Comparación de variables en poblaciones pediátricas y no pediátricas[[1]](#footnote-1)

Esteban Paduro[[2]](#footnote-2)

Junio 2024

# Introducción

El problema de la estimación de la demanda hospitalaria, particularmente la asociada a problemas respiratorios corresponde a un importante desafía que deben enfrentar los diferentes centros de salud pues afecta la asignación de manera eficiente de recursos. Es en este contexto que la pregunta del nivel de agrupación apropiado para el análisis de datos pues, por un lado, el tener más datos permite realizar mejores análisis estadísticos, sin embargo, el agrupar datos que posean comportamientos significativamente diferentes puede dificultar la extracción de buenas conclusiones a partir de las observaciones. Para estudiar esta disyuntiva es relevante identificar cuáles son las características principales que se desea rescatar de las observaciones.

La confección de este informe se enmarca en el proceso de generar una plataforma de predicción de hospitalizaciones por causas respiratorias. En la primera etapa dicho proyecto se enfoca en estudiar poblaciones pediátricas, ya que análisis preliminares sugieren que su comportamiento es cualitativamente diferente al de poblaciones no pediátricas. En el presente informe daremos exploraremos más a fondo esta disyuntiva además de estudiar posibles diferencias asociadas a distribución geográfica.

## Objetivos del proyecto

El objetivo del presente documento es realizar una comparación entre las características de las curvas de demanda de atenciones de urgencia asociada a Infecciones Respiratorias Agudas (IRA) correspondientes a poblaciones pediátricas y no pediátricas. Los objetivos específicos son los siguientes:

* Se desea decidir si es apropiado analizar de manera agrupado ciertos grupos poblacionales correspondientes a diferentes grupos etarios.
* Se busca identificar variables que expliquen las diferencias observadas en el comportamiento del máximo de demanda de atenciones de urgencias respiratorias.

## Análisis preliminar

Al estudiar el perfil de demanda de atenciones usualmente [ref] se asume que este estacional, que en el caso de las atenciones correspondientes a enfermedades respiratorias tiende a presentar un máximo de demanda durante los meses más fríos de año. Es por esto que resulta conveniente estudiar la demanda máxima (*peak*) de atenciones de urgencia. Dependiendo del contexto puede resultar útil el estudiar el peak corresponde a las diferentes escalas de tiempo, tales como: mes, semana o día en que se presenta la máxima demanda de atenciones. Es particularmente importantes para la planificación y asignación de recursos el tener buenas estimaciones de la fecha en que ocurre el peak de demanda de atenciones y la magnitud de dicho peak.

Información histórica [ref] indica que es insuficiente el estimar estos datos solo con información de años anteriores, pues la evidencia muestra que la fecha de este peak puede tener variaciones significativas (Ver Sección X.X) de un año a otro, entre sectores geográficos o diferentes grupos etarios.

Debido a los motivos antes mencionados resulta relevante estudiar con cuidado la pregunta de la agrupación apropiada de los datos de la demanda de atenciones.

## Hipótesis para el análisis

Para guiar nuestro análisis estableceremos las siguientes hipótesis de investigación.

H1: existen diferencias significativas entre las poblaciones pediátricas y no pediátricas en términos de demanda de atenciones hospitalarias de urgencia.

Una posible sutileza este análisis es que observaciones preliminares siguieres lo siguiente

* El peak de demanda hospitalaria en las poblaciones menores de un año y mayores a 65 años parecen ser menos pronunciados que en los otros grupos.
* Al mirar total de atenciones, el perfil de demanda a lo largo del año en los grupos 1-4 años, 5-14 años y 15-64 años son más parecidos, pero la composición del tipo de atenciones parece ser diferente. Es necesario hacer un análisis más detallado para poder discriminar ente estos grupos y establecer diferencias que hagan necesario un análisis diferenciado en término de predicción de demanda de atención hospitalaria.

H2: Es posible estudiar el perfil de demanda hospitalaria porcentual de manera independiente de la estimación del total de atenciones anuales de cierto grupo de población.

