se evalúa el rendimiento de cada variable y se establece la precisión promedio como el indicador de importancia dentro de la clasificación (Li et al., 2011)

Es importante señalar que el uso de estas técnicas de selección de características tradicionales no son efectivas en conjuntos de datos que poseen una alta dimensionalidad, como es el caso de esta investigación, debido a la gran cantidad de combinaciones distintas en el espacio de soluciones posibles, sin embargo se utilizarán como parámetros de referencia a la hora de evaluar los resultados obtenidos por la implementación de una heurística basada en algoritmo genético.

2.1.3 Técnicas de agrupamientos

Distintas técnicas de agrupamientos o *clustering* son utilizadas para encontrar grupos o *clusters* entre los elementos de un conjunto. Los agrupamientos identificados con estas técnicas,

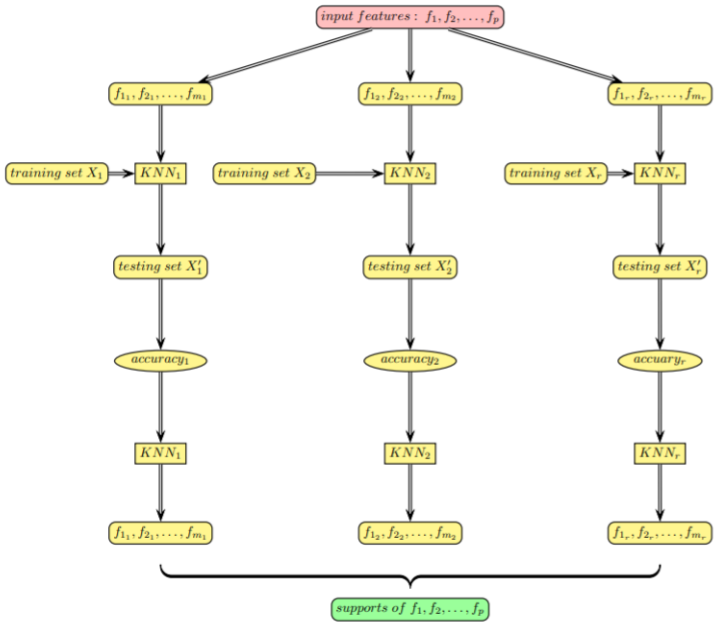

Figura 2.2: Implementación de KNN para selección de características
Fuente: (Li et al., 2011)

permiten establecer una relación de parentesco entre aquellos elementos que pertenecen a un mismo grupo y disimilitud entre aquellos que pertenecen a grupos distintos. Los grupos generados son útiles en la toma de decisiones relacionadas a la clasificación y predicción de nuevos elementos, esto dado que la cantidad de representantes es menor a la cantidad total de elementos de un conjunto, lo que reduce la cantidad de datos a utilizar (Murty & Susheela Devi, 2011).

Se definen cuatro técnicas de agrupamientos utilizadas, tres de ellas correspondientes a técnicas del tipo *Partitioning Clustering*, que necesitan de un valor previo de agrupamientos a generar y sus elementos pertenecen exclusivamente a un grupo, y otra del tipo *fuzzy*, en donde los elementos podrían eventualmente pertenecer a más de un grupo.

Partitioning Around Medoids (PAM)

PAM es un algoritmo de agrupamientos desarrollado por Kaufman y Rousseeuw (Kaufman & Rousseeuw, 1990) cuyo objetivo es determinar k agrupamientos, con k definido previamente. PAM determina un objeto como representante para cada clase, estos objetos representantes, llamados medoides, deben ser los objetos centrales de cada agrupamiento. Una vez que los medoides son seleccionados, cada objeto que no fue seleccionado como medoide,

es agrupado con su medoide más cercano. En un proceso iterativo, la selección de los medoides, que inicialmente es arbitraria, es corregida utilizando un algoritmo para ello. El algoritmo PAM se describe de la siguiente forma (Ng & Han, 1994):

1. Seleccionar k objetos representativos de forma arbitraria.
2. Calcular la matriz de distancia DM_{ih} en base a una métrica de distancia, para todos los pares de objetos O_i, O_h donde O_i es un objeto seleccionado, y O_h no.
3. Asignar a cada O_h, el O_i más cercano.
4. Para cada objeto del grupo definido por O_i, se debe realizar el cálculo de distancia promedio. Si existe un objeto O_h que posee una distancia promedio menor a la de O_i, este objeto gana la condición de representante del grupo y es necesario volver al paso 2.
5. En caso, que ningún O_i sea sustituido o la cantidad máximas de iteraciones definidas, finaliza la ejecución del algoritmo.

K-Means

K-Means es el algoritmo más popular de la categoría *Partitioning Clustering*, diseñado por J.B.Macqueen (Murty & Susheela Devi, 2011). Al igual que el algoritmo PAM 2.1.3, se debe definir previamente la cantidad de agrupamientos a generar. El objetivo del algoritmo es determinar los centroides del grupo, este elemento se define como el elemento central de cada grupo, no necesariamente pertenece al conjunto inicial de elementos, a diferencia de los medoides en el algoritmo PAM. La descripción de su algoritmo es la siguiente:

1. Seleccionar k elementos de los n elementos que posee el conjunto, estos elementos serán los centroides iniciales.
2. Asignar cada elementos no seleccionado inicialmente a uno de los k agrupamientos. Cada elemento debe ser asignado a su centroide más cercano.
3. Calcular el centro del agrupamiento generado, seleccionar este nuevo elemento (que puede pertenecer o no al conjunto n inicial) como centroide del grupo.
4. Asignar cada elemento del conjunto n a su centroide más cercano. En caso de existir cambios, volver al paso 3.
5. En caso de no existir cambio en la última iteración de reasignación de los centroides o cumplir una cantidad de iteraciones determinadas, finalizar.

Clustering LAR Applications (CLARA)

Kaufman & Rousseeuw (1990) crearon un método de muestreo simple, utilizando PAM. CLARA genera múltiples muestras del conjunto inicial, sobre ellas se aplica PAM y de todos los agrupamientos creados, selecciona el mejor. En la aplicación del algoritmo, es necesario determinar la calidad del agrupamiento obtenido, esto se realiza mediante la suma total $intra - clusters$, en otras palabras, la suma total de la distancia de cada elemento del grupo con su medoide. En este sentido el algoritmo se describe de la siguiente forma Ng & Han (1994):

1. Para $i = 1$ hasta s, donde s es el número de iteraciones definidas previamente, repetir los siguientes pasos:
2. Generar una muestra de elementos de forma aleatoria del conjunto inicial, esta muestra siempre es del mismo tamaño, en todas las iteraciones. Sobre ellos aplicar el algoritmo PAM y encontrar k medoides de la muestra.
3. Para cada elemento O_i del conjunto inicial, se debe determinar el medoide O_k más similar según alguna métrica de distancia.
4. Calcular el promedio de disimilitud del agrupamiento obtenido en paso anterior. Si este valor es menor al mínimo actual, se debe usar este valor como mínimo actual y mantener los k medoides obtenidos como los mejores medoides, de forma de compararlos con los obtenidos en futuras iteraciones.
5. Volver al paso 1 para una siguiente iteración.

Fuzzy c-means

Este método propuesto por (Bezdek et al., 1984), posee bases similares a las utilizadas en *k-means* en donde adicionalmente se asocia un valor de pertenencia de un elemento a cada grupo.

El algoritmo corresponde a un proceso iterativo que busca minimizar la función de costo definida de la siguiente forma:

$$F = \sum_{j=1}^{N} \sum_{i=1}^{c} u_{ij} ||x_j - v_i||^2 \tag{2.4}$$

Donde N representa a la cantidad de elementos, c representa la cantidad de agrupamientos deseados, v_i en el centro del clúster i

2.1.4 Métricas de calidad

Las métricas de calidad permiten evaluar el nivel de similitud entre agrupamientos. Todas las métricas descritas en este documento presentan conceptos en común definidos de la siguiente forma:

Sean V_1 y V_2 dos grupos de n elementos con el mismo número de clases k, se define:

- yy: par de elementos que pertenecen a la misma clase en V_1 y a la misma clase en V_2.
- ny: par de elementos que pertenecen a distintas clases en V_1 pero a la misma clase en V_2.
- yn: par de elementos que pertenecen a la misma clase en V_1 pero a diferentes clases en V_2.
- nn: par de objetos que pertenecen a distintas clases en V_1 y a distintas clases en V_2.

Considerando esta definición, se describe el índice de Jaccard, índice de Rand ajustado e índice Fowlkes-Mallows.

Índice de Jaccard

El índice de Jaccard es una medida de similitud entre agrupamientos que describe la presencia y ausencia de los objetos en pares. Cuenta el número de objetos en pares que pertenecen a la misma clase en ambos grupos y la divide por el número de objetos en pares que pertenecen a la misma clase en al menos uno de ellos (Thomas W. Yee , auth.). Se define de la siguiente forma:

$$Jaccard = \frac{yy}{yy + ny + yn} \tag{2.5}$$

El índice de Jaccard está en el rango de [0,1]. Cuando este es 0, no hay dos objetos juntos en la clase V_1 y la clase V_2. Si los grupos V_1 y V_2 son iguales, ny = yn = 0 por lo que alcanza su valor máximo (Brouwer, 2009).

Índice Rand

Este medida es comúnmente utilizada para comparar dos grupos y se describe en detalle en (Hubert & Arabie, 1985). El rango de valores que alcanza es [0,1], en donde 0 indica que ambos grupos no comparten pares de objetos en las mismas clases, y 1 indica que los dos grupos son exactamente iguales. Se define de la siguiente manera:

$$RI = \frac{(yy + nn)}{(yy + ny + ny + nn)} \tag{2.6}$$

Índice Rand ajustado

Una versión mejorada del índice Rand se propone en (Hubert & Arabie, 1985), y se define de la siguiente forma:

$$ARI = \frac{2(yy \times nn - ny \times yn)}{(yy + ny)(ny + nn) + (yy + yn)(yn + nn)} \tag{2.7}$$

Índice Fowlkes-Mallows

Utilizando la misma definición en común con los demás indicadores, el índice Fowlkes Mallows se define de la siguiente forma (Fowlkes & Mallows, 1983):

$$FM = \frac{yy}{\sqrt{(yy + ny)(yy + yn)}} \tag{2.8}$$

El índice Fowlkes Mallows se encuentra en el rango [0,1]. Toma su valor mínimo (0) cuando no existen dos pares de objetos en los grupos V_1 y V_2 que pertenecen a la misma clase simultáneamente o el valor 1 cuando ambos grupos son exactamente iguales.

2.1.5 Metodología knowledge discovery in databases (KDD)

La metodología knowledge discovery in databases, corresponde a una metodología para la búsqueda de conocimiento no trivial en bases de datos, entendiendo como base de datos un conjunto de hechos, y un patrón una expresión que define parte de estos hechos o bien define un modelo aplicable en ellos de igual forma (Fayyad et al., 1996).

Esta metodología posee sus fases definidas, y si bien es parte de un proceso secuencial, también forma parte de un proceso iterativo continuo que mejora la calidad de los patrones identificados y el conocimientos extraído de los datos.

Las fases que definen a KDD son las siguientes:

- Selección: entendiendo el dominio de aplicación y el objetivo definido en el uso de KDD, esta etapa selecciona los datos considerados para la aplicación de la metodología, en donde dependiendo del objetivo de la misma, puede considerar algunos subconjuntos de hechos y no la base de datos completa.

- Preprocesamiento: una vez que el conjunto de datos o hechos es seleccionado, se aplica un preprocesamiento en los datos, que busca la limpieza de los mismos y algún procesamiento menor, que puede incluir estrategias para abordar datos nulos, datos faltantes, etc.

- Transformación: distintos indicadores de interés pueden ser requeridos por etapas siguientes, para ello la etapa de transformación se utiliza en la creación de nuevos valores, a través de los hechos ya existentes o de igual forma para cambios en la estructura de los datos.

- *Data mining*: en primera etapa los análisis del tipo exploratorio ayudan la definición de los métodos a utilizar, parametrización de los mismos o incluso la definición de ciertas hipótesis. Adicionalmente esta etapa considera la búsqueda de patrones mediante la aplicación de técnicas de minería de datos, en ellas podemos encontrar las técnicas de agrupamientos, arboles de decisión o regresión.

- Interpretación y evaluación: La interpretación de los datos es un paso fundamental dentro de la metodología, y en ocasiones requiere la iteración sobre alguna etapa previa, adicionalmente la visualización de información permite la identificación de patrones y ayuda en su evaluación.

2.2 ESTADO DEL ARTE

Distintas investigaciones se han realizado en el ámbito de gestión hospitalaria, evaluando y comparando establecimientos de salud públicos en base a distintos indicadores (Castro, 2007) (Barahona-Urbina, 2011) (Santelices et al., 2013) (Villalobos-Cid et al., 2016). Uno de ellos es la medición de la eficiencia técnica en los establecimientos de salud, donde todos los estudios han coincidido en el uso de agrupamientos previos que permiten comparar instituciones equivalentes.

Castro (2007) propone la medición de eficiencia técnica en los hospitales públicos en Chile utilizando la clasificación propuesta por el MINSAL, y en base a ella, efectúa la evaluación de eficiencia técnica. Dentro de los supuestos planteados para su investigación, menciona que el uso de los GRD es fundamental para obtener la casuística hospitalaria, sin embargo, no hizo uso de ellos porque aún no estaban implementados en Chile. Asume que esta casuística se encuentra implícita en la clasificación de la época dada por el MINSAL.

Esta agrupación propuesta por el MINSAL posee las siguientes subdivisiones: (1) **Baja complejidad**, que equivale a hospitales ubicados en pequeñas ciudades o área rurales, (2) **Alta complejidad adultos**, que involucra hospitales y centro de referencia ubicado en ciudades principales, (3) **Mediana complejidad** se asocia a aquellos que no están ubicados en ciudades grandes (MINSAL, 2013). De ella también se pueden identificar (4) **Hospitales pediátricos**, que contiene hospitales de alta complejidad y especializados en área de pediatría y (5) **Hospitales psiquiátricos** cuyo foco es la salud mental (Villalobos-Cid et al., 2016).

Barahona-Urbina (2011) realizó un Análisis envolvente de datos (AED) para analizar eficiencia hospialaria, usando datos relacionados a médicos, enfermeras, matronas, camas disponibles y egresos hospitalarios. Esta última es considerada como variable de salida, y como tal, se asume que el objetivo que cada establecimiento de salud es producir egresos hospitalarios. El origen de datos utilizado corresponde al DEIS, sin embargo, las agrupaciones generadas en contraste con la investigación de Castro (2007), fue a nivel geográfico por región en donde se ubica cada establecimiento de salud.

En el mismo año, se plantea la idea de medir la eficiencia técnica hospitalaria considerando los GRD (específicamente la versión IR-GRD) (Santelices et al., 2013). Solo se estudió un total de 28 hospitales en esta investigación, ya que eran los establecimientos que poseían a la fecha esta herramienta implementada.

Finalmente, existen investigadores que plantean la opción de categorizar a los establecimientos de salud en base a su casuística (Villalobos-Cid et al., 2016), en donde se utilizan técnicas de minería de datos, específicamente algoritmos de clasificación (MST-kNN, k-means e

isodata), para categorizar previamente los establecimientos de salud y establecer un parámetro de comparación ideal. Es acá en donde los GRD toman gran importancia, sin embargo, tal como se ha mencionado anteriormente, no se encuentran implementados en el 100% de los establecimientos de salud, teniendo como cobertura en la actualidad 61 hospitales (FONASA, 2015) (Villalobos-Cid et al., 2016).

Considerando las distintas investigaciones realizadas referentes a la medición de eficiencia técnica hospitalaria, se hace fundamental la definición de una clasificación de establecimientos que considere criterios de disponibilidad de recursos, complejidad, niveles de atención de cada establecimiento. Esto se establece en la norma dictada por el MINSAL (MINSAL, 2013), en donde parte de estos criterios son abordados en su definición. Lo que, de alguna forma queda pendiente en esta propuesta, es determinar si esta clasificación administrativa también se relaciona con la casuística de los establecimientos de salud.

En la actualidad no se han llevado a cabo investigaciones respecto a la selección de características en el contexto planteado, sin embargo, técnicas utilizadas en trabajos similares han sido empleadas en la literatura. Por ejemplo, técnicas de análisis exhaustivo han sido utilizadas para estudiar todas las combinaciones de características, y así seleccionar la combinación que cumple con criterios de calidad que se definan, este tipo de métodos nos aseguran encontrar la mejor solución dentro del conjunto, sin embargo su aplicación se limita al análisis de conjuntos de datos que no superen las 20 características (Doak, 1992). Otras técnicas empleadas se basan en búsqueda exhaustiva, como, por ejemplo, Branch and Bound y Beam Search (Doak, 1992). Estas técnicas de análisis son utilizadas en conjuntos de datos que no poseen más de 20 variables dada su naturaleza de búsqueda. El orden de estos algoritmos es de $O(2^n)$ (Doak, 1992). Adicionalmente, técnicas de clasificación son utilizadas para el problema de selección de características. Al igual que las anteriores, estas técnicas consideran todas las variables del problema y dependiendo de la cantidad de ellas, pueden ser costosas en cuanto a tiempo de ejecución o limitar el número de variables seleccionadas. Ejemplo de ellos son: K-nearest neighbors (KNN), en su variante Random k-nearest neighbors (RKNN) (Li et al., 2011) y Random Forest (RF) (Genuer et al., 2010).

Por otro lado, para realizar análisis sobre un elevado número de variables algunos autores han empleado técnicas heurísticas (McShan et al., 2003) (Yoo & Lafortune, 1989). Estas corresponden a estrategias de búsqueda de soluciones a problemas de una manera no exhaustiva, utilizando estrategias inteligentes para encontrar soluciones acordes a una función de rendimiento. Es importante destacar que las técnicas heurísticas, en la mayoría de los casos, no garantizan la obtención de la mejor solución para atender un problema en particular, aunque si son capaces de encontrar buenas soluciones según su configuración y tiempos de ejecución (Gómez, 2014).

En este contexto existen técnicas heurísticas con distintas estrategias, tales como,

métodos de descomposición, los cuales consisten en la división de un problema en sub-problemas más pequeños de fácil resolución (Lian & Castelain, 2009), métodos inductivos, los cuales consisten en identificar propiedades en sub-problemas menores, que sean extensibles al problema global, métodos de reducción, los cuales consisten en reducir el espacio de búsqueda de soluciones, métodos constructivos, los cuales se basan en construir una solución seleccionando elementos identificados como 'mejores opciones', en un proceso iterativo que obtiene como resultado una solución final (Bräsel et al., 2008), finalmente los métodos de búsqueda local, los cuales consisten en realizar una búsqueda de soluciones cercanas a una solución ya encontrada (Noman & Iba, 2007) (Gómez, 2014).

Adicionalmente a estas técnicas, existen las metaheurísticas, que orquestan las técnicas heurísticas en su búsqueda a una solución. Su aplicación es genérica, y puede ser en distintos contextos (Gómez, 2014). Entre ellas destacan los algoritmos genéticos (Masilamani et al., 2010), búsqueda Tabú (Zhang & Sun, 2002), optimización por colonia de hormigas (Chen et al., 2010), entre otras estrategias.

En la actualidad el DEIS dispone de alrededor de 6000 variables que se registran anualmente desde los hospitales: egresos, número de procedimientos, patologías registradas, categorización de pacientes y otros indicadores (DEIS, Año 2014). Esto es determinante desde el punto de vista computacional, ya que esta investigación busca determinar aquellas que, individualmente o combinadas, permiten alcanzar desde el punto de vista de la casuística hospitalaria, la clasificación administrativa propuesta por el MINSAL para los hospitales. Una evaluación de todas las combinaciones posibles tendría efectuar el siguiente número de evaluaciones.

$$\sum_{i=1}^{n} \binom{n}{i} \tag{2.9}$$

En este contexto se propone abordar el problema con la aplicación de una técnica metaheurística (Kira et al., 1992). Las principales conclusiones luego de la revisión literaria se pueden resumir como:

- No se encontraron publicaciones que hayan abordado el problema en el contexto estudiado, no obstante, sí en otras áreas de aplicación.
- El uso de una técnica exhaustiva es computacionalmente no recomendable para abordar el problema actual, dado el total de variables involucradas en el problema.
- El uso de una técnica metaheurística podría adaptarse de mejor forma para abordar el problema en relación con otros tipos de estrategias, en donde se buscarán buenas soluciones, utilizando heurísticas de búsqueda inteligente. En este aspecto, se pretende identificar las variables que definen de mejor forma la categorización establecida por el

MINSAL, ya que se desconoce cómo esta definición tiene relación con la casuística de los diferentes establecimientos.

CAPÍTULO 3. DESARROLLO DE INVESTIGACIÓN UTILIZANDO KDD

La investigación será realizada bajo la metodología KDD, la cual se detalla de la siguiente forma:

3.1 SELECCIÓN

El proceso de selección se define como aquel, en donde se seleccionan los datos a trabajar de forma inicial. En este caso se describirán los archivos iniciales seleccionados disponibles en la plataforma del DEIS.

3.1.1 Descripción del conjunto de datos

Los datos utilizados en el estudio se pueden categorizar en 3 grupos:

1. Altas realizadas por prestación de salud en cada establecimiento de salud.
2. Estadísticos hospitalarios de interés, o también llamado registro estadístico mensual (REM 20).
3. Categorización de pacientes según niveles de cuidado, asociado al instrumento de categorización de usuarios por dependencia y riesgo (CUDYR) (García G & Castillo F, 2000).

Una de las problemáticas señaladas en un comienzo, corresponde a la cantidad de variables involucradas en la investigación, y dicha cantidad se detalla en la tabla 3.1.

Los datos seleccionados corresponden a aquellos generados para los años 2014, 2015, 2016, 2017 y 2018. En donde es necesario considerar que algunas variables pueden cambiar entre años, y en ese sentido, el análisis a realizar se debe considerar una evaluación año a año. Para detalles en el registro estadístico mensual, ver Anexo A

Serie	Rem	Conjunto de datos #Variables				
		2014	2015	2016	2017	2018
A	REM-A01	56	58	58	58	57
	REM-A02	14	14	14	14	14
	REM-A03	93	100	100	111	114
	REM-A04	37	47	46	56	57
	REM-A05	158	168	160	169	169
	REM-A06	65	66	67	61	61
	REM-A07	138	138	144	96	94
	REM-A08	84	95	94	107	112
	REM-A09	188	175	178	181	182
	REM-A11	66	66	66	96	182
	REM-A19a	112	112	112	128	128
	REM-A19b	30	30	30	32	32
	REM-A21	55	51	51	47	47
	REM-A23	88	88	91	91	91
	REM-A24	18	20	21	24	27
	REM-A25	88	88	90	90	90
	REM-A26	46	48	48	51	51
	REM-A27	53	66	79	79	79
	REM-A28	143	149	149	162	162
	REM-A29	0	0	0	60	62
	REM-A30	0	0	0	69	86
	REM-A31	0	0	0	0	39
BS	REM-BS0	210	213	213	2990	2877
	REM-BS17	127	129	129	168	191
	REM-BS17a	1881	1881	1881	0	0
	REM-BS17c	441	511	512	0	0
	REM-BS17d	2202	2338	2339	0	0
D	REM-D15	42	42	42	65	45
	REM-D16	10	10	15	15	15
REM-20		275	275	286	286	286
CUDYR		300	300	312	312	312

Tabla 3.1: Conjunto de datos
Fuente: Elaboración propia, 2019.

3.1.2 Descripción de clases

Las clases utilizadas para el desarrollo de la investigación, que definen los grupos establecidos por el MINSAL en su norma (MINSAL, 2013), están asociadas a las siguientes categorías:

1. Alta complejidad Adulto: Hospitales y centros de referencia locales en ciudades principales.
2. Mediana complejidad: Hospitales ubicados en ciudades de carácter secundario.

3. Baja complejidad: Hospitales ubicados en ciudades pequeñas o áreas rurales.
4. Alta complejidad niños: Hospitales de alta complejidad cuyo foco es la atención de tipo pediátrica.
5. Alta complejidad psiquiatría: Hospitales de alta complejidad cuyo foco es la atención de tipo psiquiátrica.

Id	Detalle de clase Descripción	#Hospitales
1	Baja complejidad	101
2	Alta complejidad adulto	57
3	Mediana complejidad	21
4	Alta complejidad niños	4
5	Alta complejidad psiquiatría	4

Tabla 3.2: Detalle de clases
Fuente: Elaboración propia, 2019.

Un punto importante en esta descripción de clases, es el desbalanceo observado en los establecimientos considerados. En donde la clase dominante (1) es cercana al 60% de la cantidad total de observaciones, situación contraría a las clases de menos observaciones (4)(5) que no superan el 3% respectivamente.

3.1.3 Modelo relacional

El uso de un motor de base de datos relacional no es obligatorio en el procesamiento de datos, sin embargo su implementación disminuye los tiempos de consulta a la hora de trabajar con los datos. Esta reducción en los tiempos de consulta está dada por la impletación de indices asignados a las variables. El almacenamiento de los datos será en un motor de base de datos MySQL según modelo de datos descrito (Figura 3.1). El modelo de datos propuesto almacenará de forma estructurada los datos disponibles en el DEIS, de forma de mejorar los tiempos de consulta en las etapas futuras dentro de la metodología de descubrimiento de conocimiento mediante minería de datos. Para este planteamiento se debe considerar que los datos tabulados se registran de forma mensual, y su consulta es recurrente en la generación de la matriz de distancias.

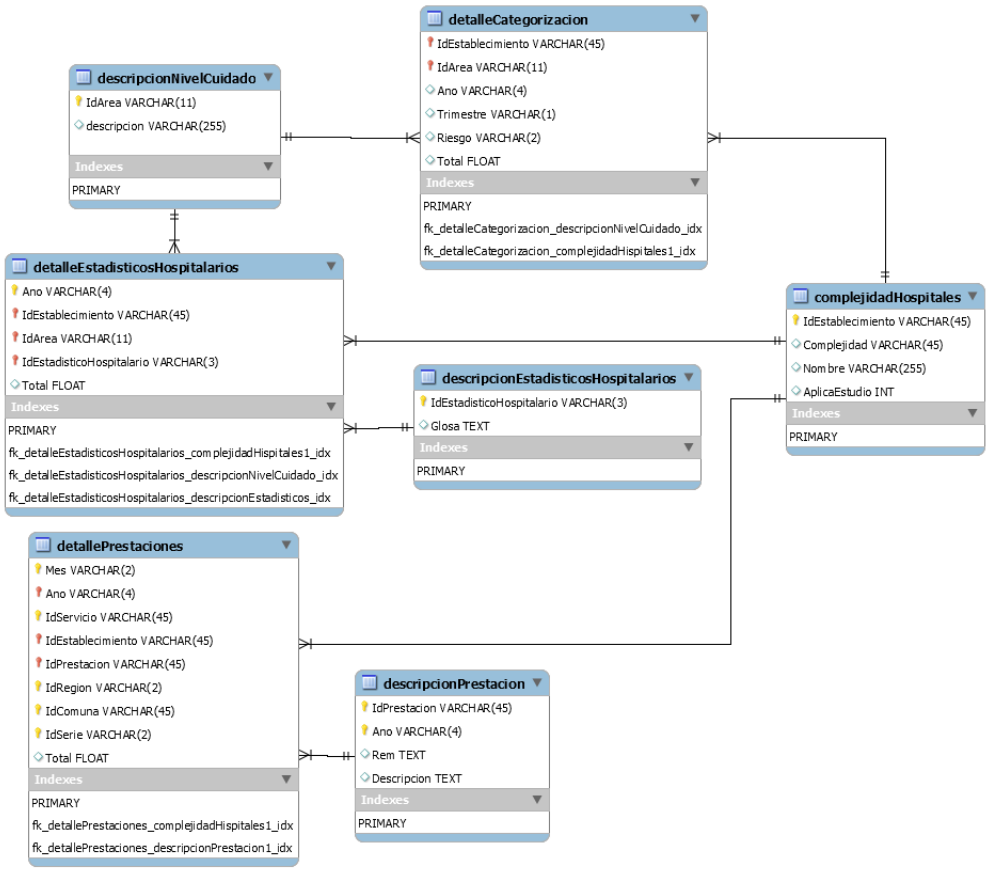

Figura 3.1: Modelo de almacenamiento de datos
Fuente: Elaboración propia, 2019.

3.2 PREPROCESAMIENTO

La base de datos generada no posee errores generales en cuanto a sus datos. Sin embargo, con respecto a la cantidad de variables relacionadas con el estudio, se hace necesario la aplicación de algún test que disminuya de forma previa la cantidad de variables consideradas de forma inicial. Esta disminución de variables tiene por función quitar del estudio aquellas variables que no aportan información adicional. Adicionalmente, la consolidación y almacenamiento de los datos en el modelo de datos propuesto (Sección 3.1.3) se hace necesario para disminuir los tiempos de respuesta en la búsqueda de datos.

3.2.1 Ordenamiento y almacenamiento

Para trabajar de forma estructurada los datos seleccionados, se hizo necesario cargar la información disponible en el DEIS en una estructura de datos acorde al problema. Se extraerán los archivos disponibles en el DEIS (en formato de texto plano, csv y excel) y se almacenarán en un motor de base de datos relacional.

Para esta etapa es necesario generar un modelo de extracción de datos (Figura 3.2):

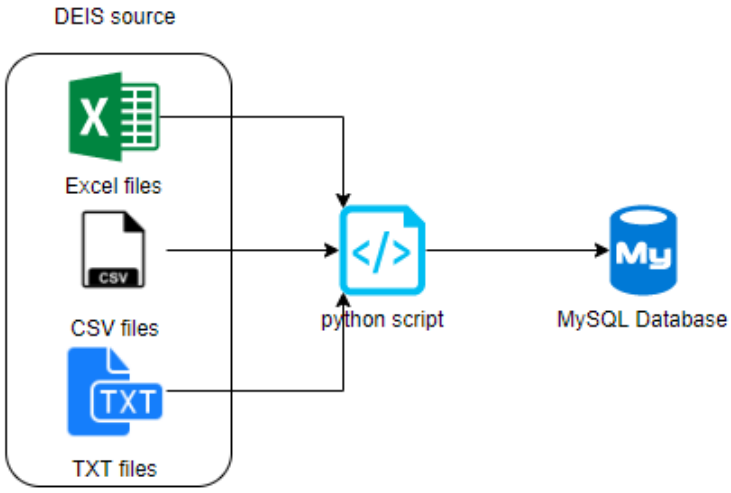

Figura 3.2: Modelo de consolidación de datos
Fuente: Elaboración propia, 2019.

3.2.2 Disminución de dimensionalidad previa

Los test utilizados para la disminución de la cantidad inicial de variables corresponden a Kruskal-Wallis (Kruskal & Wallis, 1952) y Chi-Squared (Romanski et al., 2018). Para el caso de Kruskal-Wallis se utiliza la función *kruskal.test* perteneciente al paquete *PMCMRplus* (Pohlert, 2018), en donde sus variables de entrada corresponden a los datos que obtuvo la variable a evaluar en los distintos establecimientos de salud, junto a su complejidad. Como salida de esta función, se obtiene el *p-value* asociado a la variable evaluada, que dada una confiabilidad de 95%, se asume que para valores de *p-value* mayores al 0.05, la variable no diferencia a los grupos de establecimientos propuestos, en estos casos la variable será descartada del estudio. Por otro lado, el test Chi-squared se aplica de forma previa sobre las variables consideradas posterior a la aplicación del test de Kruskal-Wallis, y aquellas que sean dependientes con la clase, serán

descartadas en la investigación para reducir la dimensionalidad del problema, ya que no entregan información adicional.

Finalmente la cantidad de variables consideradas para el estudio se redujo de la siguiente manera (Tabla 3.3)

Tipo de test	Conjuntos de datos #Variables por año				
	2014	2015	2016	2017	2018
Cantidad inicial	6445	6703	6729	5020	5064
Cantidad posterior a test Kruskal-Wallis	4081	4159	4142	4095	3681
Cantidad posterior a test Chi-squared	1685	1723	1633	1483	1488

Tabla 3.3: Conjuntos de datos preprocesados
Fuente: Elaboración propia, 2019.

3.3 TRANSFORMACIÓN

Esta etapa los datos son tratados con el fin de generar la matriz inicial requerida por el análisis mediante minería de datos. En la siguiente sección se detalla la estructura de la matriz objetivo, y su generación mediante el uso de base de datos (modelo relacional) y algoritmos de ordenamiento.

3.3.1 Matriz de datos

La técnica de minería de datos, en particular la heurística basada en algoritmo genético, requiere de una matriz de datos que considere todos los datos relacionados a los establecimientos y su valor asociado a cada variable de las categorías detalladas en la sección 3.1.1.

La construcción de esta matriz es mediante la consolidación de los datos de forma anual para cada establecimiento y variable asociada, en consecuencia, se construyen cinco matrices, una por cada año de estudio. Cada fila dentro de esta matriz representa a un establecimiento de salud, mientras que las columnas representan las variables de estudio. En la figura 3.3 se detalla su estructura, en donde i corresponde a la cantidad de establecimientos de salud considerados y j la cantidad de variables consideradas para el año de estudio.

$$M^{i \times j} = \begin{pmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,j} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ M_{i,1} & M_{i,2} & \cdots & M_{i,j} \end{pmatrix}$$

Figura 3.3: Matriz inicial de datos construida
Fuente: Elaboración propia, 2019

3.4 MINERÍA DE DATOS

Para abordar el problema planteado en esta investigación es posible utilizar técnicas asociadas a la selección de características. Sin embargo, técnicas tradicionales utilizadas en diversos contextos no son de utilidad, dada la gran cantidad de variables consideradas en este estudio (Sección 2.2). Dada esta dificultad, la implementación de un algoritmo genético para la selección de características surge como una estrategia viable para tratar el problema planteado.

3.4.1 Algoritmo genético para selección de características

Un algoritmo genético posee fases principales que pueden ser modificadas dada la naturaleza del problema abordado. El algoritmo genético propuesto para abordar el problema se detalla en el algoritmo 3.1.

Caracterización de un individuo

Cada individuo es representado por un vector de valores 0 y 1, que representan la consideración (en el caso de ser 1) de la variable que se identifica en la posición dentro del vector. Además un valor entre [KME, CLA, PAM, FAN] que representa a la técnica de agrupamientos utilizada para llegar a la solución (Sección 2.1.3). De forma gráfica, un individuo se caracteriza según lo representado en la figura 3.4, en donde i corresponde a la cantidad de variables consideradas para el año de estudio, X_i es el indicador de consideración de la variable correspondiente a la posición i, y su valor puede ser 0 o 1, finalmente T representa a la técnica de agrupamientos utilizada.

Algoritmo 3.1: Propuesta de algoritmo genético

input : F, d, ps, cr, mr, rr
output: Una población P de características seleccionadas y la clasificación generada por ellas

```
                                                    // Crea población inicial
1  P ← InitialPopulation (F, ps, d)
2  while stop condition not reached do
3  |  for each cs ∈ P do
   |  |                                            // Aplicación de Crossover
4  |  |  begin Operaciones en Crossover
5  |  |  |  [F1, F2] ← TournamentSelection (P)
6  |  |  |  Q[cs] ← CrossOver (F1,F2)
7  |  |  end
8  |  end
9  |  for each ms ∈ P do
   |  |                                            // Aplicación de Mutación
10 |  |  begin Operaciones en Mutación
11 |  |  |  Q[ms] ← Mutation (P)
12 |  |  end
13 |  end
14 |  for each rs ∈ P do
   |  |                                  // Aplicación de operaciones aleatorias
15 |  |  begin Operaciones aleatorias
16 |  |  |  Q[rs] ← Random (P)
17 |  |  end
18 |  end
19 |  P ← Selection (P,Q,ps)
20 end
21 return (P)
```

$$Individuo = \begin{pmatrix} X_1 & X_2 & \cdots & X_i & T \end{pmatrix}$$

Figura 3.4: Caracterización de un individuo
Fuente: Elaboración propia, 2019

Matriz de distancia

Una matriz de distancia es aquella que describe la distancia en distintos elementos en un espacio. Para su cálculo es necesario considerar dos elementos fundamentales, en primera instancia la métrica de distancia, particularmente en este estudio se utilizan dos métricas de cálculo que serán evaluadas y comparadas en sus resultados. Estas métricas corresponden a Euclidiana (sección 2.2) y basada en correlación de Pearson (sección 2.3). Por otro lado, también es necesario considerar la construcción de esta matriz utilizando los datos parciales que la solución evaluada está representando. Según lo visto en la caracterización de cada individuo (21), este

individuo solo representa la consideración de ciertas variables y una técnica de clúster para la representación de una solución, en dicho aspecto, la construcción de la matriz de distancia solo considera los valores normalizados de dichas variables.

La matriz de distancia es una matriz de $n \times n$, en donde n representa la cantidad de establecimientos de salud considerados en el estudio, y cada valor en la intersección i, j, con $0 < i < n$ y $0 < j < n$, representa la distancia que existe en el espacio entre el establecimiento i y el establecimiento j, utilizando alguna métrica de distancia para ello.

$$MD^{n \times n} = \begin{pmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,j} & \cdots & M_{1,n} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,j} & \cdots & M_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ M_{i,1} & M_{i,2} & \cdots & M_{i,j} & \cdots & M_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ M_{n,1} & M_{n,2} & \cdots & M_{n,j} & \cdots & M_{n,n} \end{pmatrix}$$

Figura 3.5: Esquema matriz de distancia
Fuente: Elaboración propia, 2019

3.4.2 Parametrización

Los parámetros de inicialización del algoritmo genético se generan mediante el paquete de R *irace* (López-Ibáñez et al., 2016). El paquete *irace* es un software que permite la evaluación de diferentes configuraciones iniciales de un algoritmo, y utilizando estos mismos, obtiene la configuración con mejores resultados. Particularmente en este estudio, los rangos de valores utilizados son: generaciones [1:300], tamaño de la población [1:300], porcentaje de la población con mutación [0:1], porcentaje de la población con *crossover* [0:1], balanceo de clases mediante SMOTE (Chawla et al., 2002), implementado en R para generar elementos y disminuir la brecha existente entre elementos de clases [TRUE,FALSE] y métodos de inicialización [KNN, RF, RANDOM].

Operaciones genéticas

Las operaciones genéticas definen la base de un algoritmo genético, la aplicación de ellas permite establecer la lógica desarrollada en torno a él. En este aspecto se definen 4

operaciones como las básicas dentro del algoritmo aplicado:

- Inicialización: la etapa de inicialización consiste en la generación de la población inicial de individuos que serán considerados por el algoritmo genético. Esta generación de los individuos iniciales puede darse de dos formas. **Generación inicial aleatoria**, cada uno de los individuos es generado de forma aleatoria en base a su caracterización. **Generación inicial con selección de características**, previo a la generación de los individuos, es posible obtener las variables más importantes dentro del estudio, en base a alguna técnica de selección de características, particularmente en este caso, se utilizarán dos técnicas para ello: *KNN* y *Random Forest* (Sección 2.1.2).

- *Crossover*: la etapa de *Crossover* en los sistemas biológicos, consiste en la mezcla del material genético de dos individuos, para la generación de un nuevo individuo con material genético de ambos.

  En el contexto del algoritmo genético, la caracterización de dos individuos es mezclado, esto quiere decir que se considera un nuevo individuo con la caracterización de dos individuos seleccionados previamente, en donde mediante una ”probabilidad de consideración” se utiliza la caracterización de un individuo u otro como valor para una determinada variable. En este estudio, el valor para la probabilidad considerado para un *crossover* es de 0,5. Adicionalmente, se considera la creación de un cuarto individuo, que considera los valores de las variables no considerados por la probabilidad de *crossover*.

$$Individuo_1 = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & \cdots & X_i & T \end{pmatrix}$$

$$Individuo_2 = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & \cdots & X_i & T \end{pmatrix}$$

$$Crossover_1 = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & \cdots & X_i & T \end{pmatrix}$$

$$Crossover_2 = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & \cdots & X_i & T \end{pmatrix}$$

Figura 3.6: Ejemplo para operación genética: crossover
Fuente: Elaboración propia, 2019

- Mutación: la etapa de mutación consiste en la modificación del material genético de un individuo, que necesariamente genera un nuevo individuo. En el contexto del algoritmo genético, la caracterización de un individuo es modificada por su valor opuesto, en donde mediante una ”probabilidad de cambio” se modifica el valor asociado a una variable dentro de la caracterización.

Para el caso particular de este estudio, el valor para la probabilidad de cambio es de 0,1. Adicionalmente, al igual que en la etapa de *crossover*, se considera la caracterización del individuo complementario al resultante en la aplicación de la mutación, que se evalúan de forma simultánea según la métrica de calidad asociada, para finalmente seleccionar al mejor individuo resultante.

$$Individuo_1 = \begin{pmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & \cdots & X_i & T \end{pmatrix}$$

$$Mutation_1 = \begin{pmatrix} -X_1 & -X_2 & X_3 & X_4 & -X_5 & -X_6 & \cdots & -X_i & T \end{pmatrix}$$

$$Mutation_2 = \begin{pmatrix} X_1 & X_2 & -X_3 & -X_4 & X_5 & X_6 & \cdots & X_i & T \end{pmatrix}$$

Figura 3.7: Ejemplo para operación genética: mutación
Fuente: Elaboración propia, 2019

- Selección: la etapa de selección consiste en la selección de los mejores individuos, esta etapa posee una dirección de búsqueda, que está basada en el *fitness*. En el contexto del algoritmo genético, cada caracterización de individuo es evaluada e interpretada como una solución del problema, para generar su evaluación mediante la métrica de calidad asociada. Se considera como individuos a evaluar los elementos de la población inicial en cada iteración y los elementos generados mediante los operadores genéticos, obteniendo una población con los mejores elementos de ambos conjuntos. Esta nueva población será de tamaño equivalente al de la población inicial y de esta forma, será el conjunto inicial de la siguiente generación.

Técnicas de clúster

Distintas técnicas de *clúster*, o también llamadas técnicas de agrupamientos, pueden ser utilizadas para determinar grupos de observaciones que comparten características y así establecer parámetros comunes entre ellos. El objetivo en cada una de ellas es el mismo, sin embargo varían en su estrategia para hacerlo, y este dice relación con agrupar las observaciones en una cantidad de grupos. En particular en esta investigación, se utilizarán 4 técnicas de agrupamientos, definida en la caracterización de cada individuo.

Evaluación de individuos

El proceso de evaluación de cada individuo en un algoritmo genético está directamente relacionado con la etapa de selección. En este estudio se realizarán distintas ejecuciones del algoritmo, con distintas métricas de calidad, dando a conocer las que obtienen mejores resultados. Para realizar la evaluación de un individuo, en primera instancia se debe considerar la caracterización del mismo, y generar mediante una técnica de agrupamiento, los grupos que el individuo está asignando a cada una de las observaciones. Finalmente aplicar una medida de comparación entre agrupamientos, en este caso las medidas de calidad 2.1.4, que describa su similitud con el agrupamiento definido por el MINSAL. En particular las métricas de evaluación para los individuos serán, índice de Jaccard (sección 2.1.4), índice ARI (sección 2.1.4) y índice Fowlkes-Mallows (sección 2.1.4).

3.5 RESULTADOS

En esta sección se presentan los resultados obtenidos en la investigación. El detalle de ellos es en relación a los años considerados.

Parametrización

Utilizando la estrategia *Irace* (sección 3.4.2), los parámetros óptimos para la ejecución del algoritmo genético son: generaciones = 100, tamaño de la población = 100, $cr = 0.6$, $mr = 0.2$, smote = $FALSE$ y inicialización = $random$. Estos parámetros fueron utilizados en las ejecuciones de todas las pruebas del algoritmo genético propuesto. La consideración de la variable smote = $FALSE$, se debe a que la incorporación de elementos utilizando este algoritmo produce ruido en los resultados obtenidos, generando valores bajos para las funciones de *fitness*

Resultados generales

Los resultados generales dicen relación con las métricas de calidad observadas en la distintas ejecuciones del algoritmo. Se debe considerar que cada una de ellas tuvo 31 iteraciones y se obtuvieron los siguientes resultados (Tabla 3.4). Esta cantidad de iteraciones se consideraron bajo distintas estrategias abordadas en la literatura (Villalobos-Cid et al., 2018), (Cant-Paz & Goldberg, 2003)

Resultados	Métrica de distancia			
	Euclidiana		Correlación	
Métricas de Calidad	Mediana	Máximo	Mediana	Máximo
	2014			
Índice adjusted rand	0.838	0.848	0.926	**0.944**
Índice Fowlkes-Mallows	0.899	0.911	0.954	**0.966**
Índice Jaccard	0.817	0.824	0.911	**0.933**
	2015			
Índice adjusted rand	0.842	0.851	0.939	**0.96**
Índice Fowlkes-Mallows	0.901	0.915	0.965	**0.977**
Índice Jaccard	0.821	0.844	0.933	**0.955**
	2016			
Índice adjusted rand	0.794	0.812	0.934	**0.953**
Índice Fowlkes-Mallows	0.875	0.889	0.958	**0.969**
Índice Jaccard	0.779	0.797	0.916	**0.938**
	2017			
Índice adjusted rand	0.825	0.847	0.855	**0.915**
Índice Fowlkes-Mallows	0.891	0.903	0.914	**0.948**
Índice Jaccard	0.804	0.82	0.84	**0.901**
	2018			
Índice adjusted rand	0.806	0.826	0.889	**0.903**
Índice Fowlkes-Mallows	0.881	0.898	0.936	**0.949**
Índice Jaccard	0.791	0.808	0.88	**0.909**

Tabla 3.4: Resultados obtenidos por años
Fuente: Elaboración propia, 2019.

Las soluciones identificadas por el algoritmo implementado permiten clasificar de buena forma a los establecimientos de salud, teniendo como referencia la clasificación propuesta por el MINSAL. La métrica de calidad utilizada dentro del algoritmo que tuvo los mejores resultados fue el índice Fowlkes-Mallows utilizando una métrica de calidad basada en correlación, obteniendo

valores en el rango de [0.948, 0.977], lo que indica alto porcentaje de establecimientos bien clasificados respecto a la propuesta, utilizando una métrica de calidad que establece una relación de comportamiento entre establecimientos. En teoría este indicador posee mejor rendimiento en grupos cuyas clases están desbalanceadas, como era el caso en particular de esta investigación, por lo que a priori se observaba como la mejor.

Características seleccionadas

Las características identificadas, son aquellas que, según el análisis realizado, describen la categorización propuesta por el MINSAL, y están presentes en al menos dos de las tres mejores soluciones identificadas utilizando distintas métricas de calidad en la ejecución del algoritmo genético. Para revisar en detalle cada una de estas variables, revisar Anexo B.

En el conjunto de variables identificadas, puede observarse una concordancia con la clase asociada. En particular, aquellos hospitales de baja complejidad (Clase 1) no quedaron representados por el conjunto de variables seleccionadas, esto dice que aquellas variables no tienen observaciones en este tipo de establecimientos ya que pertenecen a patologías más complejas. Por otro lado, los establecimientos de alta complejidad adultos (Clase 2) presentan variaciones en un conjunto amplio de patologías e indicadores asociados al cuidado de pacientes, considerando aquellas patologías de tipo más complejas y que en general, no son atendidas por establecimientos de una complejidad baja. Como por ejemplo, *Litiasis renal, trat. quir. percutáneoc/s ultrasonido*, un procedimiento complejo que no se realiza en establecimientos de baja complejidad. Los establecimientos de mediana complejidad (Clase 3) se ven representados por un grupo muy acotado de variables, que son capaces de distinguir a dichos establecimientos y en determinados casos están solo asociadas a este tipo de establecimientos. Aquellos establecimientos de alta complejidad pediátricos (Clase 4) de igual forma fueron representados por un grupo acotado de variables, aunque con valores pequeños, esto dado que las patologías identificadas también se dan en otros establecimientos de alta complejidad y los recintos con foco pediátrico son de alta complejidad de igual forma. Finalmente los hospitales psiquiátricos fueron representados por variables que solamente se identifican en ellos, con foco en aquellos indicadores de estadía de pacientes en el recinto, asociando esto como una característica particular.

Análisis anual de casuística hospitalaria

El análisis por año es realizado de esta forma por la cantidad de variaciones que existen año a año en las variables registradas por el DEIS (Tabla 3.1), existe un gráfico global para todos los establecimientos de salud, que considera el total de variables identificadas por la técnica implementada, para luego detallar el perfil por grupo.

Análisis global para año 2014, Figura 3.8, en donde se observa un total de 40 variables, estas fueron seleccionadas por la estrategia y sus valores normalizados para el conjunto completo de establecimientos de salud. Adicionalmente, se agrega una columna (costado izquierdo del gráfico) que representa al agrupamiento definido previamente. Además, en el eje horizontal, las variables identificadas por la estrategia, y en el eje vertical, los establecimientos de salud. Finalmente se encuentra, al costado superior izquierdo, el rango de valores representados por sus respectivos colores.