## Descripción de los datos

Nuestra fuente de información corresponde las principales causas de atenciones de urgencia, de manera agrupada, realizadas dentro de los establecimientos públicos, pertenecientes al Sistema Nacional de Servicios de Salud (SNSS)[[3]](#footnote-3), las cuales se reportaron al sistema DEIS [[4]](#footnote-4). Las causas de atenciones de urgencia consideradas corresponden a 6 categorías descritas en los datos del ministerio de salud detalladas en la Tabla 1.

|  |  |
| --- | --- |
| **IdCausa** | **Glosa Causa** |
| 3 | Bronquitis/bronquiolitis aguda (J20-J21) |
| 4 | Influenza (J09-J11) |
| 5 | Neumonía (J12-J18) |
| 6 | Otra causa respiratoria (J22, J30-J39, J47, J6... |
| 10 | IRA Alta (J00-J06) |
| 11 | Crisis obstructiva bronquial (J40-J46) |

Tabla 1 Lista de causas respiratorias según definición MINSAL [[5]](#footnote-5)

Los datos se encuentran desagregados a nivel diario con datos desde el 2017 hasta el 2024, sin embargo, en este estudio nos enfocaremos en los datos hasta 2023. Adicionalmente daremos un trato especial a los datos correspondiente al periodo 2020-2021 por encontrarse fuertemente afectado por la pandemia COVID-19.

En términos de grupos etarios, los datos están desagregados en 5 categorías

* Menores de 1 año
* Entre 1 y 4 años
* Entre 5 y 14 años
* Entre 15 y 64 años
* Mayores de 65 años

En términos geográficos, se disponen datos de centros de salud de 16 regiones, además de contar información desagregada a nivel de servicio de salud.

Para determinados análisis nos restringiremos a centro de salud tipo hospitales o ciertas agrupaciones geográficas como todos los servicios de salud de la región metropolitana o agrupación por regiones en zonas.

# Análisis de Niveles de Agrupación

## Análisis preliminar

En objetivo de esta sección es recopilar algunos totales que nos permitan nuestro análisis. A lo largo de este estudio consideraremos diferentes niveles de agrupaciones que nos permitirán identificar variables bajo que niveles de agrupación

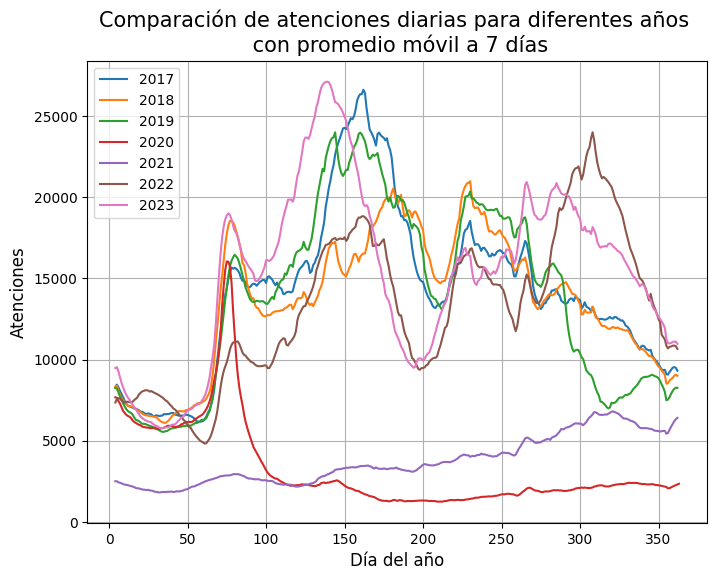


Ilustración 1 Atenciones diarias de urgencia para cada día del año correspondientes al periodo 2017 -2023

En la Ilustración 1, se muestra el total de atenciones diarias de urgencia en que se han agrupado datos de establecimientos públicos de todo el país separado a nivel de día e para cada año. Es posible observar que el comportamiento de los años 2020 y 2021, es muy diferente al resto de los años. Adicionalmente es posible identificar que hay un alza en la demanda de atenciones alrededor del día 150, el cual varía a lo largo de los años, y además es posible identificar un alza de la demanda, con menor magnitud durante la segunda mitad del año. Notamos que incluso a este nivel de agrupación es posible observar diferencias importantes en cuanto a la fecha del peak de atenciones de urgencia.