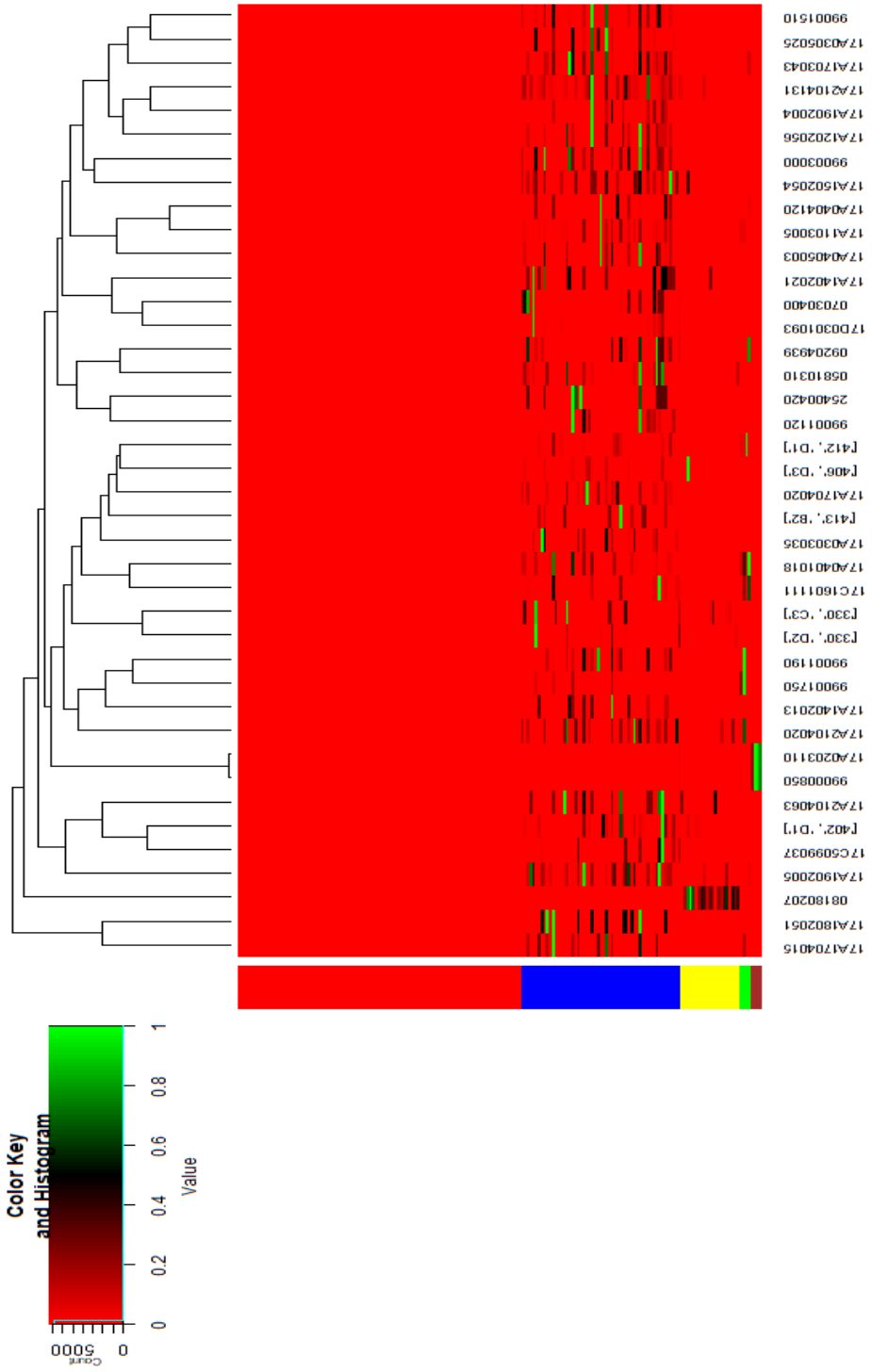

Figura 3.8: Heatmap para variables identificadas en año 2014
Fuente: Elaboración propia, 2019

Perfil para clase 1, establecimientos de baja complejidad, en año 2014 (figura 3.9 y 3.10). En donde es posible visualizar que ninguna de las variables seleccionadas se identifica con este tipo de establecimientos, se asocia este perfil a la selección de variables propias de establecimientos con complejidad más alta, que finalmente son aquellas capaces de diferenciar a los agrupamientos.

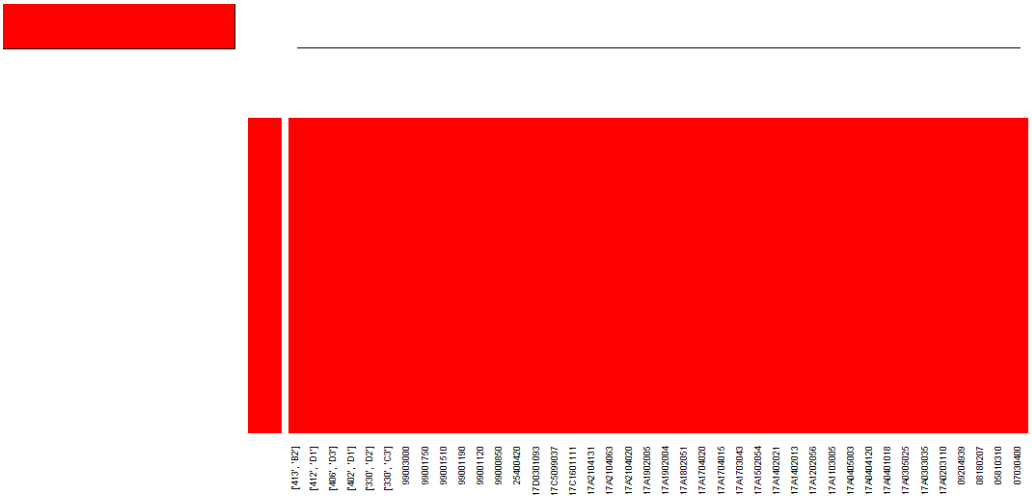

Figura 3.9: Perfil para clase 1 en año 2014
Fuente: Elaboración propia, 2019

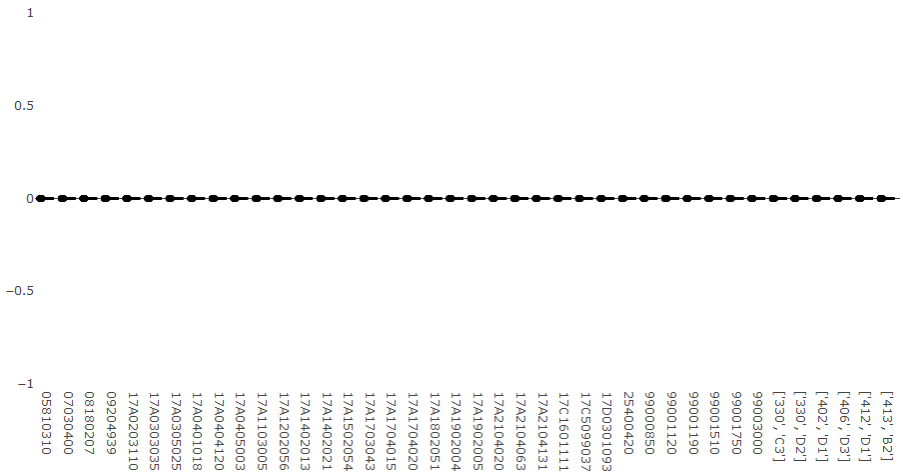

Figura 3.10: Perfil para clase 1 en año 2014
Fuente: Elaboración propia, 2019

Perfil para clase 2, establecimientos de Alta complejidad (Adultos), en año 2014 (figura 3.11 y 3.12). En donde es posible visualizar un conjunto amplio de variables identificadas con este grupo, se asocia este perfil a la cantidad de variables de alta complejidad en los establecimientos.

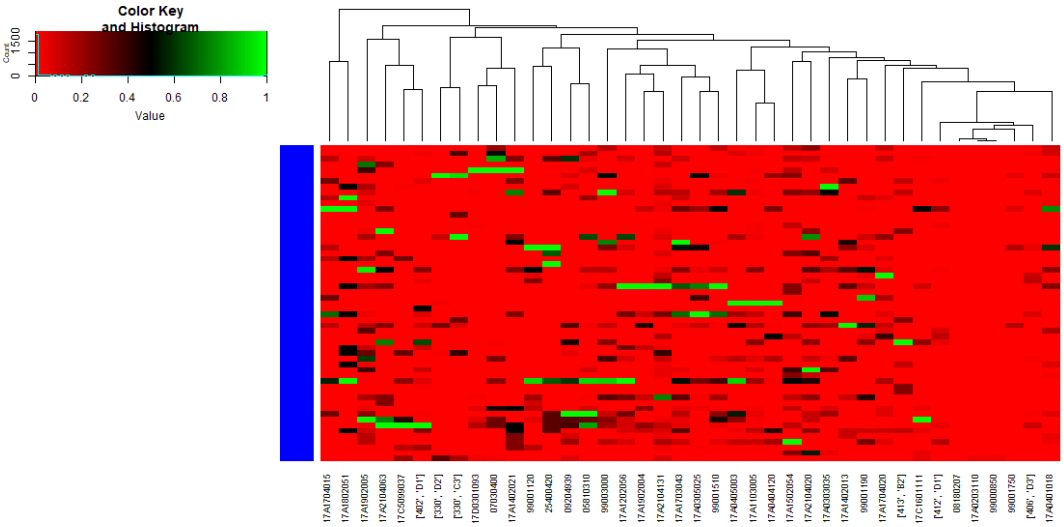

Figura 3.11: Perfil para clase 2 en año 2014
Fuente: Elaboración propia, 2019

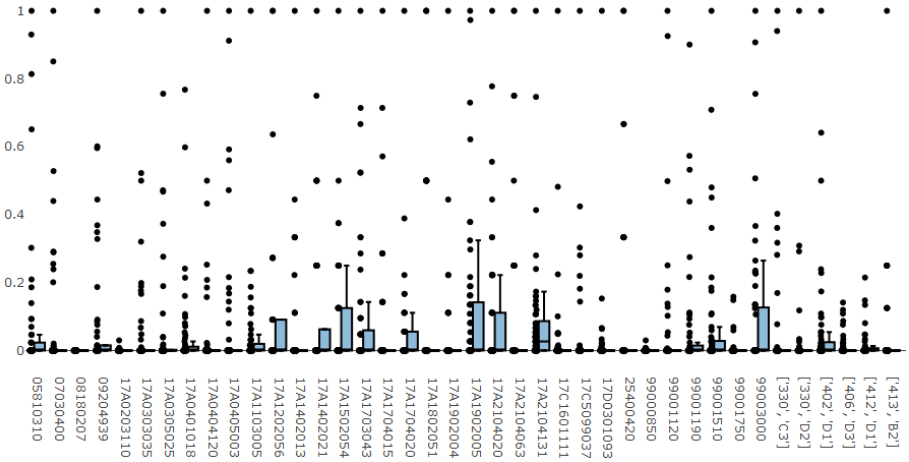

Figura 3.12: Perfil para clase 2 en año 2014
Fuente: Elaboración propia, 2019

Perfil para clase 3, establecimientos de Media complejidad, en año 2014 (figura 3.13 y 3.14). En donde es posible visualizar un conjunto acotado de variables, particularmente la consideración de la variable *"08180207 - ATENCIONES REALIZADAS EN UEH DE HOSPITALES DE MEDIANA COMPLEJIDAD. MATRONA /ÓN)"* es de importancia ya que permite diferenciar a este tipo de establecimientos de los demás, considerando que esta patología solo es registrada en hospitales de mediana complejidad. Este tipo de variables presenta una particularidad dentro del estudio, ya que son variables solamente observadas en algún tipo de establecimientos, y en términos de describir la clasificación, son variables que diferencian de forma natural a los agrupamientos.

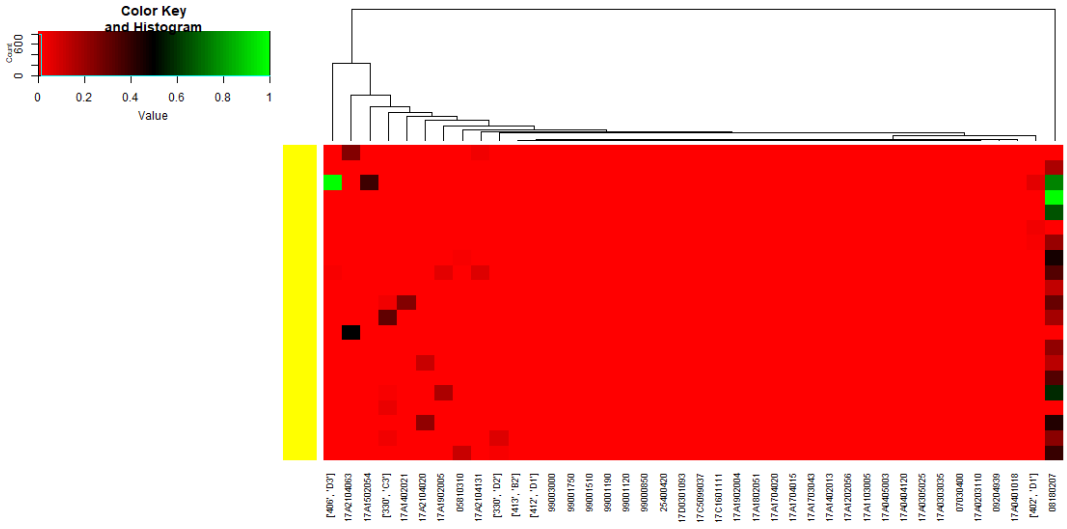

Figura 3.13: Perfil para clase 3 en año 2014
Fuente: Elaboración propia, 2019

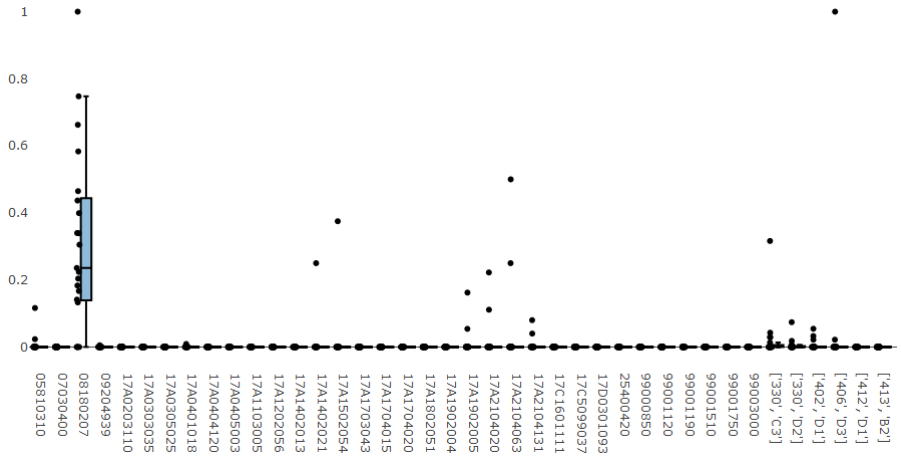

Figura 3.14: Perfil para clase 3 en año 2014
Fuente: Elaboración propia, 2019

Perfil para clase 4, establecimientos de Alta complejidad (Pediatría), en año 2014 (figura 3.15 y 3.16). Acá es posible visualizar el conjunto de variables que describe el perfil de los establecimientos de alta complejidad con foco en el área de pediatría. No se identifican variables relativas específicamente a la atención pediátrica, aunque existen dos patologías identificadas que logran separar a este grupo de establecimientos de los demás, ellas son *"99001190 - Aortografía, en adultos o niños (Incluye proc. rad.)"* y *"99001750 - Recaída tumores sólidos"*.

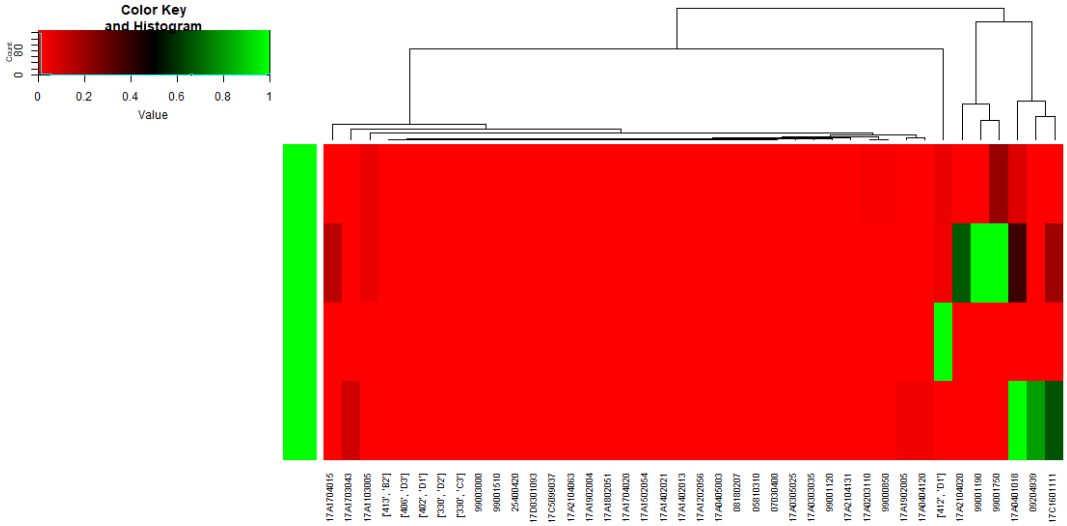

Figura 3.15: Perfil para clase 4 en año 2014
Fuente: Elaboración propia, 2019

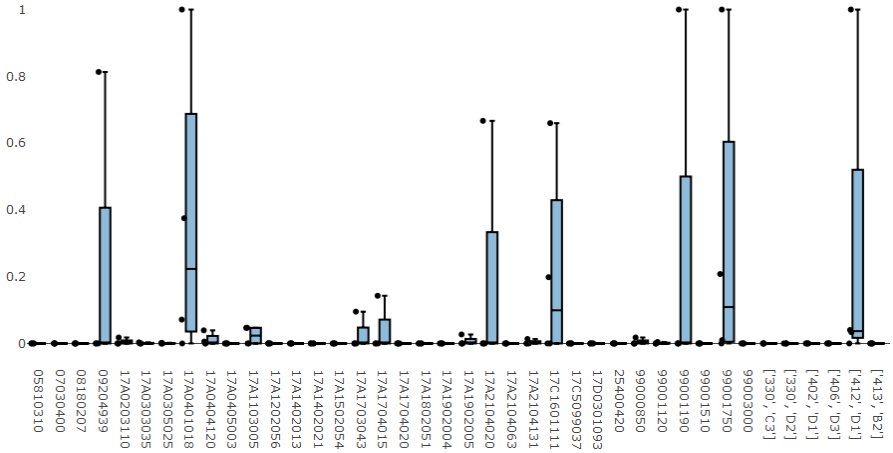

Figura 3.16: Perfil para clase 4 en año 2014
Fuente: Elaboración propia, 2019

Perfil para clase 5, establecimientos de alta complejidad (psiquiátricos), en año 2014 (figura 3.17 y 3.18). El perfil correspondiente a hospitales de alta complejidad con foco en el área psiquiátrica presenta la consideración de dos variables claramente pertenecientes al área de psiquiatría, que permiten diferenciar a este grupo de establecimientos de los demás. Estas variables son *"99000850 - Día cama hospitalización integral psiquiatría mediana estadía"* y *"17A0203110 - Día cama hospitalización integral psiquiatría mediana estadía"*.

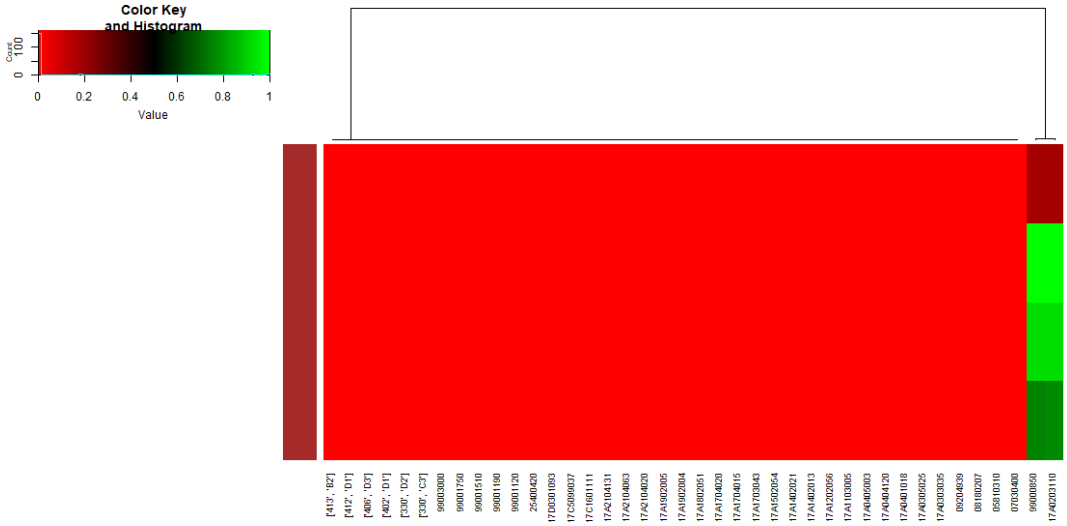

Figura 3.17: Perfil para clase 5 en año 2014
Fuente: Elaboración propia, 2019

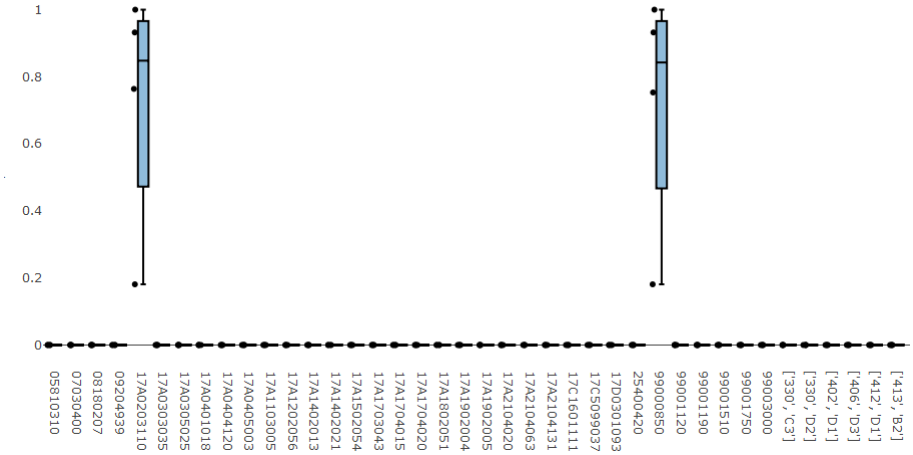

Figura 3.18: Perfil para clase 5 en año 2014
Fuente: Elaboración propia, 2019

Análisis global para año 2015, Figura 3.19, en donde se observa un total de 39 variables seleccionadas por la estrategia. Adicionalmente, se agrega una columna (costado izquierdo del gráfico) que representa al agrupamiento definido previamente.