De manera similar en la Ilustración 2 consideramos la comparación de las curvas de demanda de atenciones correspondientes a diferentes grupos etarios. Esta vez consideramos las atenciones correspondientes a todos los establecimientos de salud pública del país, pero escalamos las curvas usando su total anual, de manera que las atenciones diarias se presentan como un porcentaje de las atenciones anuales. Como ya identificamos que existen variaciones importantes de año a año, decidimos fijar el año (2019 en este caso) y comparamos las curvas de demanda. A partir de esta Ilustración es posible identificar que los datos para menores de un año y para mayores de 65 presentan un comportamiento con peaks menos marcados, pero para realizar observaciones más precisar necesitaremos herramientas más elaboradas que desarrollaremos en la siguiente sección.

Gráfico, Gráfico de líneas, Histograma

Descripción generada automáticamente

Ilustración 2 Año2019, suavizado a 7 días

El siguiente factor que queremos analizar es la dependencia geográfica de las demandas, para esto separaremos nuestros datos en 6 zonas geográficas identificadas por lo correspondiente

s servicios de salud asociados

1. Zona norte grande: Arica y Parinacota, Tarapacá y Antofagasta. Regiones 15, 1, 2. Servicios de salud con ID: 1, 2, 3
2. Zona norte chico: Atacama y Coquimbo (3,4). Regiones 3,4. Servicios de Salud con ID: 4, 5
3. Zona Núcleo Central: Valparaíso, Metropolitana de Santiago, O'Higgins, Maule y Ñuble. Regiones 5,13,6,7,16. Servicios de Salud con ID: 6,7,8,9,10,11,12,13,14,15, 16, 17.
4. Zona Concepción y La Frontera: Biobío y La Araucanía. Regiones 8,9. Servicios de Salud ID: 18,19,20,21, 28,29,
5. Zona Región de los Lagos: Los Ríos y Los Lagos (continental) . Regiones 14,10. Servicios de Salud ID: 22,23,24.
6. Zona Región de los Canales: Los Lagos (Provincia de Chiloé), Aysén y Magallanes. Regiones 10, 11,12. Servicios de Salud ID: 25, 26, 33.

A partir de esta clasificación construimos la ilustración 3, es ella es posible visualizar las atenciones diarias, escaladas para mostrar las atenciones diarias como un porcentaje de las atenciones anuales. Esta visualización nos permite identificar diferencias significativas.

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 3 Atenciones (%) diarias agrupadas por zona geográfica para cada día del año, Años 2017,2018,2019,2022,2023

Para un año fijo, podemos hacer comparaciones entre grupo etario y zona geográfica, estos resultados se muestran en la Ilustración 4

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 4 atenciones diarias como porcentaje de las atenciones anuales, correspondientes al año 2019, separado por grupos etarios y zonas geográficas

## Diferencias por Grupo Etario

Una de las preguntas claves del análisis es el poder decidir cuál es el nivel de agrupación apropiado de los datos. Adicionalmente puede ser relevante analizar de forma independiente los datos de IRA, pues ellos son el objetivo principal del proyecto.

Diferentes niveles de agrupación permiten estudiar diferentes aspectos de los datos bajo consideración para evaluar nuestra hipótesis de que los datos de diferentes grupos etarios deben ser estudiados de manera diferenciada, propondremos medidas que permitan identificar que permitan identificar que ellos significativamente diferentes.

Para este propósito consideremos el siguiente procedimiento para preparar la información: consideraremos series de datos correspondientes a datos diarios asociados a un cierto año y características asociadas a los grupos que deseamos comparar.

Para el primer análisis agruparemos las atenciones de urgencias respiratorias agrupadas por año, región y grupo etario. De esta forma, por ejemplo, para la combinación 2022, Región 5 y la población de entre 5 y 14 años le corresponderá una lista de atenciones diaria que se indizará por el número de día del año.

Considerando la influencia de factores externos, y debido a que se tienen suficientes datos es conveniente separar el análisis por años. Una vez que se tienen las atenciones diarias correspondientes a cada uno de estos triples (año, región, grupo etario), notamos que los datos tienen un comportamiento bastante ruidoso. Una forma de disminuir los efectos de las variaciones diarias es considerar la cantidad de atenciones acumuladas a cierto día del año, adicionalmente para poder comparar la forma del perfil de demanda de atenciones, escalaremos la curva de tal manera que se represente el porcentaje de las atenciones anuales que han ocurrido hasta cierta fecha.

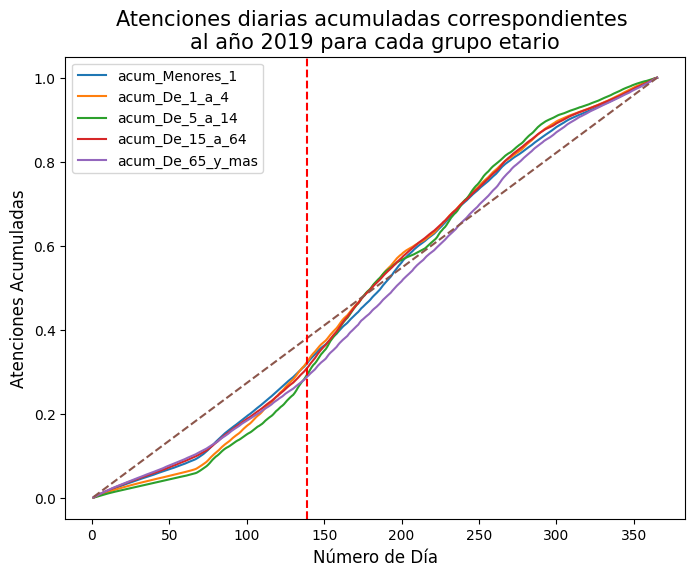


Ilustración 5: Atenciones acumuladas año 2019, por grupo etario, V región. En rojo se indica la fecha correspondiente al peal de atenciones de urgencias respiratorias para niños de 1 a 4 años. La línea roja indica la fecha en que se produce el peak de demanda de atenciones para el grupo etario De\_1\_a\_4

A partir de esta curva Ilustración 5 ya es posible observar diferencias cualitativas entre los perfiles de demanda de atenciones, así un perfil de más cercano a la diagonal, como lo es el caso de atenciones de mayores de 65, representa que las atenciones se distribuyen de manera más uniforme a lo largo del año. Adicionalmente es posible estudiar la época del año en que ocurre la mayor demanda de atenciones fijando una ventana temporal (digamos una semana) y buscar el rango de 5 días consecutivos en que se produjo la mayor demanda de atenciones indicado con una línea roja vertical en la Ilustración 1 el correspondiente al grupo etario entre 1 y 4 años.

La ventaja de exponer los datos de esta manera es que permite comparar de manera más directa diferentes grupos. A partir de esta información nos gustaría evaluar nuestra hipótesis de similitud o diferencias de comportamientos de diferentes grupos poblacionales, para este objetivo utilizaremos una técnica de reducción de dimensionalidad conocida como análisis de componentes principales (PCA) [ref].

Cada observación corresponde a un vector con largo la cantidad de días del año. Las atenciones correspondientes a un mismo día son estandarizadas a lo largo de todas las observaciones para tener media cero y varianza 1. Para nuestro análisis consideraremos las 3 primeras componentes principales [explicar por qué tiene sentido tomar 3, ¿gráfico de la varianza explicada?].

Para nuestro primer análisis consideraremos la mayor cantidad de datos que tenemos a disposición. Se considerarán los datos de los años 2017, 2018, 2019, 2022, 2023, para todas las regiones. La agrupación se realizará a nivel de servicio de salud (29 en total). Se considerarán por separado las atenciones correspondientes a cada uno de los grupos etarios definidos en la introducción, adicionalmente nos restringiremos a las atenciones diarias por urgencias respiratorias.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Ilustración 6 Componentes principales asociados a los datos de urgencias respiratorias agrupados por año, servicio de salud, grupo etario.

La Ilustración 6 nos permite identificar que existe un comportamiento marcadamente diferente entre la agrupación De\_1\_a\_4 y De\_5\_a\_14 vs el resto de los individuos, también a partir de las gráficas notamos que el comportamiento del grupo Menores\_1 no parece poder separarse de las demás clases. Para poder formalizar estas observaciones aplicaremos un método de clasificación binaría basada en métodos de Machine Learning conocida como Support Vector Machine (SVM) [ref], a partir de ella buscaremos un plano en el subespacio descrito por las componentes principales (PC1, PC2, PC3) el cual intenta separar de los elementos de una clase vs los elementos fuera de ella. La motivación es que, si somos capaces de hallar un plano que separe las clases un alto nivel de precisión, entonces es una indicación razonable que los conjuntos de datos tienen comportamientos cuantitativamente diferentes, y por tanto hay razones para estudiarlos de forma. Para el caso antes considerado consideraremos como una clase a la agrupación De\_1\_a\_4 y De\_5\_a\_14, y trataremos de separarla del resto. En este caso se obtiene una precisión del clasificador del 89.7%, lo cual sugiere que