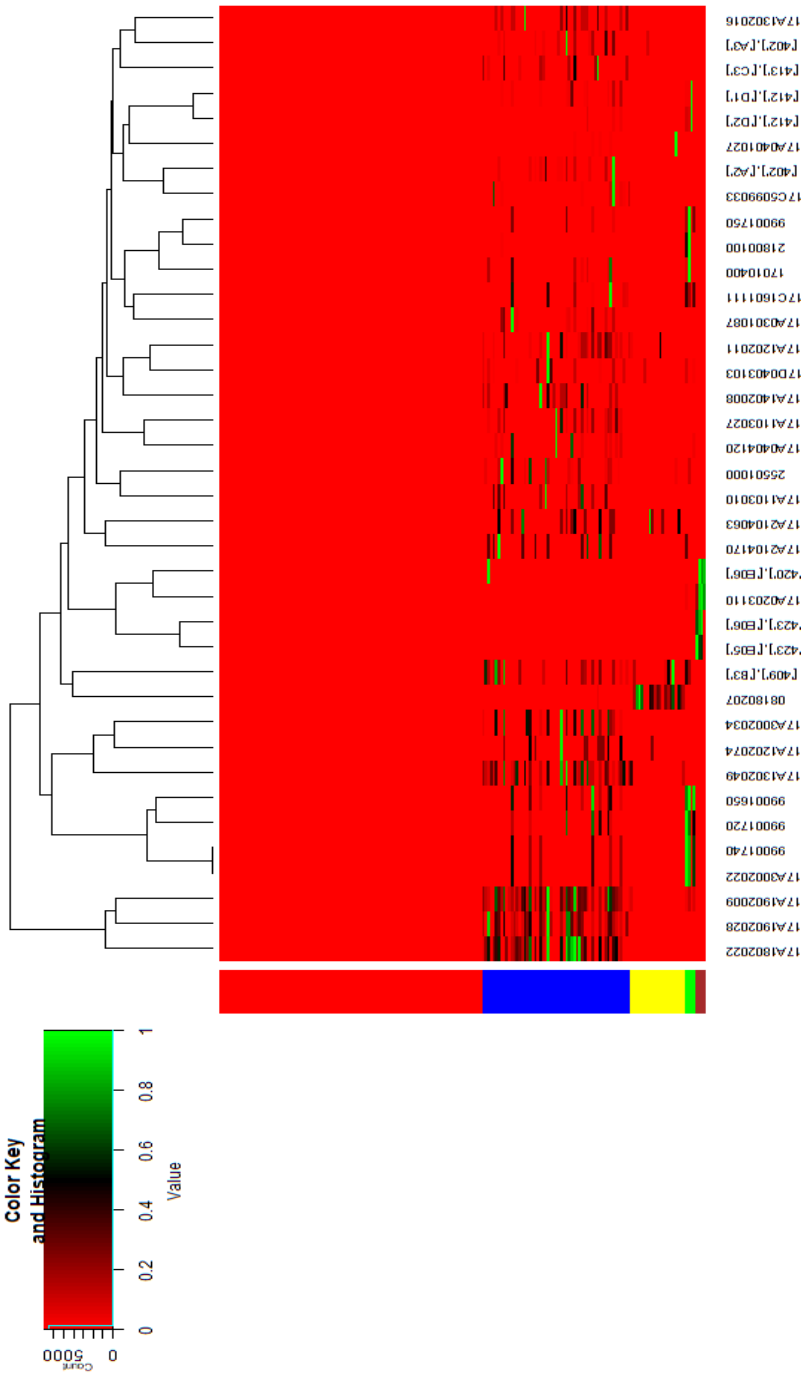

Figura 3.19: Heatmap para variables identificadas en año 2015
Fuente: Elaboración propia, 2019

Perfil para clase 1, establecimientos de baja complejidad, en año 2015 (figura 3.20 y 3.21). En donde se observa el mismo comportamiento del año 2014, ya que las variables consideradas por la estrategia corresponden a establecimientos de mayor complejidad en su operación.

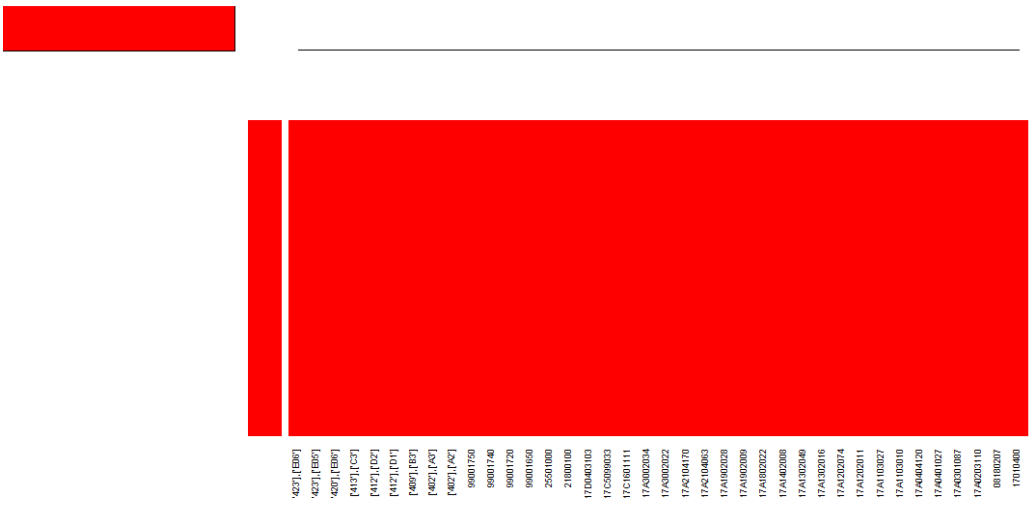

Figura 3.20: Perfil para clase 1 en año 2015
Fuente: Elaboración propia, 2019

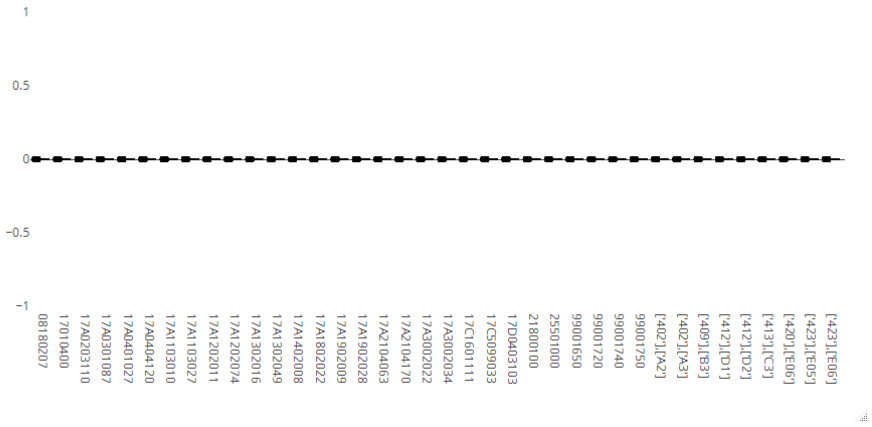

Figura 3.21: Perfil para clase 1 en año 2015
Fuente: Elaboración propia, 2019

Perfil para clase 2, establecimientos de Alta complejidad (Adultos), en año 2015 (figura 3.22 y 3.23). En donde es posible visualizar un conjunto amplio de variables identificadas con este grupo, se asocia este perfil a la cantidad de variables de alta complejidad en los establecimientos, sin embargo destaca la patología *"17A1802051 - Operación de etapificación (incluye esplenectomía, biopsias hepáticas ,de ganglios abdominales y de cresta ilíaca)"* como aquella permite separar a los establecimientos de alta complejidad en la clasificación.

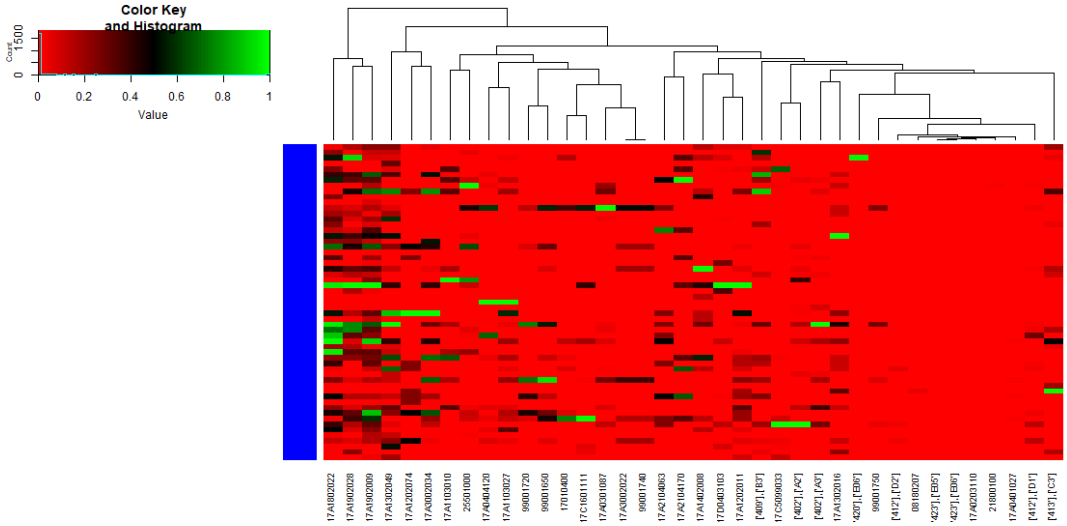

Figura 3.22: Perfil para clase 2 en año 2015
Fuente: Elaboración propia, 2019

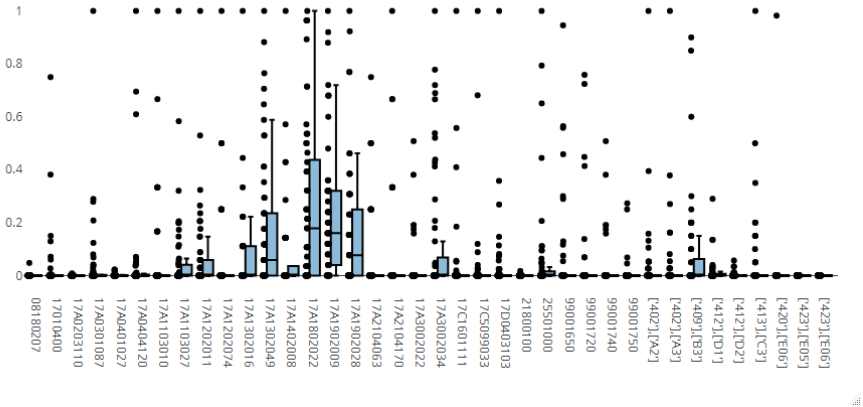

Figura 3.23: Perfil para clase 2 en año 2015
Fuente: Elaboración propia, 2019

Perfil para clase 3, establecimientos de Media complejidad, en año 2015 (figura 3.24 y 3.25). Se observa el mismo patrón que en el año 2014, con una variable que permite distinguir a este tipo de establecimientos de los demás grupos, esta variable es *"08180207 - ATENCIONES REALIZADAS EN UEH DE HOSPITALES DE MEDIANA COMPLEJIDAD. MATRONA /ÓN"*.

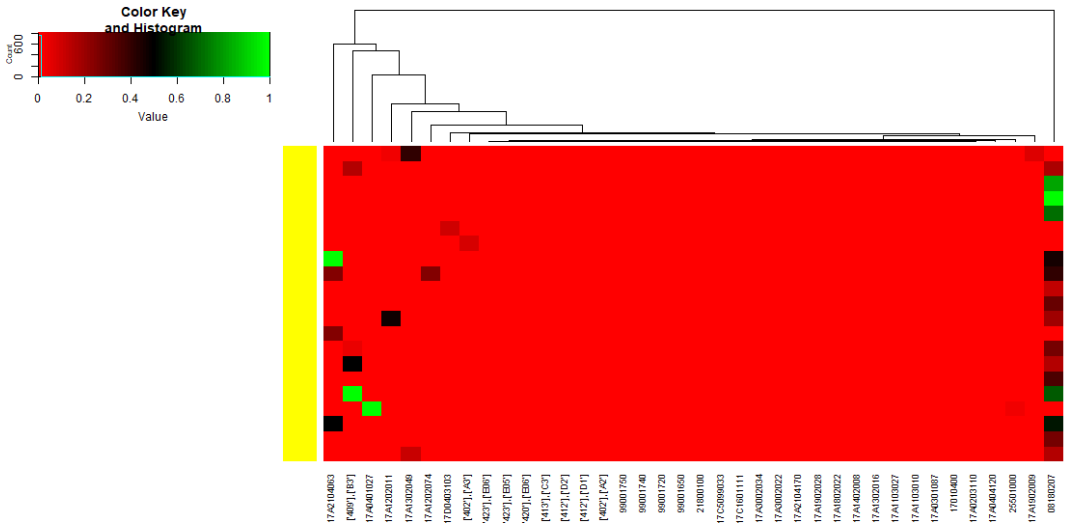

Figura 3.24: Perfil para clase 3 en año 2015
Fuente: Elaboración propia, 2019

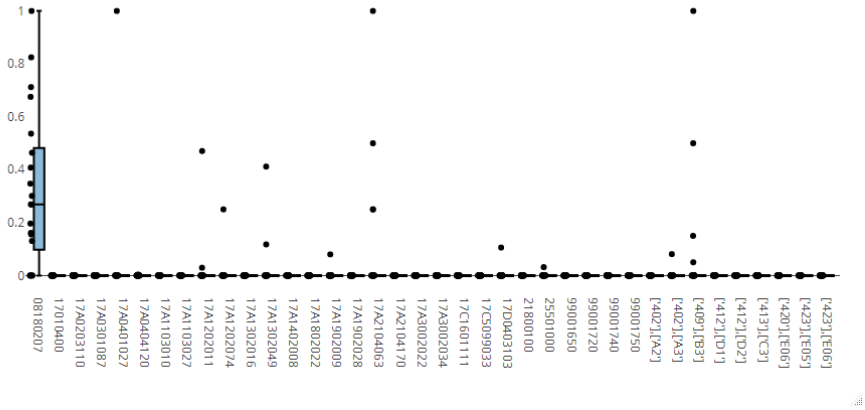

Figura 3.25: Perfil para clase 3 en año 2015
Fuente: Elaboración propia, 2019

Perfil para clase 4, establecimientos de Alta complejidad (Pediatría), en año 2015 (figura 3.26 y 3.27). Acá es posible visualizar el conjunto de variables que describe el perfil de los establecimientos de alta complejidad con foco en el área de pediatría. No se identifican variables relativas específicamente a la atención pediátrica, aunque si destaca la participación de las variables pertenecientes a la categoría de REM-BS0 como diferenciadores de estos establecimientos frente a los demás grupos.

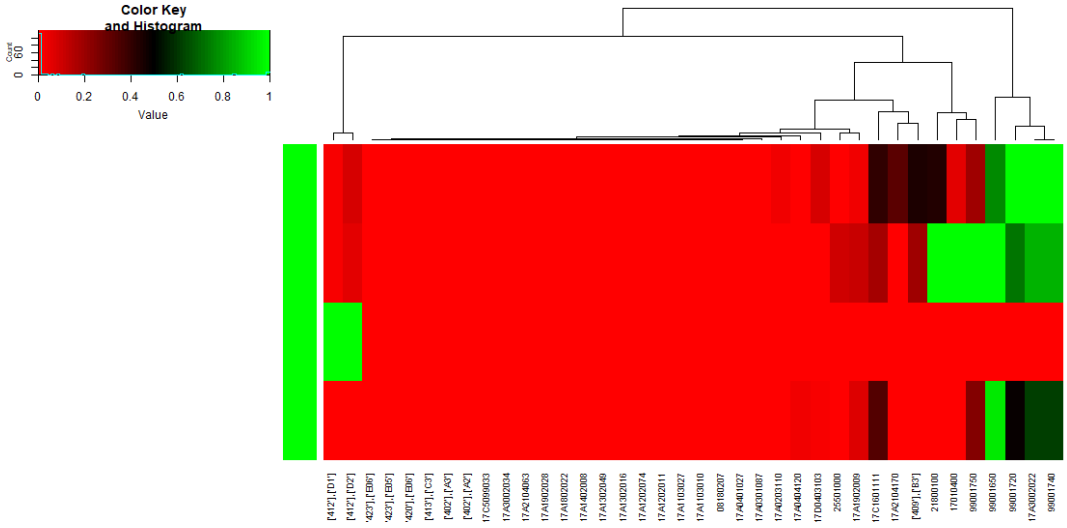

Figura 3.26: Perfil para clase 4 en año 2015
Fuente: Elaboración propia, 2019

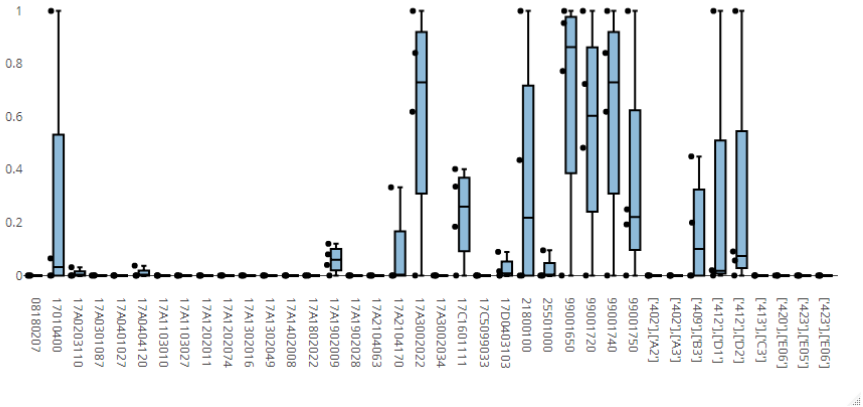

Figura 3.27: Perfil para clase 4 en año 2015
Fuente: Elaboración propia, 2019

Perfil para clase 5, establecimientos de alta complejidad (psiquiátricos), en año 2015 (figura 3.28 y 3.29). Las variables con mayor presencia en el perfil de este grupo, pertenecen de forma exclusiva a establecimientos con foco en la psiquiatría, y en particular hacen referencia a la estadía de los pacientes.

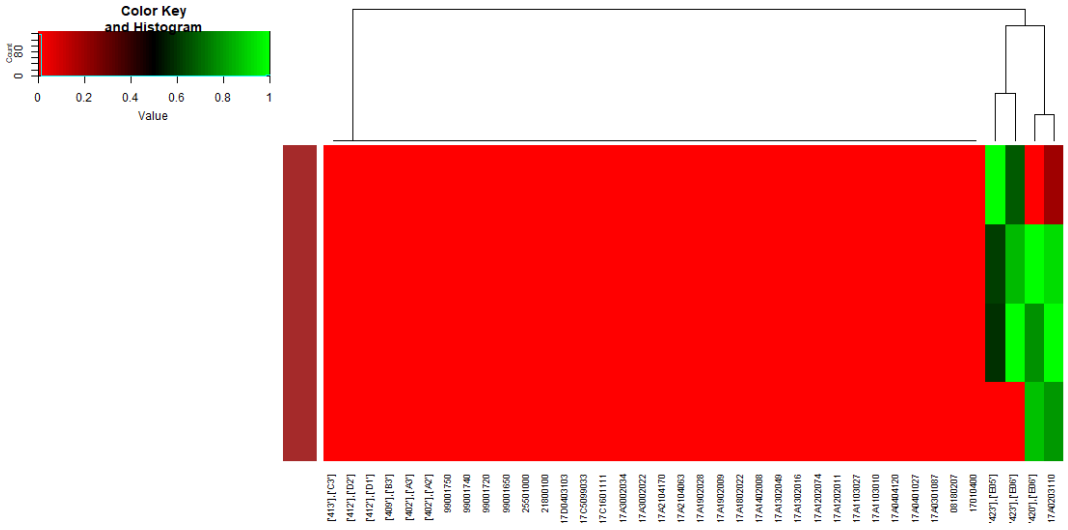

Figura 3.28: Perfil para clase 5 en año 2015
Fuente: Elaboración propia, 2019

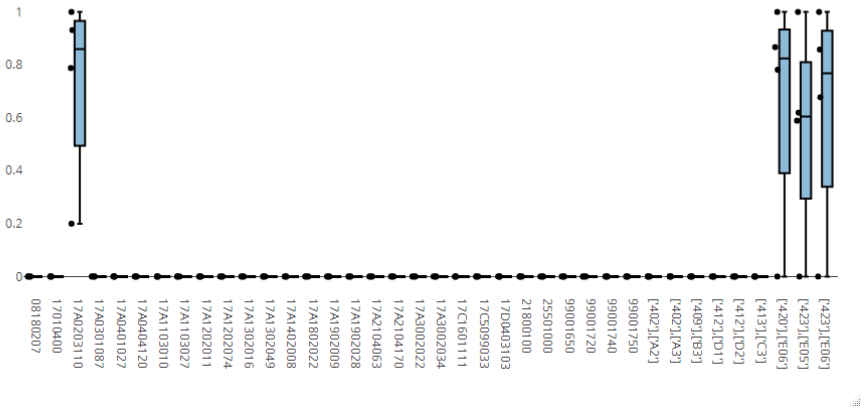

Figura 3.29: Perfil para clase 5 en año 2015
Fuente: Elaboración propia, 2019

Análisis global para año 2016, Figura 3.30, en donde se observa un total de 53 variables seleccionadas por la estrategia. Adicionalmente, se agrega una columna (costado izquierdo del gráfico) que representa al agrupamiento definido previamente.

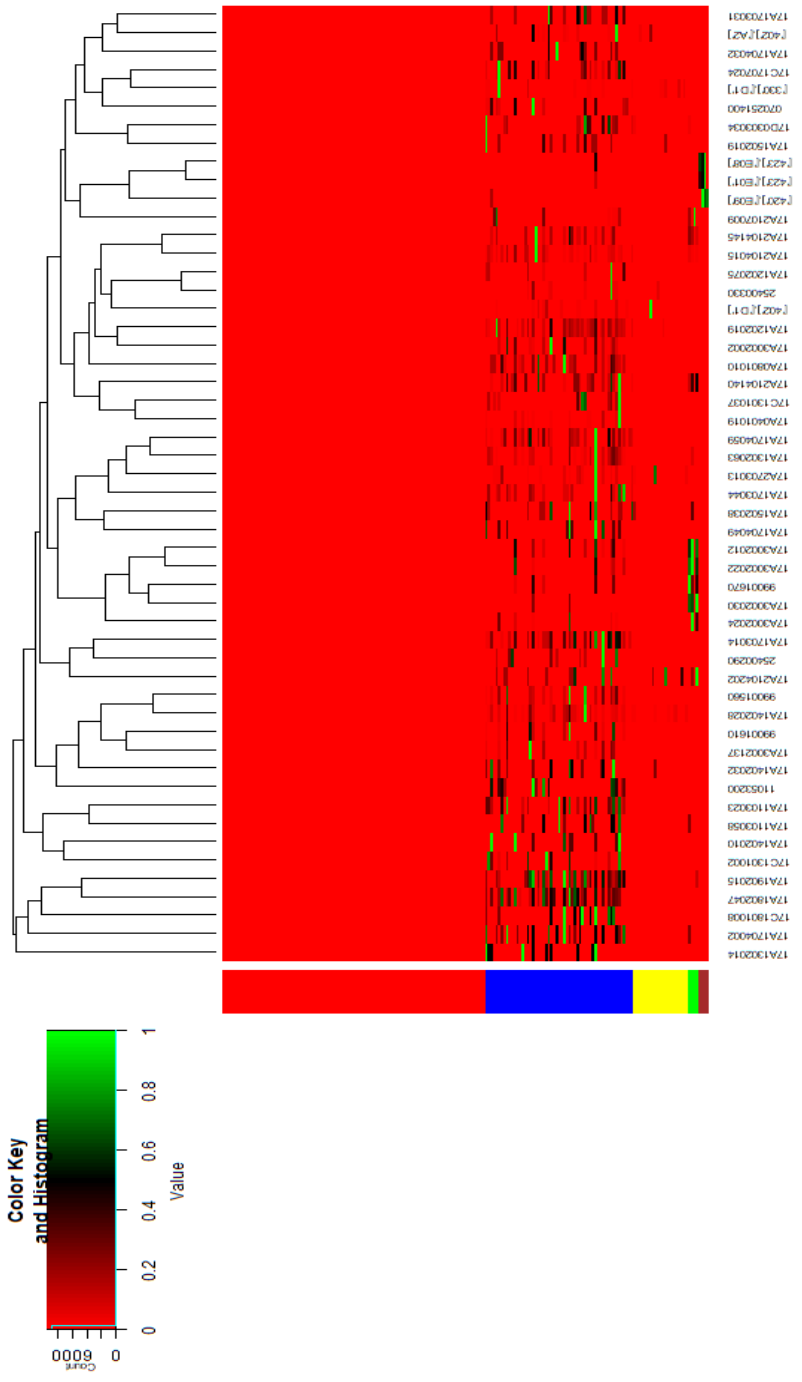

Figura 3.30: Heatmap para variables identificadas en año 2016
Fuente: Elaboración propia, 2019

Perfil para clase 1, establecimientos de baja complejidad, en año 2016 (figura 3.31 y 3.32). En donde se observa el mismo comportamiento del año 2014 y 2015, ya que las variables consideradas por la estrategia corresponden a establecimientos de mayor complejidad en su operación.