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Ilustración 7 Diagrama de la clasificación Binaria y el correspondiente plano que separa las clases. En este caso se indican en rojo el grupo formado al agrupar las clases De\_1\_a\_4 y De\_5\_a\_14

No es claro que la separación hecha anteriormente sea la mejor, entonces consideraremos una procedimiento más exhaustivo. Como tenemos cinco grupos etarios consideraremos todas las posibles clasificaciones en dos clases entre ellos. Debido a la simetría del problema, existen solo dos casos posibles, separar uno de los otros 4 y separar 2 de los otros 3. De esta forma la Tabla [] indica en la fila i, columna j el porcentaje de clasificación correcta cuando se intenta separar el conjunto (i,j) del resto, además los elementos en la diagonal representan que solo separamos una clase de las otras cuatro.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Menores\_1 | De\_1\_a\_4 | De\_5\_a\_14 | De\_15\_a\_64 | De\_65\_y\_mas |
| Menores\_1 | 0.717362 |  |  |  |  |
| De\_1\_a\_4 | 0.577845 | 0.738000 |  |  |  |
| De\_5\_a\_14 | 0.604345 | **0.867310** | 0.812241 |  |  |
| De\_15\_a\_64 | 0.600138 | 0.544828 | 0.529879 | 0.728672 |  |
| De\_65\_y\_mas | 0.655603 | 0.527138 | 0.571655 | **0.821172** | **0.843552** |

Tabla 2 Precisión de la clasificación como promedio de 1000 repeticiones. Observamos que los tres mejores resultados corresponden a separar De\_1\_a\_4 y De\_5\_a\_14 del resto, remover De\_65\_y\_mas del resto y finalmente separa De\_15\_a\_64 con De\_65\_y\_mas del resto

Los resultados presentados en la Tabla 2 indican dentro de nuestros datos existen subclases que tienen comportamientos significativamente diferentes. Visto de otra forma los grupos identificados tienen una clara interpretación en términos de nuestros datos, nos dicen datos correspondientes a grupos pediátricos (De\_1\_a\_4, De\_5\_a\_14) o (Menores\_1, De\_1\_a\_4, De\_5\_a\_14) tienen comportamientos cuantitativamente diferentes a los correspondientes de a no pediátricos, lo que sugiere que es apropiado estudiarlos de forma separa con respecto a el análisis de la curvas de demanda.

## Influencia de la Agrupación Geográfica

Siguiendo en análisis de la sección anterior esta vez consideraremos cual es la mejor segregación según zona geográfica. Motivados por los resultados de la sección anterior consideraremos separar nuestro análisis entre Población pediátrica (Menores\_1, De\_1\_a\_4, De\_5\_a\_14) vs Población no pediátrica (De\_15\_a\_64, De\_65\_y\_mas).

Para el análisis consideraremos 3 años de manera simultánea (2017,2018,2019) pues la Ilustración 3 sugiere comportamiento parecido. En primera instancia miraremos los datos correspondientes a todos los grupos etarios y luego separaremos entre pediátricos y no pediátricos, y haremos un análisis de componentes principales sobre estos conjuntos de datos. Consideraremos diferentes agrupaciones para las zonas geográficas, de manera de tratar de determinar si alguna de ellas parece apropiada para el conjunto de datos bajo consideración

Tabla 3 Agrupaciones por zonas geográficas. datos de todas las grupos etarios simultáneamente para los años 2017, 2018 y 2019