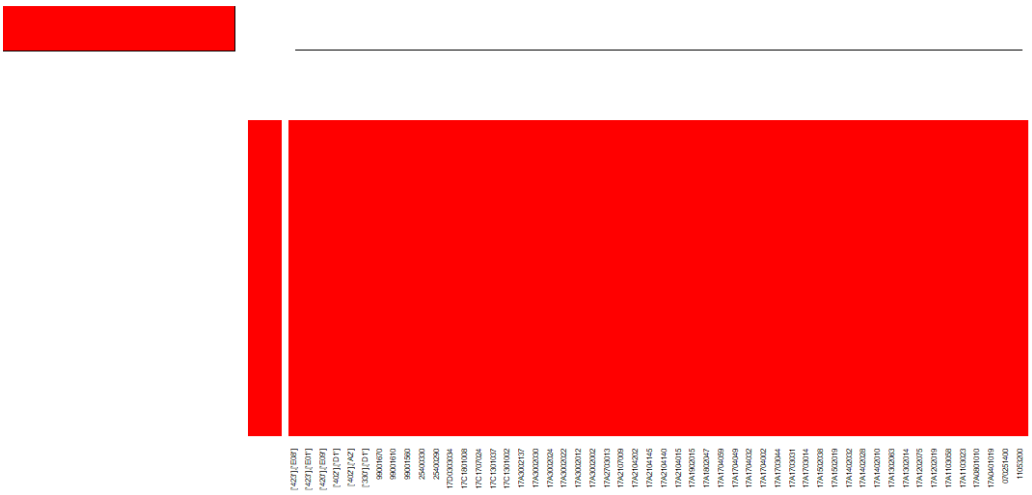

Figura 3.31: Perfil para clase 1 en año 2016
Fuente: Elaboración propia, 2019

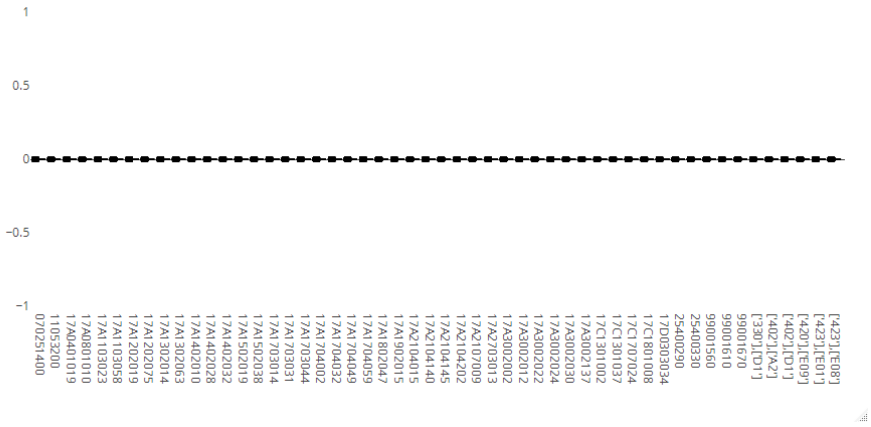

Figura 3.32: Perfil para clase 1 en año 2016
Fuente: Elaboración propia, 2019

Perfil para clase 2, establecimientos de Alta complejidad (Adultos), en año 2016 (figura 3.33 y 3.34). En donde es posible visualizar un conjunto amplio de variables identificadas con este grupo, se asocia este perfil a la cantidad de variables de alta complejidad en los establecimientos.

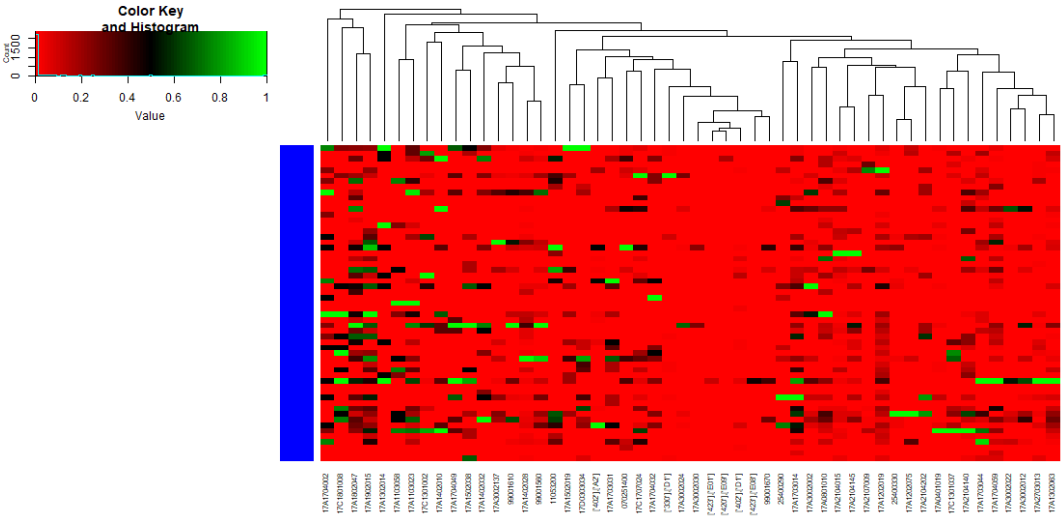

Figura 3.33: Perfil para clase 2 en año 2016
Fuente: Elaboración propia, 2019

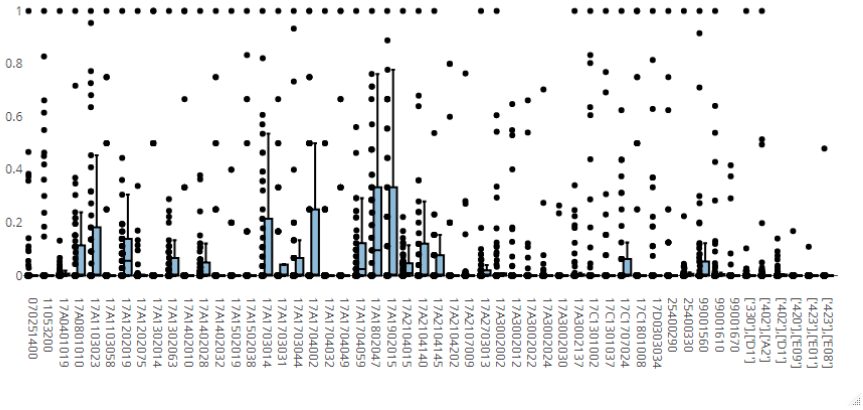

Figura 3.34: Perfil para clase 2 en año 2016
Fuente: Elaboración propia, 2019

Perfil para clase 3, establecimientos de Media complejidad, en año 2016 (figura 3.35 y 3.36). No se observa un patrón de las variables identificadas para este grupo.

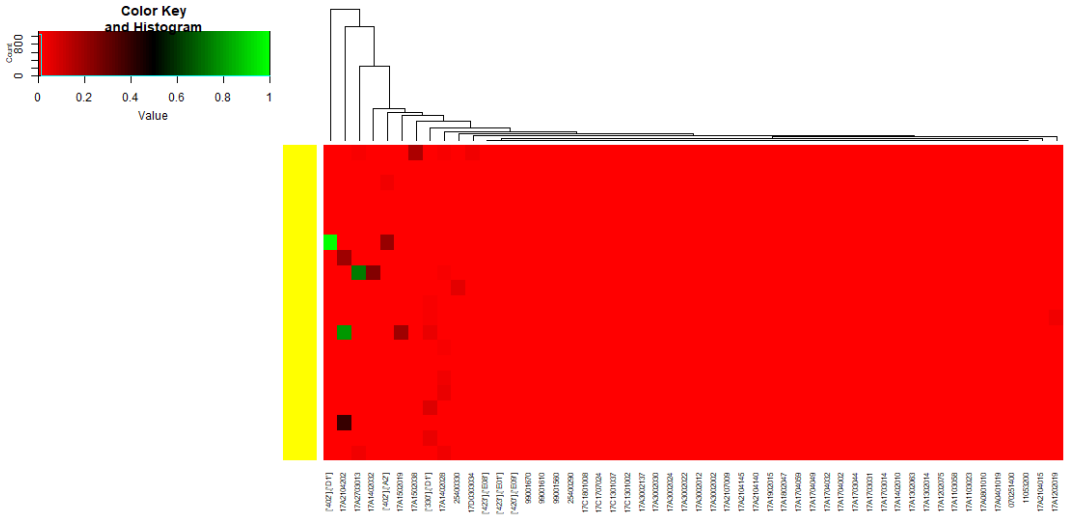

Figura 3.35: Perfil para clase 3 en año 2016
Fuente: Elaboración propia, 2019

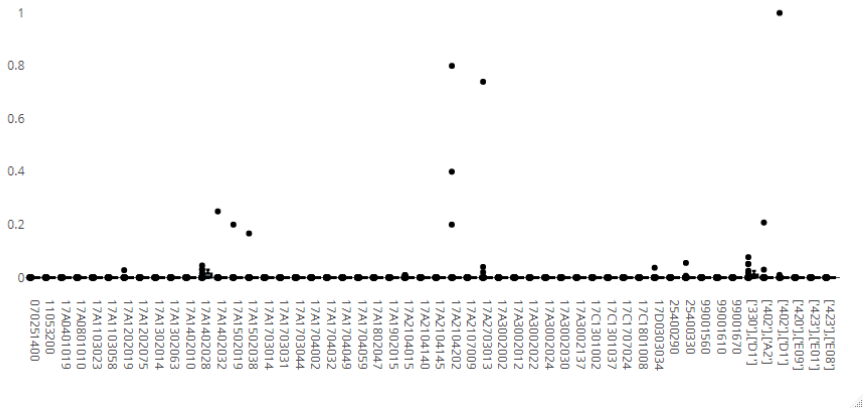

Figura 3.36: Perfil para clase 3 en año 2016
Fuente: Elaboración propia, 2019

Perfil para clase 4, establecimientos de Alta complejidad (Pediatría), en año 2016 (figura 3.37 y 3.38). Acá es posible visualizar el conjunto de variables que describe el perfil de los establecimientos de alta complejidad con foco en el área de pediatría. No se identifican variables relativas específicamente a la atención pediátrica, aunque si destaca la participación de las variables pertenecientes a la categoría de REM-BS17A como diferenciadores de estos establecimientos frente a los demás grupos.

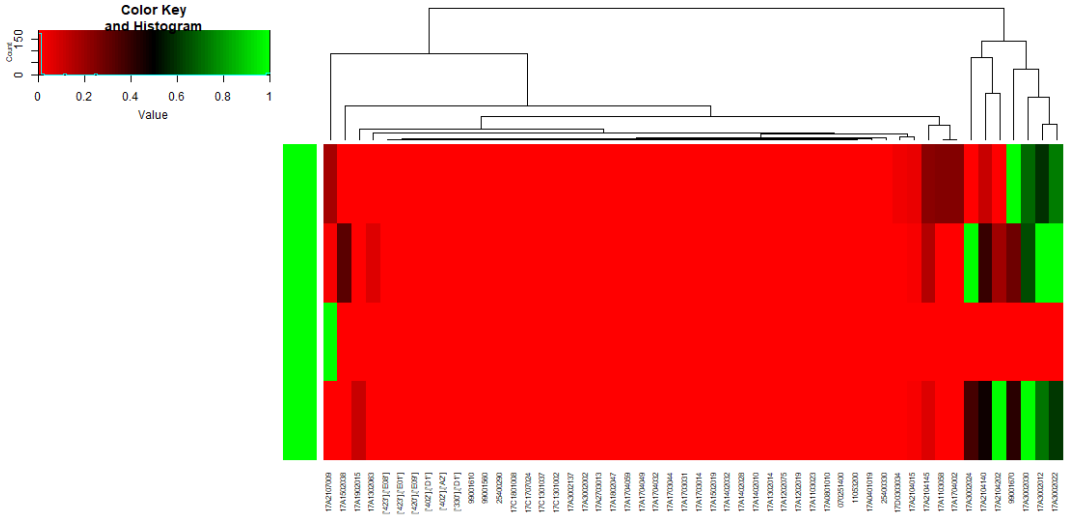

Figura 3.37: Perfil para clase 4 en año 2016
Fuente: Elaboración propia, 2019

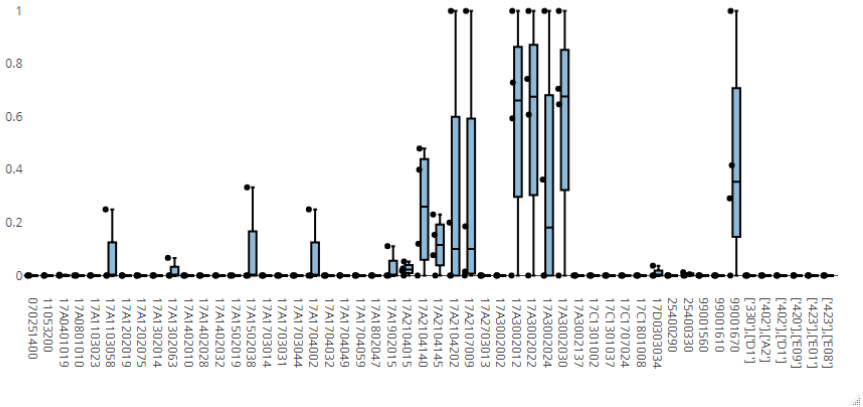

Figura 3.38: Perfil para clase 4 en año 2016
Fuente: Elaboración propia, 2019

Perfil para clase 5, establecimientos de alta complejidad (psiquiátricos), en año 2016 (figura 3.39 y 3.40). Las variables con mayor presencia en el perfil de este grupo, pertenecen de forma exclusiva a establecimientos con foco en la psiquiatría, y en particular hacen referencia a la estadía de los pacientes.

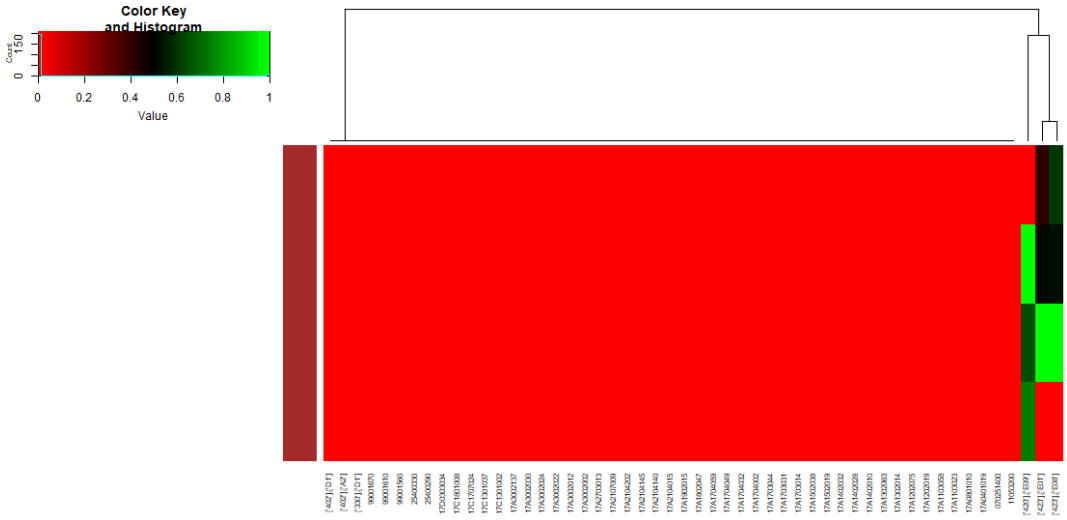

Figura 3.39: Perfil para clase 5 en año 2016
Fuente: Elaboración propia, 2019

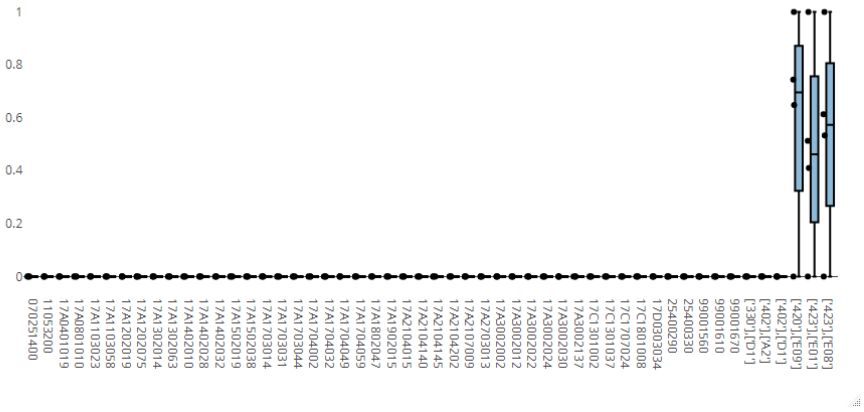

Figura 3.40: Perfil para clase 5 en año 2016
Fuente: Elaboración propia, 2019

Análisis global para año 2017, Figura 3.41, en donde se observa un total de 38 variables seleccionadas por la estrategia. Adicionalmente, se agrega una columna (costado izquierdo del gráfico) que representa al agrupamiento definido previamente.

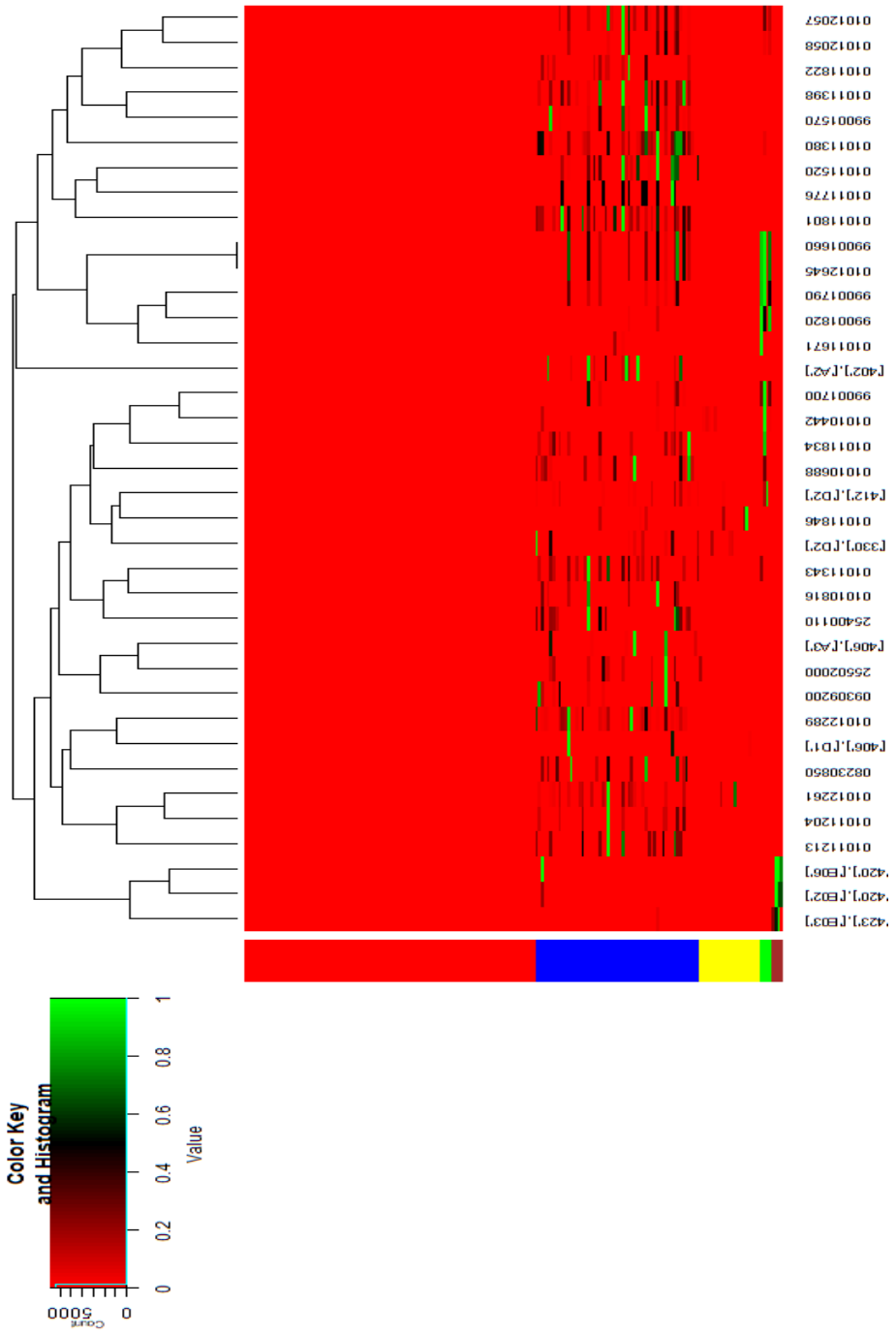

Figura 3.41: Heatmap para variables identificadas en año 2017
Fuente: Elaboración propia, 2019

Perfil para clase 1, establecimientos de baja complejidad, en año 2017 (figura 3.42 y 3.43). En donde se observa el mismo comportamiento de los años anteriores, ya que las variables consideradas por la estrategia corresponden a establecimientos de mayor complejidad en su operación, aunque con un valor correspondiente a un *ouliers*.

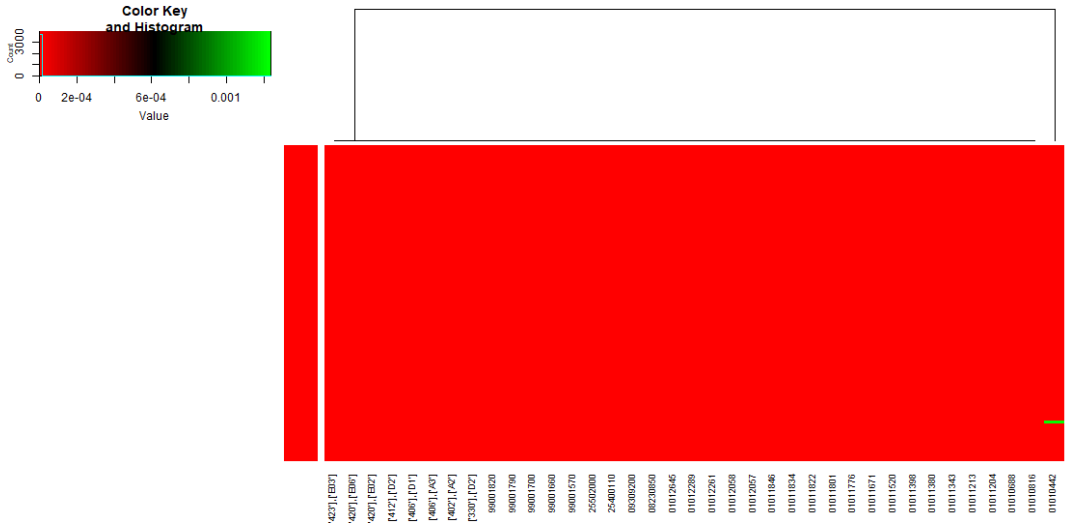

Figura 3.42: Perfil para clase 1 en año 2017
Fuente: Elaboración propia, 2019

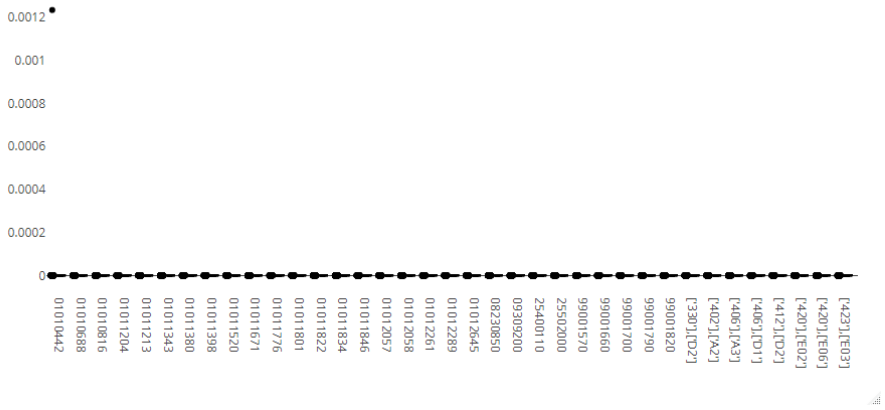

Figura 3.43: Perfil para clase 1 en año 2017
Fuente: Elaboración propia, 2019

Perfil para clase 2, establecimientos de Alta complejidad (Adultos), en año 2017 (figura 3.44 y 3.45). En donde es posible visualizar un conjunto amplio de variables identificadas con este grupo, se asocia este perfil a la cantidad de variables de alta complejidad en los establecimientos.

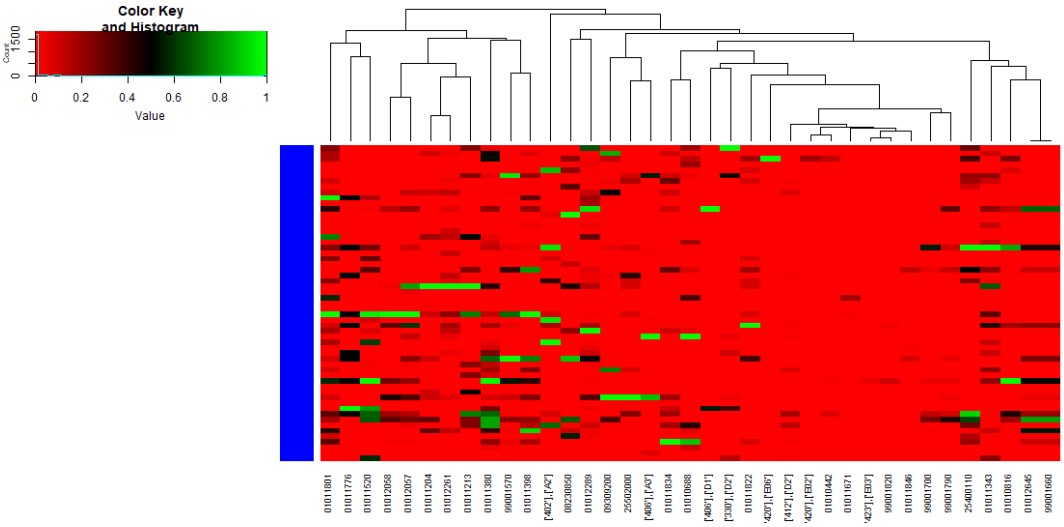

Figura 3.44: Perfil para clase 2 en año 2017
Fuente: Elaboración propia, 2019

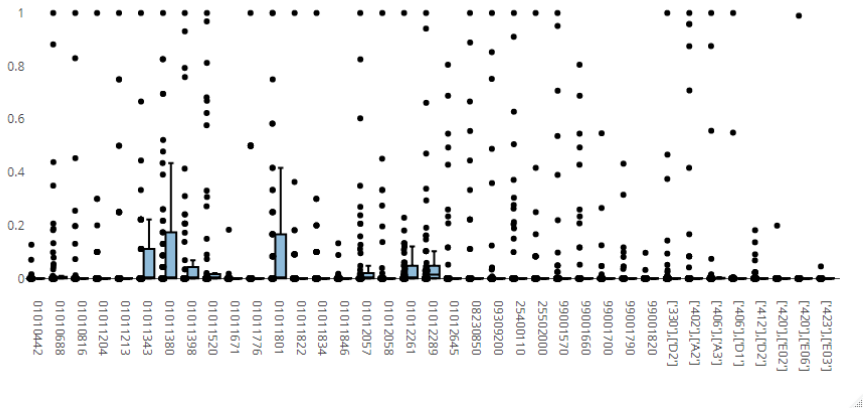

Figura 3.45: Perfil para clase 2 en año 2017
Fuente: Elaboración propia, 2019

Perfil para clase 3, establecimientos de media complejidad, en año 2017 (figura 3.46 y 3.47). No se observa un patrón de las variables identificadas para este grupo.

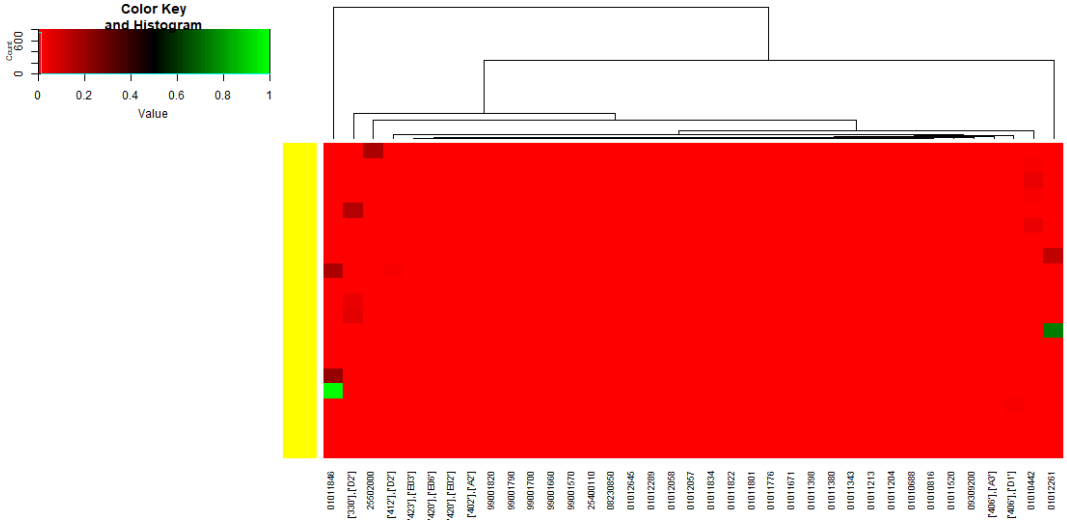

Figura 3.46: Perfil para clase 3 en año 2017
Fuente: Elaboración propia, 2019

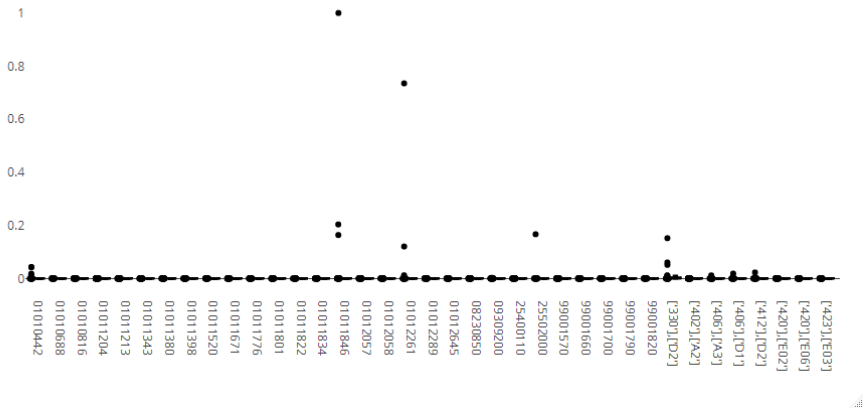

Figura 3.47: Perfil para clase 3 en año 2017
Fuente: Elaboración propia, 2019

Perfil para clase 4, establecimientos de Alta complejidad (Pediatría), en año 2017 (figura 3.48 y 3.49). Acá es posible visualizar el conjunto de variables que describe el perfil de los establecimientos de alta complejidad con foco en el área de pediatría. No se identifican variables relativas específicamente a la atención pediátrica, aunque si destaca la participación de las variables pertenecientes a la categoría de REM-BS0 como diferenciadores de estos establecimientos frente a los demás grupos.

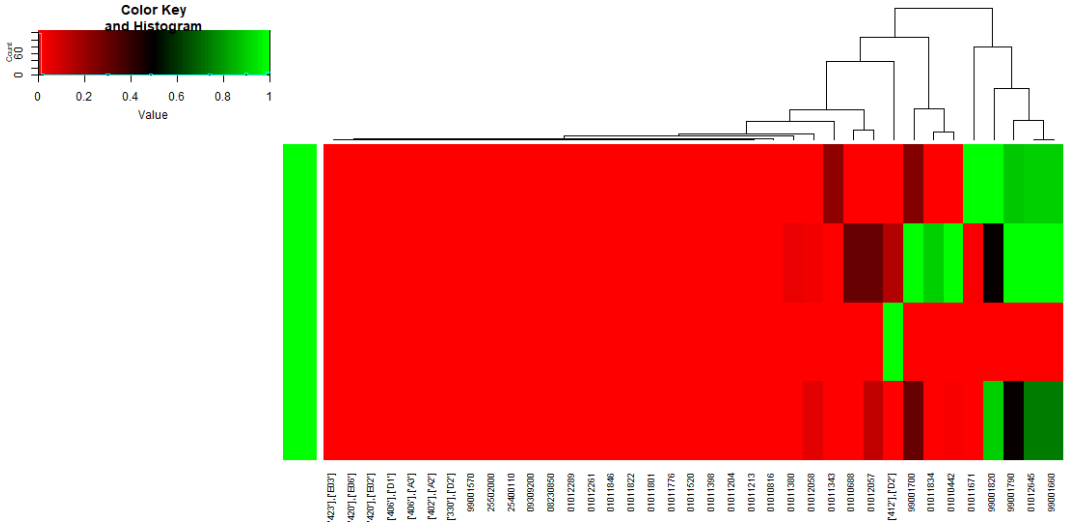

Figura 3.48: Perfil para clase 4 en año 2017
Fuente: Elaboración propia, 2019

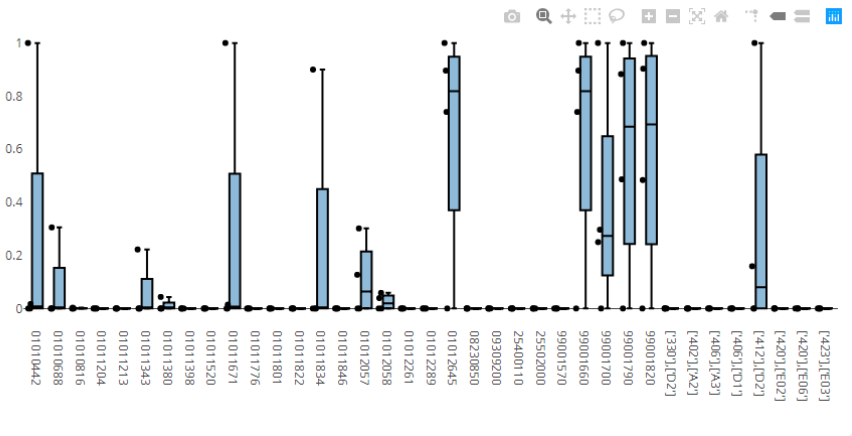

Figura 3.49: Perfil para clase 4 en año 2017
Fuente: Elaboración propia, 2019

Perfil para clase 5, establecimientos de alta complejidad (psiquiátricos), en año 2017 3.50 y 3.51). Las variables con mayor presencia en el perfil de este grupo, pertenecen de forma exclusiva a establecimientos con foco en la psiquiatría, y en particular hacen referencia a la estadía de los pacientes.

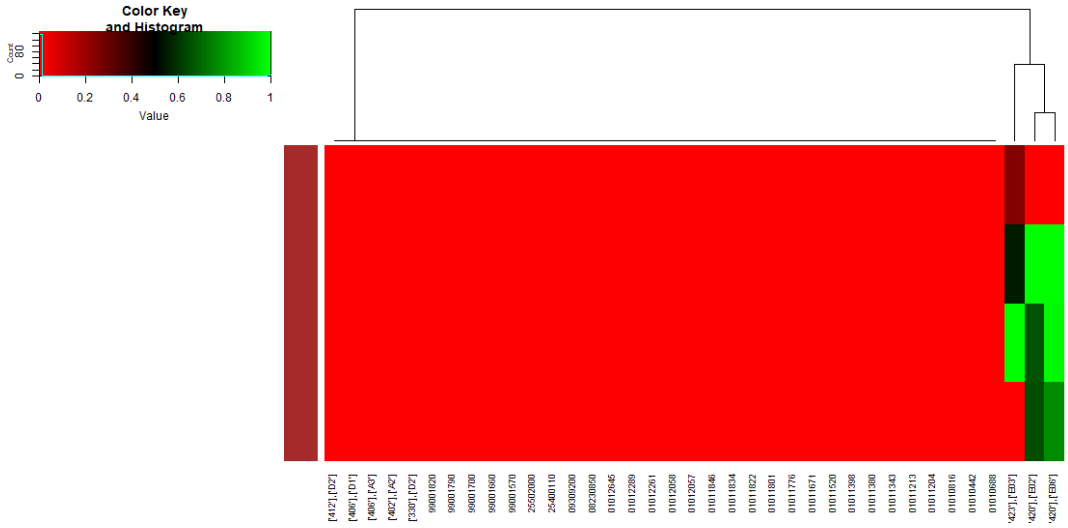

Figura 3.50: Perfil para clase 5 en año 2017
Fuente: Elaboración propia, 2019

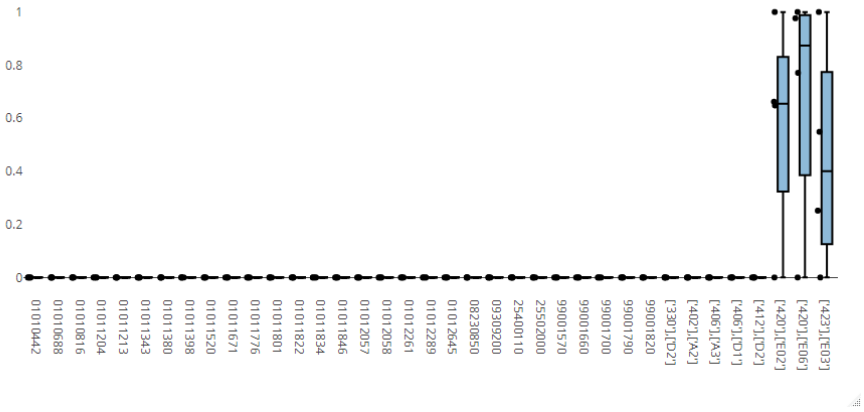

Figura 3.51: Perfil para clase 5 en año 2017
Fuente: Elaboración propia, 2019

Análisis global para año 2018, Figura 3.52, en donde se observa un total de 38 variables seleccionadas por la estrategia. Adicionalmente, se agrega una columna (costado izquierdo del gráfico) que representa al agrupamiento definido previamente.

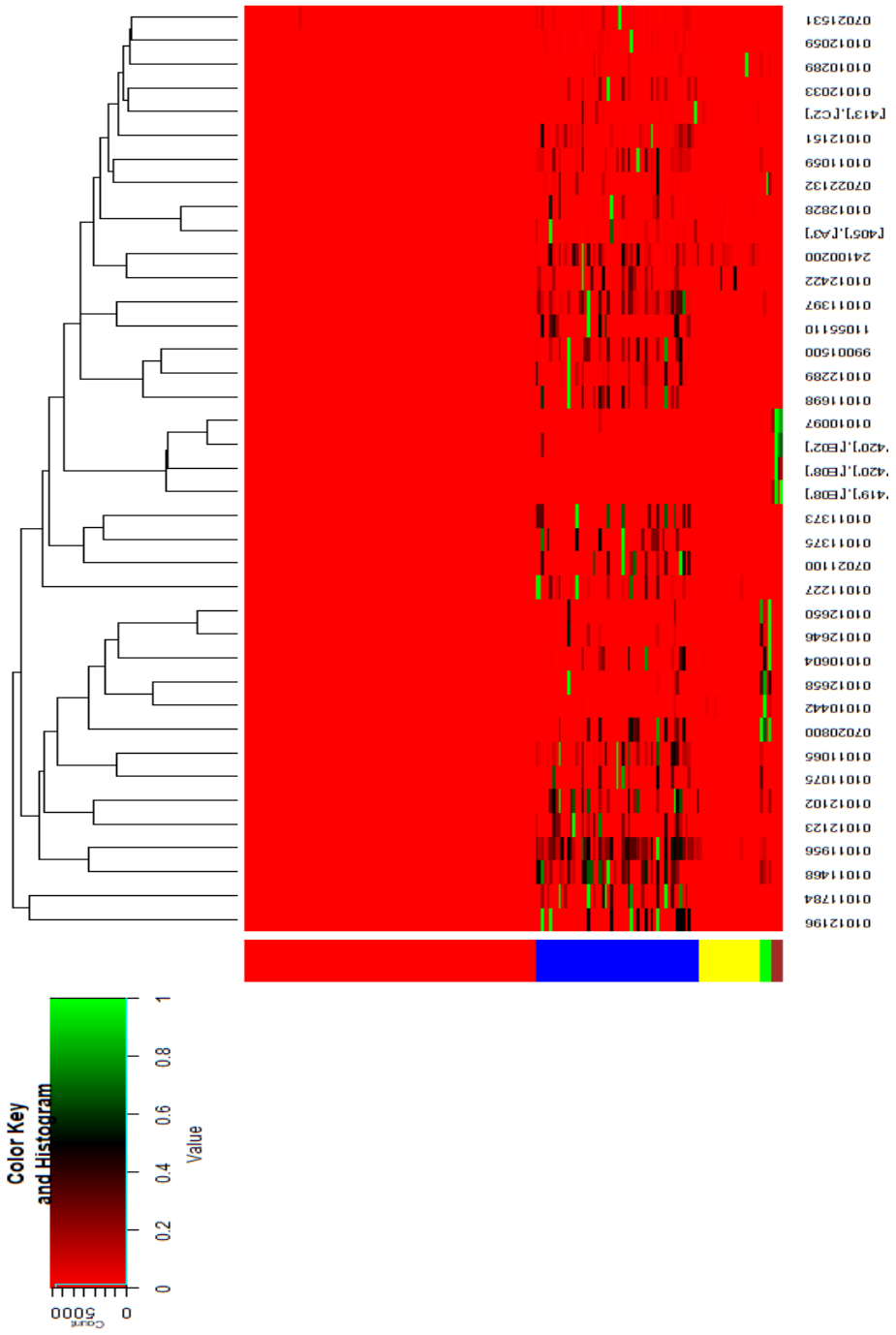

Figura 3.52: Heatmap para variables identificadas en año 2018
Fuente: Elaboración propia, 2019

Perfil para clase 1, establecimientos de baja complejidad, en año 2018 (figura 3.53 y 3.54). En donde se observa el mismo comportamiento de los años anteriores, ya que las variables consideradas por la estrategia corresponden a establecimientos de mayor complejidad en su operación, aunque con un valor correspondiente a un *valor atípico*. Estos valores, se presume que sean errores en los registros del DEIS.

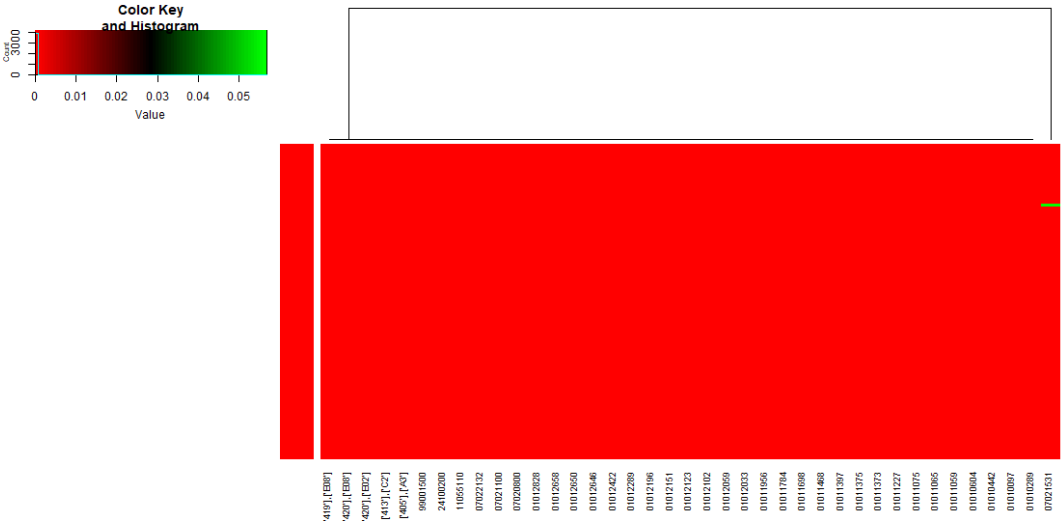

Figura 3.53: Perfil para clase 1 en año 2018
Fuente: Elaboración propia, 2019

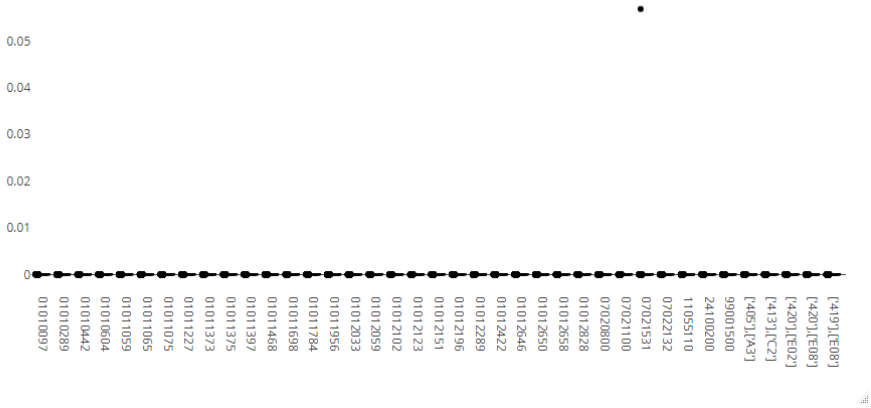

Figura 3.54: Perfil para clase 1 en año 2018
Fuente: Elaboración propia, 2019

Perfil para clase 2, establecimientos de Alta complejidad (Adultos), en año 2018 (figura 3.55 y 3.56). En donde es posible visualizar un conjunto amplio de variables identificadas con este grupo, se asocia este perfil a la cantidad de variables de alta complejidad en los establecimientos.

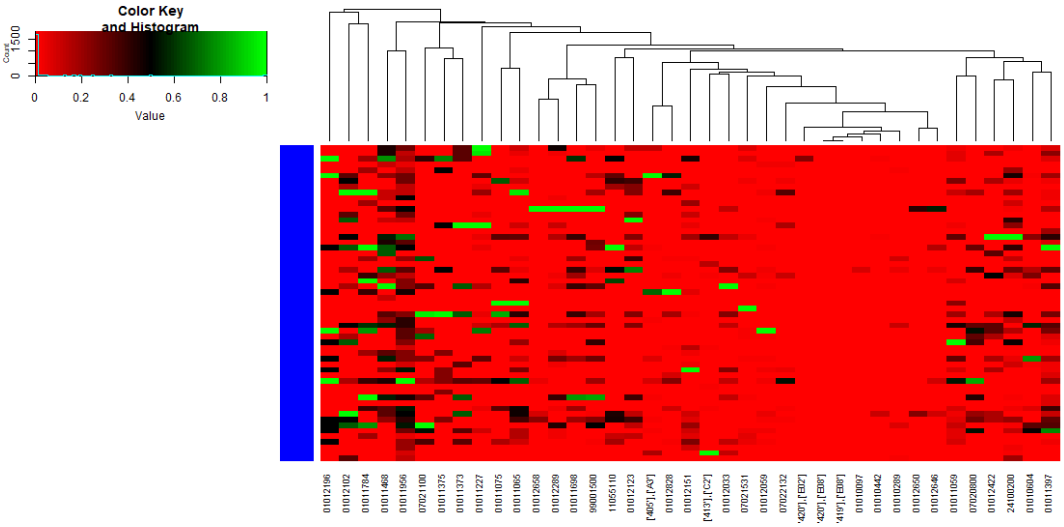

Figura 3.55: Perfil para clase 2 en año 2018
Fuente: Elaboración propia, 2019

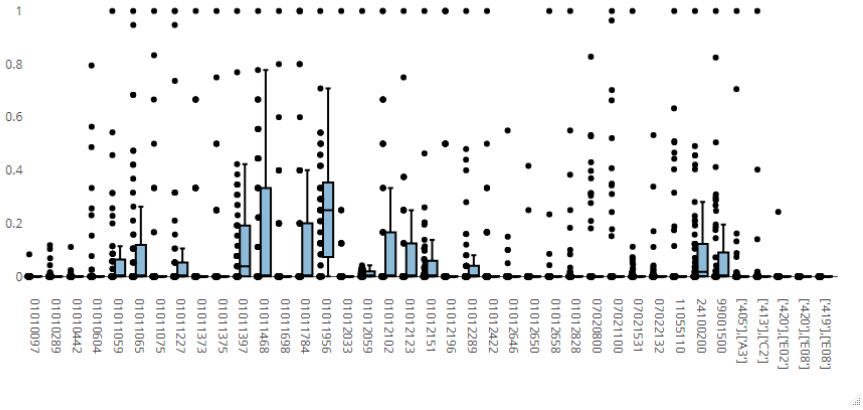

Figura 3.56: Perfil para clase 2 en año 2018
Fuente: Elaboración propia, 2019

Perfil para clase 3, establecimientos de media complejidad, en año 2018 (figura 3.57 y 3.58). No se observa un patrón de las variables identificadas para este grupo.

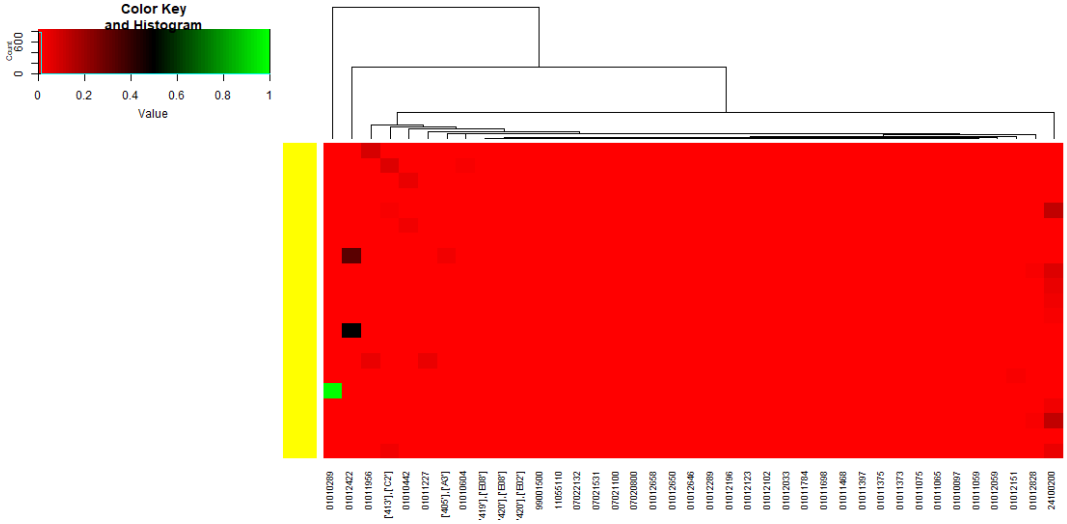

Figura 3.57: Perfil para clase 3 en año 2018
Fuente: Elaboración propia, 2019

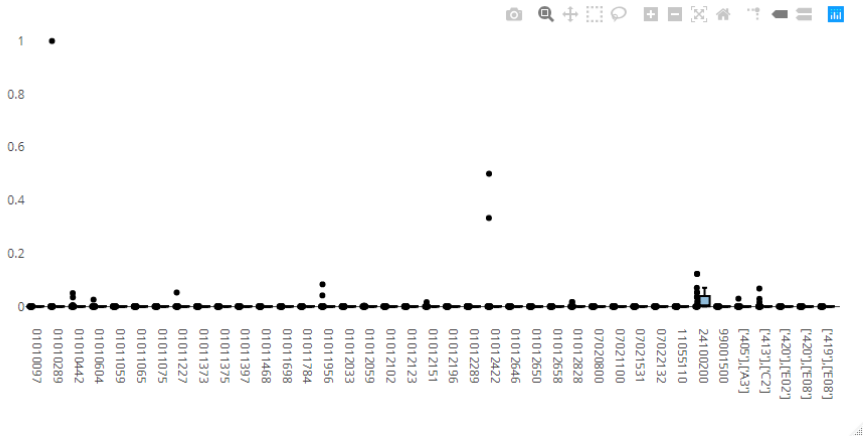

Figura 3.58: Perfil para clase 3 en año 2018
Fuente: Elaboración propia, 2019

Perfil para clase 4, establecimientos de Alta complejidad (Pediatría), en año 2018 (figura 3.59 y 3.60). Acá es posible visualizar el conjunto de variables que describe el perfil de los establecimientos de alta complejidad con foco en el área de pediatría. No se identifican variables relativas específicamente a la atención pediátrica, aunque si destaca la participación de las variables pertenecientes a la categoría de REM-BS0 como diferenciadores de estos establecimientos frente a los demás grupos.

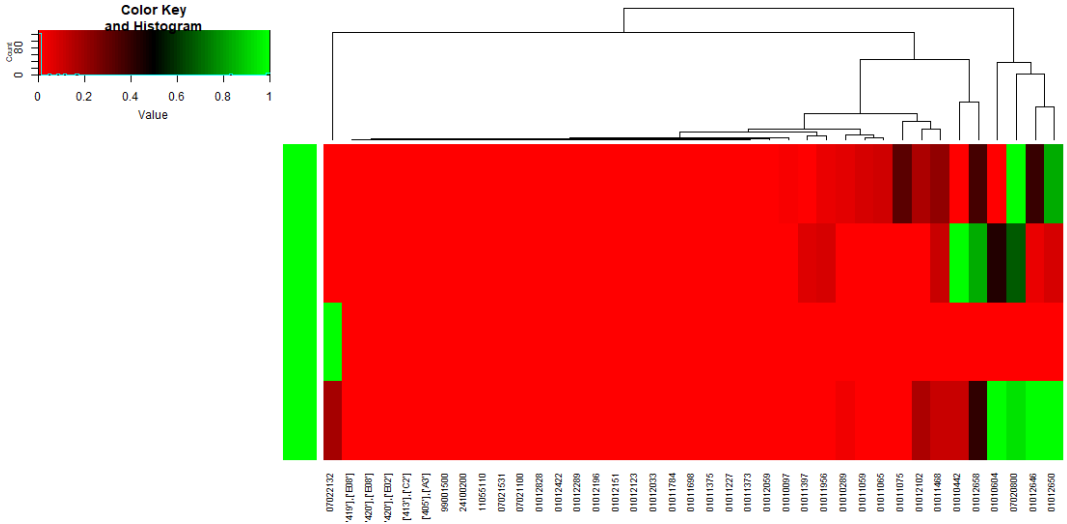

Figura 3.59: Perfil para clase 4 en año 2018
Fuente: Elaboración propia, 2019

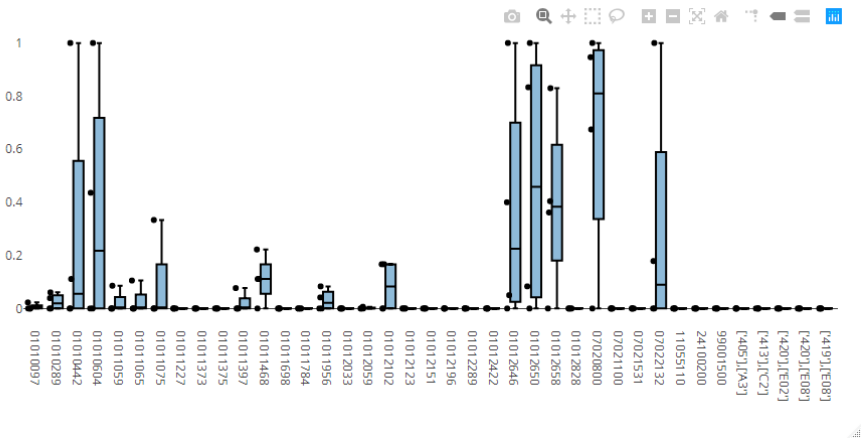

Figura 3.60: Perfil para clase 4 en año 2018
Fuente: Elaboración propia, 2019

Perfil para clase 5, establecimientos de alta complejidad (psiquiátricos), en año 2018 (figura 3.61 y 3.62). Las variables con mayor presencia en el perfil de este grupo, pertenecen de forma exclusiva a establecimientos con foco en la psiquiatría, y en particular hacen referencia a la estadía de los pacientes.

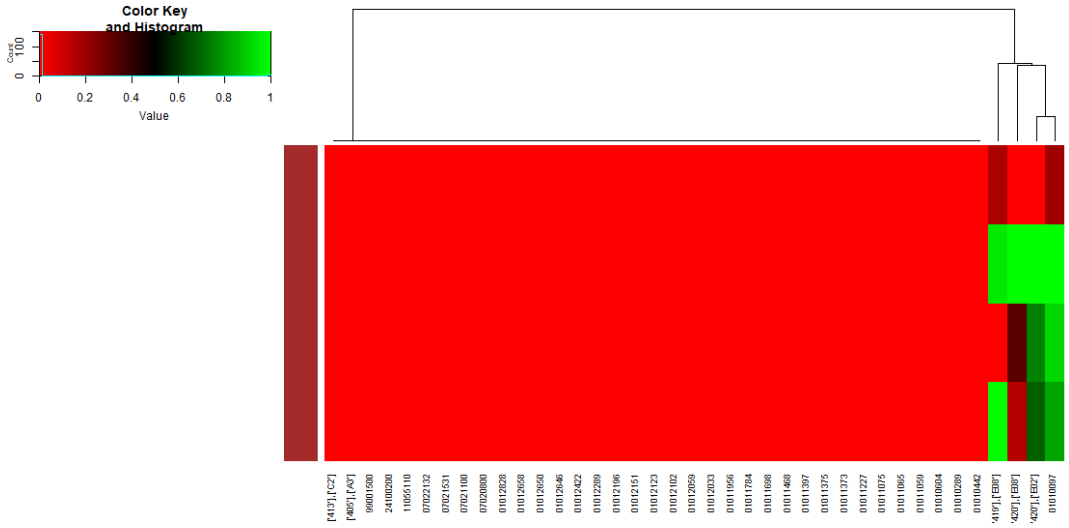

Figura 3.61: Perfil para clase 5 en año 2018
Fuente: Elaboración propia, 2019

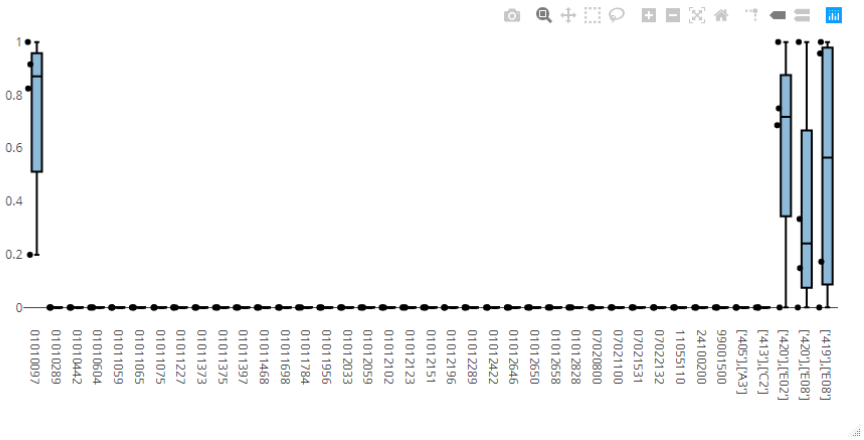

Figura 3.62: Perfil para clase 5 en año 2018
Fuente: Elaboración propia, 2019

3.5.1 Análisis de intersección de variables por año

El análisis de intersección de variables por año debe ser dividido en dos partes, esto debido al cambio general en los códigos utilizados por el DEIS en el registro del REM de la serie BS. En este aspecto se evaluará los años 2014-2015-2016 y 2017-2018.

En el primer caso, se puede identificar las variables:

- 17A0203110 - ”Día cama hospitalización integral psiquiatría mediana estadía”
- 17A0404120 - ”Ecotomografía trancraneal”
- 17A2104063 - ”Epicondilitis, trat. quir. (cualquier técnica)”
- 17C1601111 - ”Aplicación de inmunomodulares, químicos y similares hasta 10 lesiones”
- 17A3002022 - ”Histiocitosis”
- 402,A2 - ”Área Médica Adulto Cuidados Medios, Máximo riesgo - dependencia parcial”.

En donde no es posible identificar un patrón en cuanto a la repetición de ellas, y empíricamente es complejo debido al cambio que existe año a año en el registro de los datos en el DEIS. Gran parte de estas variables caracterizan al grupo de alta complejidad (Clase 1) salvo 17A0203110 que está directamente ligada a los hospitales psiquiátricos.

Por otro lado, en los años finales, se pueden identificar las siguientes variables:

- 01010442 - ”Coproparasitario seriado con técnica para Cryptosporidium sp o para Dianta-moeba fragilis”
- 01012289 - ”Amputación antebrazo”
- 420,E02 - ”Área Psiquiatría Adulto Larga estadía, Días Cama Ocupados”

Se repite el caso anterior aunque con otras variables identificadas, en donde se puede observar una variable asociada a los hospitales de tipo psiquiátrico, y otras que definen a los hospitales de alta complejidad.

Comparación con otras técnicas de la literatura

Para realizar la comparación de los resultados con otras técnicas de selección de características, se utilizó las técnicas de KNN y *Random forest*. Si bien estas técnicas poseen una

limitante en cuanto al tamaño de los grupos utilizados para establecer la relación entre variables, ya que la cantidad de combinaciones posibles es demasiada (Sección 2.2), se utiliza un muestreo aleatorio que permite evaluar parte del espacio de búsqueda de soluciones. Los resultados, que corresponden al valor máximo identificado en la ejecución de estas técnicas en 31 iteraciones, al igual que la estrategia basada en algoritmo genético, se detallan en la Tabla 3.5.

Resultados	Tipo de algoritmo utilizado		
	Random Forest	KNN	Algoritmo genético
Métricas de Calidad	Máximo		
	2014		
Índice adjusted rand	0.815	0.804	**0.944**
Índice Fowlkes-Mallows	0.870	0.900	**0.966**
Índice Jaccard	0.784	0.760	**0.933**
	2015		
Índice adjusted rand	0.768	0.775	**0.960**
Índice Fowlkes-Mallows	0.858	0.858	**0.977**
Índice Jaccard	0.808	0.781	**0.955**
	2016		
Índice adjusted rand	0.733	0.740	**0.953**
Índice Fowlkes-Mallows	0.821	0.841	**0.969**
Índice Jaccard	0.689	0.698	**0.938**
	2017		
Índice adjusted rand	0.683	0.681	**0.915**
Índice Fowlkes-Mallows	0.814	0.824	**0.948**
Índice Jaccard	0.680	0.701	**0.901**
	2018		
Índice adjusted rand	0.689	0.695	**0.903**
Índice Fowlkes-Mallows	0.828	0.809	**0.949**
Índice Jaccard	0.713	0.693	**0.909**

Tabla 3.5: Comparación con técnicas de selección de características, por años de estudio. Fuente: Elaboración propia, 2019.

Como se puede apreciar en los resultados, las técnicas tradicionales de selección de características fueron aplicadas obteniendo resultados muy bajos en relación a los obtenidos por la técnica de algoritmo genético. Este resultado tiene relación con lo esperado, debido a la selección aleatoria y combinación de grupos de características que realizan las técnicas implementadas, en este caso *Random forest* y KNN, lo que les da un enfoque aleatorio en la búsqueda de sus soluciones. La aplicación de un direccionamiento mediante la función de *fitness* en el algoritmo genético permite orientar la búsqueda de mejores soluciones.

CAPÍTULO 4. CONCLUSIONES Y TRABAJOS FUTUROS

En esta sección se presentan las conclusiones obtenidas en el estudio, realizadas en base a los objetivos planteados de manera inicial y también respecto a parámetros específicos relacionados a la implementación de un algoritmo genético y técnicas de clasificación.

4.1 CONCLUSIONES

En este trabajo se propuso una estrategia para identificar las variables de la casuística hospitalaria que permiten la clasificación definida por el MINSAL, esto teniendo como base de datos disponible los registros del DEIS. Considerando la gran cantidad de datos y variables involucradas en este problema, se implementó una estrategia de heurística basada en algoritmo genético, la cual realiza una búsqueda sobre un espacio de soluciones más acotado y orientado por la función *fitness* implementada. Utilizar la metodología KDD para la implementación del algoritmo genético fue beneficioso debido a su naturaleza iterativa, corrigiendo errores o considerando nuevos aspectos que mejoraban los resultados en la búsqueda de soluciones.

4.1.1 Parametrización

La utilización de *irace* como herramienta de parametrización permitió encontrar los valores iniciales de los parámetros, y su definición descartó problemáticas que apriori se veian dificiles de resolver, particularmente el caso de las clases desbalanceadas, en donde su resolución mediante *SMOTE* era perjudicial para la estrategia y agregaba ruido a las soluciones encontradas.

Respecto a los resultados obtenidos en la caracterización de la clasificación de establecimientos de salud, si bien los resultados no hablan de algunas variables que año a año se repitan en este proceso, si se puede encontrar un patrón general respecto a la caracterización de los grupos por separado, en donde la clase 1, correspondiente a establecimientos de baja complejidad, no tiene participación en las variables encontradas, asumiendo que en su funcionamiento se atienden patologías comunes en todos los tipos de establecimientos. Los establecimientos de alta complejidad adultos, posee un número de variables asociadas a su funcionamiento, que son observadas en otros tipos de establecimientos de alta complejidad pero en menor magnitud, como es el caso de los establecimiento con su foco en el área de pediatría. Establecimientos de mediana complejidad poseen determinadas variables que los definen y

algunas fueron identificadas por la estrategia utilizada. Finalmente los establecimientos de alta complejidad con foco en el área de psiquiatría poseen variables claras en su funcionamiento, que fueron identificadas de forma clara por la estrategia.

4.1.2 Objetivos

Referente a los objetivos planteados (Sección 1.4) para el desarrollo de la investigación es posible concluir lo siguiente:

Un modelo de datos bien construido disminuye en gran forma los tiempos de respuesta en el acceso a los datos, particularmente útil en el desarrollo de técnicas de carácter iterativo y posterior análisis de resultados obtenidos.

Muchos datos fueron eliminados de forma previa a la aplicación del algoritmo genético por ser considerados como datos que no entregaban información adicional al estudio. En términos del tiempo de ejecución asociado al procesamiento de los datos se ve una mejora considerable y respecto al ruido que estas variables ingresaban a las primeras soluciones.

La heurística basada en algoritmo genético fue diseñada e implementada para la identificación de las variables que caracterizan la clasificación definida por el MINSAL en donde se evaluó distintas soluciones y conforme a la parametrización utilizada, se identificó soluciones con altos valores en cuanto a las métricas de calidad definidas.

Distintas técnicas tradicionales en la selección de características fueron implementadas con ciertas excepciones, recordando que la cantidad de variables consideradas en el estudio impide la evaluación de todas las combinaciones de variables posibles, con resultados que no fueron superiores a los encontrados por la heurística basada en algoritmo genético, que en parte se debe al uso de una función que guía la búsqueda de las mejores soluciones.

Las variables identificadas en las mejores soluciones encontradas por el algoritmo genético se relacionan, en parte, a las características de cada grupo evaluado. Encontrando tipos de variables y un comportamiento similiar en los distintos años de estudio.

Respecto a la pregunta de investigación, fue posible identificar un conjunto de variables asociados a la casuística hospitalaria, las cuales se relacionan con la clasificación propuesta por el MINSAL. Distintos subconjuntos fueron identificados año a año considerando la variación de los conjuntos de datos registrados. De igual forma, cada conjunto de variables identificado posee variables que se identifican directamente con los tipos de establecimientos de salud definidos.

Finalmente, si bien se ha identificado un conjunto de variables correspondientes a la casuística hospitalaria que caracterizan la clasificación definida por el MINSAL, corresponde a un

conjunto muy acotado de variables dentro del universo de características, identificando además características específicas de cada clase, en algunos casos que solo se registran en tipos de establecimientos, y en particular los establecimientos de baja complejidad con nula participación en la caracterización.

4.2 TRABAJOS FUTUROS

El desarrollo de esta investigación permitió identificar trabajos futuros que ayuden a la gestión del MINSAL, respecto a la clasificación de sus establecimientos de salud, los cuales son:

- Identificación de una mejor clasificación de los establecimientos: una mejor clasificación de los establecimientos de salud permite generar características común entre ellos que estén realmente ligadas a su funcionamiento. En aspectos referentes a la gestión y también considerando estudios de eficiencia técnica hospitalaria, permitiría establecer parámetros de comparación más justos entre establecimientos de salud.
- Homologación de base de datos por año: una de las dificultades en la investigación, es la variabilidad en el registros de las variables realizado por el DEIS, y si bien los códigos en si cambian año a año, existe una relación que , de conocerse, permitiría realizar análisis más globales.
- Identificar relaciones con GRD: establecer la relación que existe entre las variables identificadas con lo establecido por los GRD en los establecimientos de alta complejidad en donde se encuentra implementado.
- Selección de características sobre datos históricos: generar un análisis de selección de características en todos los años que cuenten con información disponible, de esta forma establecer patrones globales.

GLOSARIO

CLARA: del inglés, clustering LAR applications.
CUDYR: categorización de usuarios por dependencia y riesgo.
DEIS: departamento de estadísticas e información de salud.
GRD: sistema de grupos relacionados por el diagnóstico.
KDD: del inglés, knowledge discoveryand data mining.
KNN: del inglés, k-nearest neighbor.
MINSAL: ministerio de salud.
PAM: del inglés, partitioning around medoids.
REM: registro estadístico mensual.

REFERENCIAS BIBLIOGRÁFICAS

Barahona-Urbina, P. (2011). Análisis de eficiencia hospitalaria en chile. *Anales de la Facultad de Medicina*.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2-3), 191–203.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brouwer, R. K. (2009). Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. *J. Intell. Inf. Syst.*, *1*(3).

Bräsel, H., Herms, A., Mörig, M., Tautenhahn, T., Tusch, J., & Werner, F. (2008). Heuristic constructive algorithms for open shop scheduling to minimize mean flow time. *European Journal of Operational Research*, *189*(3), 856–870.

Cant-Paz, E., & Goldberg, D. E. (2003). Are multiple runs of genetic algorithms better than one? In *Genetic and Evolutionary Computation Conference*, (pp. 801–812). Springer.

Castro, R. (2007). Midiendo la eficiencia de los hospitales públicos en chile.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, Y., Miao, D., & Wang, R. (2010). A rough set approach to feature selection based on ant colony optimization. *Pattern Recognition Letters*, *31*(3), 226–233.

Coll, V., & Blasco, O. (2006). *Evaluación de la eficiencia mediante el análisis envolvente de datos: introducción a los modelos básicos*. Eumed.net.
URL `https://books.google.cl/books?id=LxCXnQAACAAJ`

Cortes-Martínez, A. E. (2010). La economía de la salud en el hospital. *Revista Gerencia y Políticas de Salud*, *9*, 138 – 149.
URL `http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-70272010000200010&nrm=iso`

Doak, J. (1992). *An Evaluation of Feature Selection Methods and Their Application to Computer Security*. University of California, Computer Science.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*, 37–54.

FONASA (2015). https://www.fonasa.cl.
URL `https://www.fonasa.cl/sites/fonasa/prestadores/convenios/red-publica/prestadores-publicos/acuerdos-grd`

foundation, R. (2019). https://www.r-project.org/.
URL `https://www.r-project.org/`

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*.

García G, M. A., & Castillo F, L. (2000). Categorización de usuarios: una herramienta para evaluar las cargas de trabajo de enfermería. *Revista médica de Chile*, *128*, 177 – 183.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.

Gómez, M. M. (2014). Las metaheurísticas: tendencias actuales y su aplicabilidad en la ergonomía. *Ingeniería Industrial. Actualidad y Nuevas Tendencias*, *4*(12), 108–120.

Hornbrook, M. C. (1982). Review article : Hospital case mix: Its definition, measurement and use: Part i. the conceptual framework. *Medical Care Review*, *39*(1), 1–43.
URL https://doi.org/10.1177/107755878203900101

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1).

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kira, K., Rendell, L. A., et al. (1992). The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, vol. 2, (pp. 129–134).

Kohl, S., Schoenfelder, J., Fügener, A., & O. Brunner, J. (2018). The use of data envelopment analysis (dea) in healthcare with a focus on hospitals. *Health Care Management Science*, *22*.

Krig, S. (2016). *Computer vision metrics*. Springer.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Li, S., Harner, E. J., & Adjeroh, D. A. (2011). Random knn feature selection-a fast and stable alternative to random forests. *BMC bioinformatics*, *12*(1), 450.

Lian, L., & Castelain, E. (2009). *A decomposition-based heuristic approach to solve general delivery problem*.

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., & Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. *Operations Research Perspectives*, *3*, 43 – 58.

Masilamani, A., , A., & Iyenger, N. C. S. N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm. *International Journal of Engineering Science and Technology*, *2*.

McShan, D. C., Rao, S., & Shah, I. (2003). Pathminer: predicting metabolic pathways by heuristic search. *Bioinformatics*, *19*(13), 1692–1698.

MINSAL (2013). Criterios de clasificación según nivel de complejidad de establecimientos hospitalarios.

Murty, M., & Susheela Devi, V. (2011). *Pattern recognition. An algorithmic approach*.

Ng, R., & Han, J. (1994). :" efficient and effective clustering methods for spatial data mining", proc. 20th int. conf. on very large data bases, santiago, chile, morgan kaufmann publishers.

Noman, N., & Iba, H. (2007). Inferring gene regulatory networks using differential evolution with local search heuristics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *4*(4), 634–647.

OCDE (2018). https://data.oecd.org/healthres/health-spending.htm.

Pohlert, T. (2018). https://cran.r-project.org/web/packages/pmcmrplus/.
URL https://cran.r-project.org/web/packages/PMCMRplus/PMCMRplus.pdf

Romanski, P., Kotthoff, L., & Kotthoff, M. L. (2018). Package 'fselector'. *Repository CRAN*.

Santelices, E., Ormeño, H., Delgado, M., Lui, C., Valdés, R., & Durán, L. (2013). Análisis de la eficiencia técnica hospitalaria 2011. *Revista médica de Chile*, *141*(3), 332–337.

Thomas W. Yee (auth.), P. D. B. R. e., Professor Dr. Wolfgang Härdle (2002). *COMPSTAT 2002 - Proceedings in Computational Statistics: 15th Symposium held in Berlin, Germany, 2002*. Physica-Verlag Heidelberg, 1 ed.

Vega M., J. (2015). Mejorando la salud hospitalaria: Alternativas para el financiamiento y la gestión.

Villalobos-Cid, M., Chacón, M., Zitko, P., & Inostroza-Ponta, M. (2016). A new strategy to evaluate technical efficiency in hospitals using homogeneous groups of casemix. *Journal of Medical Systems*.
URL https://doi.org/10.1007/s10916-016-0458-9

Villalobos-Cid, M., Dorn, M., Ligabue-Braun, R., & Inostroza-Ponta, M. (2018). A memetic algorithm based on an nsga-ii scheme for phylogenetic tree inference. *IEEE Transactions on Evolutionary Computation*.

Yang, X.-S., & Koziel, S. (2011). *Computational Optimization, Methods and Algorithms*, vol. 356.

Yoo, H., & Lafortune, S. (1989). An intelligent search method for query optimization by semijoins. *IEEE Transactions on Knowledge and Data Engineering*, *1*(2), 226–237.

Zhang, H., & Sun, G. (2002). Feature selection using tabu search method. *Pattern recognition*, *35*(3), 701–711.

ANEXO A. DESCRIPCIÓN REM

Descripción de la categorización de los registros estadísticos mensuales (REM).

REM	Descripción
REM-A01	CONTROLES DE SALUD
REM-A02	EXAMEN DE MEDICINA PREVENTIVA EN MAYORES DE 15 AÑOS
REM-A03	APLICACIÓN Y RESULTADOS DE ESCALAS DE EVALUACIÓN
REM-A04	CONSULTAS
REM-A05	INGRESOS Y EGRESOS POR CONDICIÓN Y PROBLEMAS DE SALUD
REM-A06	PROGRAMA DE SALUD MENTAL ATENCIÓN PRIMARIA Y ESPECIALIDADES
REM-A07	ATENCIÓN DE ESPECIALIDADES
REM-A08	ATENCIÓN DE URGENCIA
REM-A09	ATENCIÓN DE SALUD ODONTOLÓGICA EN APS Y ESPECIALIDADES
REM-A11	EXÁMENES DE PESQUISA DE ENFERMEDADES TRASMISIBLES
REM-A19a	ACTIVIDADES DE PROMOCIÓN Y PREVENCIÓN DE LA SALUD
REM-A19b	ACTIVIDADES DE PARTICIPACIÓN SOCIAL

Tabla A.1: Descripción REM parte 1

REM	Descripción
REM-A21	PABELLONES QUIRÚRGICOS Y OTROS RECURSOS HOSPITALARIOS
REM-A23	SALAS: IRA, ERA Y MIXTAS EN APS
REM-A24	ATENCIÓN EN MATERNIDAD
REM-A25	SERVICIOS DE SANGRE
REM-A26	ACTIVIDADES EN DOMICILIO Y OTROS ESPACIOS
REM-A27	EDUCACIÓN PARA LA SALUD
REM-A28	REHABILITACIÓN
REM-A29	PROGRAMA DE IMÁGENES DIAGNÓSTICAS Y/O RESOLUTIVIDAD EN ATENCION PRIMARIA
REM-A30	ATENCIONES POR TELEMEDICINA EN LA RED ASISTENCIAL
REM-A31	MEDICINA COMPLEMENTARIA
REM-BS0	FACTURACION PAGO POR PRESTACIONES INSTITUCIONALES
REM-BS17	ACTIVIDADES DE APOYO DIAGNOSTICO Y TERAPEUTICO
REM-BS17A	LIBRO DE PRESTACIONES DE APOYO DIAGNÓSTICO Y TERAPÉUTICO (USO EXCLUSIVO ESTABLECIMIENTOS DEL SERVICIO DE SALUD Y DELEGADOS)
REM-BS17C	LIBRO DE PRESTACIONES DE APOYO DIAGNÓSTICO Y TERAPÉUTICO
REM-BS17D	COMPRA DE SERVICIO DE PRESTACIONES DE APOYO DIAGNÓSTICO Y TERAPÉUTICO REALIZADAS
REM-D15	PROGRAMA NACIONAL DE ALIMENTACION COMPLEMENTARIA (P.N.A.C.)
REM-D16	PROGRAMA NACIONAL DE ALIMENTACIÓN COMPLEMENTARIA DEL ADULTO MAYOR (P.A.C.A.M.)

Tabla A.2: Descripción REM parte 2

ANEXO B. DESCRIPCIÓN DE VARIABLES SELECCIONADAS

Variables asociadas al REM, 2014		
Categoría	ID	Descripción
REM-A05	05810310	INGRESOS Y EGRESOS A PROGRAMA INFECCIÓN POR TRANSMISIÓN SEXUAL (Uso de establecimientos que realizan atención de ITS) CHLAMYDIAS
REM-A07	07030400	CONSULTAS INFECCIÓN TRANSMISIÓN SEXUAL (ITS) Y CONTROLES DE SALUD SEXUAL EN EL NIVEL SECUNDARIO (Incluidos en Sección B)-CONTROL VIH SIN TAR-MATRONA
REM-A08	08180207	ATENCIONES REALIZADAS EN UEH DE HOSPITALES DE MEDIANA COMPLEJIDAD. MATRONA /ÓN
REM-A09	09204939	ACTIVIDADES EN ATENCIÓN DE ESPECIALIDADES Ortopedia prequirúrgica, actividad (fisura labiopalatina)
REM-BS17C	17C1601111	Aplicación de inmunomodulares, quimicos y similares hasta 10 lesiones
	17C5099037	Dental Endodoncia: Desobturacion de conductos
REM-BS17D	17D0301093	Resistencia Proteína C
REM-A25	25400420	REACCIONES ADVERSAS POR ACTO TRANSFUSIONAL (UMT-BS), SOBRECARGA CIRCULATORIA
REM-BS0	99000850	Día cama hospitalización integral psiquiatría mediana estadía
	99001120	Ecocardiograma bidimensional (incluye registro modo M, papel fotosensible y fotografía), en adultos o niños (proc. aut.)
	99001190	Aortografía, en adultos o niños (Incluye proc. rad.)
	99001510	Linfoma No Hodgkin no agresivo
	99001750	Recaída tumores sólidos
	99003000	Ca mama etapa IV

Tabla B.1: Variables seleccionadas para el año 2014, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2014		
Categoría	ID	Descripción
REM-BS17A	17A0203110	Día cama hospitalización integral psiquiatría mediana estadía
	17A0303035	Cortisol libre urinario
	17A0305025	Inmunofijación de inmunoglobulina, c/u
	17A0401018	Enema baritada del colon (incluye llene y control postvaciamiento; 8-10 exp.)
	17A0404120	Ecotomografía trancraneal
	17A0405003	Orbitas
	17A1103005	Craneoplastía con prótesis (no incluye valor de la prótesis)
	17A1202056	Desprendimiento retinal, cirugía convencional (exoimplantes)
	17A1402013	Parotidectomía- Total ampliada (incluye músculos, ganglios, articulaciones y rama vertical de la mandíbula)
	17A1402021	Fístula salival, trat. quir.
	17A1502054	Con resección ósea c/s colgajo de rotación
	17A1703043	Radical clásica o modificada de cuello
	17A1704015	Timectomía:- Vía torácica medioesternal
	17A1704020	Hernioplastía diafragmática por vía torácica c/ prótesis (no incluye valor de la prótesis)
	17A1802051	Operación de etapificación (incluye esplenectomía, biopsias hepáticas , de ganglios abdominales y de cresta ilíaca)
	17A1902004	Cirugía de banco, (proc. completo) (micro-extracorpórea), autotrasplante
	17A1902005	Litiasis renal, trat. quir. percutáneo c/s ultrasonido (incluye todo el procedimiento)
	17A2104020	Injertos esponjosos o córtico-esponjosos de cresta ilíaca
	17A2104063	Epicondilitis, trat. quir. (cualquier técnica)
	17A2104131	Fractura de cuello de fémur, osteosíntesis, cualquier técnica (no incluye elementos de osteosíntesis)

Tabla B.2: Variables seleccionadas para el año 2014, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas a nivel de cuidado, 2014 Categoría	ID
Área Pensionado Riesgo medio, autosuficiencia	330, C3
Área Pensionado Bajo riesgo, dependencia parcial	330, D2
Área Médica Adulto Cuidados Medios Bajo riesgo, dependencia total	402,D1
Área Cuidados Intermedios Adultos Bajo riesgo, autosuficiencia	406,D3
Área Cuidados Intermedios Pediátricos Bajo riesgo, dependencia total	412,D1
Área Neonatología Cuidados Básicos Alto riesgo, dependencia parcial	413,B2

Tabla B.3: Variables seleccionadas para el año 2014.
Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2015		
Categoría	ID	Descripción
REM-A08	08180207	ATENCIONES REALIZADAS EN UEH DE HOSPITALES DE MEDIANA COMPLEJIDAD. MATRONA /ÓN
REM-BS17	17010400	EXAMENES DE DIAGNOSTICO - GENETICA
REM-BS17A	17A0203110	Día cama hospitalización integral psiquiatría mediana estadía
	17A0301087	Vitamina B12, absorción de (Co 57 o similar)
	17A0401027	Pielografía de eliminación o descendente: incluye renal y vesical simples previas, 3 placas post inyección de medio de contraste, controles de pie y cistografía pre y post miccional. (7 a 9 exp.)
	17A0404120	Ecotomografía trancraneal
	17A1103010	Craneotomías lineales
	17A1103027	Aneurismas, malformaciones arteriovenosas encefálicas u orbitarias, fístulas durales
	17A1202011	Biopsia de párpado y/o anexos (proc. aut.)
	17A1202074	Hernia de iris y/o fístulas, reparación de
	17A1302016	Reconstitución de conducto auditivo externo, c/s timpanoplastía (incluye revisión de cadena osicular)
	17A1302049	Pólipo nasal y/o coanal, trat. quir.
	17A1402008	Adenoma y/o hiperplasia, trat. quir.- Explor. cervical mas esternotomía por hiperparatiroidismo
	17A1802022	Gastrectomía total
	17A1902009	Nefrectomía parcial y/o cirugía de traumatismo renal
	17A1902028	Cistectomía radical, proc. completo
	17A2104063	Epicondilitis, trat. quir. (cualquier técnica)
	17A2104170	Osteotomía del peroné
	17A3002022	Histiocitosis
	17A3002034	Ca. Mama etapa I y II

Tabla B.4: Variables seleccionadas para el año 2015, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2015		
Categoría	ID	Descripción
REM-BS17C	17C1601111	Aplicación de inmunomodulares, quimicos y similares hasta 10 lesiones
	17C5099033	Dental Ortodoncia: Instalacion aparato removible
REM-BS17D	17D0403103	Angiotac de abdomen
REM-A21	21800100	GESTIÓN DE PABELLON (CIRUGÍA ELECTIVA INSTITUCIONAL)-CIRUGÍA CARDIOVASCULAR
REM-A25	25501000	REACCIONES ADVERSAS A LA DONACIÓN
REM-BS0	99001650	(CS - UMT - BS) CON SINTOMAS LOCALES HEMATOMA Leucemia linfoblástica aguda
	99001720	Tumor de Wilms
	99001740	Histiocitosis
	99001750	Recaída tumores sólidos

Tabla B.5: Variables seleccionadas para el año 2015, asociadas al registro estadístico mensual.
Fuente: Elaboración propia, 2019.

Variables asociadas a nivel de cuidado, 2015	
Categoría	ID
Área Médica Adulto Cuidados Medios Máximo riesgo dependencia parcial	402, A2
Área Médica Adulto Cuidados Medios Máximo riesgo autosuficiencia	402, A3
Área Médico-Quirúrgico Pediátrica Cuidados Básicos Alto riesgo autosuficiencia	409, B3
Área Cuidados Intermedios Pediátricos Bajo riesgo dependencia total	412,D1
Área Cuidados Intermedios Pediátricos Bajo riesgo dependencia parcial	412,D2
Área Neonatología Cuidados Básicos Riesgo medio autosuficiencia	413,C3
Área Psiquiatría Adulto Larga estadía Índice Ocupacional	420,E06
Área Psiquiatría Forense Adulto evaluación e inicio tto. Índice de Rotación	423,E05
Área Psiquiatría Forense Adulto evaluación e inicio tto. Índice Ocupacional	423,E06

Tabla B.6: Variables seleccionadas para el año 2015.
Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2016		
Categoría	ID	Descripción
REM-A07	070251400	CONSULTAS INFECCIÓN TRANSMISIÓN SEXUAL (ITS) Y CONTROLES DE SALUD SEXUAL EN EL NIVEL SECUNDARIO (Incluidos en Sección B)-CONSULTAS VIH/SIDA-PSICOLOGO
REM-A11	11053200	EXÁMENES DE VIH POR GRUPOS DE USUARIOS (Uso exclusivo de establecimientos con Laboratorio que procesan)-DONANTES DE SANGRE-Familiar o Reposición
REM-BS17C	17C1301002	& rinomanometria c/s vasocontrictor
	17C1301037	Dilatacion esofagica por sesion
	17C1707024	Pleuroscopia (toracoscopia) c/s biopsia
	17C1801008	Coledocoscopia intraoperatoria c/s extraccion de calculos
REM-BS17D	17D0303034	Catecolaminas
REM-A25	25400290	COMPONENTE SANGUINEO DISTRIBUIDO (CS) O TRANSFERIDOS (BS Y UMT) PLAQUETAS POOL
	25400330	COMPONENTES SANGUINEOS DISTRIBUIDOS (CS) O TRANSFERIDOS (BS Y UMT) - CRIOPRECIPITADOS
REM-BS0	99001560	Cáncer de Testículo y Germinales extragonadales
	99001610	Ca. Cérvico Uterino
	99001670	Neuroblastoma
REM-BS17A	17A2107009	Luxación congénita de cadera, trat. ortopédico completo (uni o bilateral)
	17A2703013	Profundización de vestíbulo o reconstrucción de rebordes, con o sin injerto
	17A3002002	Linfoma No Hodgkin no agresivo
	17A3002012	Leucemia linfoblástica aguda
	17A3002022	Histiocitosis
	17A3002024	Recaída tumores sólidos
	17A3002030	Glioma
	17A3002137	Ca mama etapa IV metástasis ósea

Tabla B.7: Variables seleccionadas para el año 2016, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2016		
Categoría	ID	Descripción
REM-BS17A	17A0401019	Enema baritada del colon o intestino delgado, doble contraste (12 exp.)
	17A0801010	Necropsia de feto o recién nacido, con estudio histopatológico corriente
	17A1103023	Hematoma intracerebral, vaciamiento de
	17A1103058	Tumor de nervio periférico, extirp. de
	17A1202019	Entropión, plastía de
	17A1202075	Retinopexia neumática
	17A1302014	Exostosis, resección retro o endoaural
	17A1302063	Cuerdas vocales, tumores benignos, trat. quir.- Por vía endoscópica
	17A1402010	Adenoma y/o hiperplasia, trat. quir.- Reintervención por hiperparatiroidismo
	17A1402028	Resección cutánea ampliada (incluye musculatura, ganglios y huesos subyacentes; desplazamiento de colgajos)
	17A1402032	Resección parcial y cirugía reparadora
	17A1502019	- Colgajos osteomusculocutáneos
	17A1502038	Reconst. Osteopl.reborde alveolar- Bilateral en un tiempo
	17A1703014	Endarterectomía carotídea, subclavia, vertebral, femoral, o similar c/s injerto (proc. aut.)
	17A1703031	Trombectomía de venas profundas
	17A1703044	Yugular simple
	17A1704002	Cirugía tórax abierto traumático y/o fijación tórax volante, osteosíntesis costales múltiples y de esternón (no incluye el valor de la prótesis)
	17A1704032	Tratamiento quirúrgico fístula bronquial por toracotomía
	17A1704049	Esofagostomía cervical (proc. aut.)
	17A1704059	Prótesis o tubo endoesofágico, colocación de (proc. aut.)
	17A1802047	Pancreatoduodenectomía
	17A1902015	Suprarrenalectomía unilateral
	17A2104015	Artrotomía hombro o cadera c/u
	17A2104140	Tenotomía aductores c/s botas, con yugo (proc. aut.)
	17A2104145	Osteotomía correctora
	17A2104202	Transplantes tendinosos (cualquier técnica)

Tabla B.8: Variables seleccionadas para el año 2016, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas a nivel de cuidado, 2016	
Categoría	ID
Área Pensionado Riesgo bajo dependencia total	303, D1
Área Médica Adulto Cuidados Medios Máximo riesgo dependencia parcial	402, A2
Área Médica Adulto Cuidados Medios Riesgo bajo dependencia total	402, D1
Área Psiquiatría Adulto Larga estadía	420,E09
Área Psiquiatría Forense Adulto evaluación e inicio tto. Días Cama Disponibles	423,E01
Área Psiquiatría Forense Adulto evaluación e inicio tto. Numero de Egresos	423,E08

Tabla B.9: Variables seleccionadas para el año 2016.
Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2017		
Categoría	ID	Descripción
REM-A08	08230850	Sin Descripción
REM-A09	09309200	CONSULTAS, INGRESOS Y EGRESOS A TRATAMIENTOS EN ESTABLECIMIENTOS DE NIVEL SECUNDARIO Y TERCIARIO-Patología Oral. -Ingresos a tratamiento
REM-A25	25400110	PRODUCCIÓN DE COMPONENTES SANGUÍNEOS (CS-BS)-PLAQUETAS ESTANDAR
	25502000	REACCIONES ADVERSAS A LA DONACIÓN (CS - UMT - BS) CON SINTOMAS GENERALES RVV INMEDIATA CON LESION
REM-BS0	01010442	Coproparasitario seriado con técnica para Cryptosporidium sp o para Diantamoeba fragilis
	01010688	Tomografía Computarizada Angio Cardíaco
	01010816	Cintigrafía ósea trifásica (incluye mediciones fase precoz y tardía)
	01011204	Simblefaron, resección de adherencias y plastía de
	01011213	Biopsia de globo ocular (proc. aut.)
	01011343	Sección simple y/o resección frenillo sublingual retrofaríngeo o faringolaríngeo
	01011380	Cuerdas vocales, tumores benignos, trat. quir.- Por vía endoscópica

Tabla B.10: Variables seleccionadas para el año 2017, asociadas al registro estadístico mensual.
Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2017		
Categoría	ID	Descripción
REM-BS0	01011398	Adenoma y/o hiperplasia, trat. quir.- Autoinjerto de paratiroides
	01011520	Fototerapia UVB, banda angosta y UVA por sesion en cabina
	01011671	Sondeo cardíaco izquierdo y derecho, en adultos o niños
	01011776	Mediastinotomía exploradora ant. o post. c/s biopsia proc. Aut vía cervical
	01011801	Bulas, trat. quir.
	01011822	Esofagectomía total con esofagostomía, gastrostomía y yeyunostomía
	01011834	Ano-recto-sigmoidescopia en niños (además anestesia cód. 22-01-001 si corresponde)
	01011846	Drenaje de la via biliar transhepatica y/o percutaneo (a.c.)
	01012057	Auto o heterotrasplante
	01012058	Cirugía de banco, (proc. completo) (micro-extracorpórea), autotrasplante
	01012261	Endoprótesis total, cualquier técnica
	01012289	Amputación antebrazo
	01012645	Leucemia Mieloide Aguda
	99001570	Enfermedad Trofoblástica Gestacional
	99001660	Leucemia Mieloide Aguda
	99001700	Ewing
	99001790	Recaídas de leucemias Linfoblasticas
	99001820	Glioma

Tabla B.11: Variables seleccionadas para el año 2017, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas a nivel de cuidado, 2017 Categoría	ID
Área Pensionado Riesgo bajo dependencia parcial	303, D2
Área Médica Adulto Cuidados Medios Máximo riesgo dependencia parcial	402, A2
Área Cuidados Intermedios Adultos Máximo riesgo autosuficiencia	406, A3
Área Cuidados Intermedios Adultos Bajo riesgo dependencia total	406,D1
Área Cuidados Intermedios Pediátricos Bajo riesgo dependencia parcial	412,D2
Área Psiquiatría Adulto Larga estadía Días Cama Ocupados	420,E02
Área Psiquiatría Adulto Larga estadía Índice Ocupacional	420,E06
Área Psiquiatría Forense Adulto evaluación e inicio tto. Días de Estada	423,E03

Tabla B.12: Variables seleccionadas para el año 2017.
Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2018		
Categoría	ID	Descripción
REM-A07	07020800	CONSULTAS MÉDICAS ESPECIALIDADES Y SUB-ESPECIALIDADES-Genética
	07021100	CONSULTAS MÉDICAS ESPECIALIDADES Y SUB-ESPECIALIDADES-Nutrición
	07021531	CONSULTAS MÉDICAS ESPECIALIDADES Y SUB-ESPECIALIDADES-Geriatría
	07022132	Sin Descripción
REM-A08	24100200	INGRESOS POR EMERGENCIA OBSTÉTRICA AL SERVICIO DE URGENCIA (Establecimientos Alta y Mediana Complejidad).-Preeclampsia severa
REM-A11	11055110	EXÁMENES SEGÚN GRUPOS DE USUARIOS POR CONDICIÓN DE HEPATITIS B, HEPATITIS C, CHAGAS, HTLV 1 Y SIFILIS (Uso exclusivo de establecimientos con Laboratorio que procesan)-DONANTES-Familiar o Reposición
REM-BS0	99001500	Linfoma de Hodgkin
	01010097	Día cama hospitalización integral psiquiatría mediana estadía
	01010289	IGF1 o Somatomedina - C (Insuline Like Growth Factor)
	01010442	Coproparasitario seriado con técnica para Cryptosporidium sp o para Diantamoeba fragilis
	01010604	Estudio radiológico de deglución faríngea
	01011059	Craniectomías descompresivas
	01011065	Hematoma o absceso extradural, vaciamiento de

Tabla B.13: Variables seleccionadas para el año 2018, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas al REM, 2018		
Categoría	ID	Descripción
REM-BS0	01011075	Extirpación de tumores y/o quistes y/o cavernoma de base de cráneo
	01011227	Recubrimiento conjuntival
	01011373	Sinequia nasal, trat. quir.
	01011375	Vaciamiento etmoidal por vía nasal c/s polipectomía
	01011397	Tiroidectomía total ampliada (incluye extirpación de estructuras anatómicas vecinas)
	01011468	Colgajos musculares o musculocutáneos
	01011698	Puentes aorto bifemoral puentes de troncos supra-aórticos
	01011784	Heridas traumáticas, trat. quir.
	01011956	Colostomía (proc. aut.)
	01012033	Cistografia por sonda (de relleno) o por puncion hipo-
	01012059	Litiasis renal, trat. quir. percutáneo c/s ultrasonido (incluye todo el procedimiento)
	01012102	Fistulectomía
	01012123	Cirugía del epidídimo y cordón (proc.aut), incluye cirugía intravaginal y/o varicocele mismo lado
	01012151	Electrodiatermo o criocoagulación de lesiones del cuello
	01012196	Vulvectomía- Radical
	01012289	Amputación antebrazo
	01012422	Pie reumatoídeo, trat. quir. completo (cualquier técnica)
	01012646	Neuroblastoma
	01012650	Tumores germinales Extra Sistema Nerviso Central (Extra SNC)
	01012658	Recidiva de leucemias Linfoblasticas
	01012828	Aspiración manual endouterina

Tabla B.14: Variables seleccionadas para el año 2018, asociadas al registro estadístico mensual. Fuente: Elaboración propia, 2019.

Variables asociadas a nivel de cuidado, 2018	
Categoría	ID
Área Cuidados Intensivos Adultos Máximo riesgo autosuficiencia	405, A3
Área Neonatología Cuidados Básicos Riesgo medio dependencia parcial	413, C2
Área Psiquiatría Adulto Larga estadía Dias Cama Ocupados	420,E02
Área Psiquiatría Adulto Larga estadía Numero de Egresos	420,E08
Área Psiquiatría Adulto Mediana estadía Numero de Egresos	419,E08

Tabla B.15: Variables seleccionadas para el año 2018.
Fuente: Elaboración propia, 2019.