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Agrupación | Precisión del clasificador |  | Agrupación | Precisión del clasificador |  | Agrupación | Precisión del clasificador |
| (0) | **0,78069** |  | (1, 3) | 0,64056 |  | (0, 1, 5) | 0,50667 |
| (1) | **0,78125** |  | (1, 4) | 0,57056 |  | (0, 2, 3) | 0,55222 |
| (2) | **0,76069** |  | (1, 5) | 0,56542 |  | (0, 2, 4) | 0,60847 |
| (3) | **0,74875** |  | (2, 3) | 0,56153 |  | (0, 2, 5) | 0,56778 |
| (4) | **0,80681** |  | (2, 4) | 0,59403 |  | (0, 3, 4) | 0,53986 |
| (5) | **0,73944** |  | (2, 5) | 0,53667 |  | (0, 3, 5) | 0,53167 |
| (0, 1) | 0,64639 |  | (3, 4) | 0,64528 |  | (0, 4, 5) | 0,63153 |
| (0, 2) | **0,71500** |  | (3, 5) | 0,64458 |  |  |  |
| (0, 3) | 0,58917 |  | (4, 5) | **0,72028** |  |  |  |
| (0, 4) | 0,63667 |  | (0, 1, 2) | **0,73736** |  |  |  |
| (0, 5) | 0,57750 |  | (0, 1, 3) | 0,57486 |  |  |  |
| (1, 2) | 0,69472 |  | (0, 1, 4) | 0,51569 |  |  |  |

Los resultados obtenidos a considerar solamente poblaciones pediátricas se muestran en la Tabla 3. Observamos que en este caso el clasificador sugiere que es razonable considerar cada una de las zonas como que tiene un comportamiento suficientemente significativo para obtener un algo nivel de separación. Además de entre los casos que efectivamente agrupan zonas, notamos que las que parecen ser más apropiadas son las que corresponden a asociar las zonas más extremas del territorio, es decir. (0,2) vs (1,3,4,5); (0,1,2) vs (3,4,5) y (0,1,2,3) vs (4,5).

Tabla 4 Resultados de las clasificaciones agrupando zonas geográficas considerando solo gropos poblacionales pediátricos correspondientes a los años 2017, 2018 y 2019

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Agrupación | Precisión del clasificador |  | Agrupación | Precisión del clasificador |  | Agrupación | Precisión del clasificador |
| (0) | **0,78930** |  | (1, 3) | 0,65977 |  | (0, 1, 5) | 0,54860 |
| (1) | **0,75791** |  | (1, 4) | 0,53256 |  | (0, 2, 3) | 0,55465 |
| (2) | **0,79302** |  | (1, 5) | 0,58698 |  | (0, 2, 4) | 0,66953 |
| (3) | **0,76535** |  | (2, 3) | 0,59070 |  | (0, 2, 5) | 0,57535 |
| (4) | **0,78767** |  | (2, 4) | 0,61628 |  | (0, 3, 4) | 0,55488 |
| (5) | **0,81651** |  | (2, 5) | 0,61279 |  | (0, 3, 5) | 0,51860 |
| (0, 1) | 0,64465 |  | (3, 4) | 0,64047 |  | (0, 4, 5) | 0,67581 |
| (0, 2) | **0,74140** |  | (3, 5) | 0,69674 |  |  |  |
| (0, 3) | 0,57558 |  | (4, 5) | **0,73349** |  |  |  |
| (0, 4) | 0,70163 |  | (0, 1, 2) | **0,74209** |  |  |  |
| (0, 5) | 0,61744 |  | (0, 1, 3) | 0,57744 |  |  |  |
| (1, 2) | 0,66047 |  | (0, 1, 4) | 0,54953 |  |  |  |

Finalmente hacemos el análisis análogo considerando solo grupos de población no pediátricos. En este caso los resultados sugieren que en el caso de adultos la separación más efectiva se logra al considerar las agrupaciones (0,4,5) vs (1,2,3), es decir las zonas más extremas versus las centrales.

Tabla 5 Análisis de dependencia geográfica, considerando solo grupos etarios correspondientes a población no pediátrica, años 2017, 2018 y 2019

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Agrupación | Precisión del clasificador |  | Agrupación | Precisión del clasificador |  | Agrupación | Precisión del clasificador |
| (0) | **0,83071** |  | (1, 3) | 0,72607 |  | (0, 1, 5) | 0,64929 |
| (1) | **0,86286** |  | (1, 4) | 0,62321 |  | (0, 2, 3) | 0,58750 |
| (2) | **0,81714** |  | (1, 5) | 0,58321 |  | (0, 2, 4) | 0,53250 |
| (3) | **0,80500** |  | (2, 3) | 0,73607 |  | (0, 2, 5) | 0,64357 |
| (4) | **0,85393** |  | (2, 4) | 0,61357 |  | (0, 3, 4) | 0,59679 |
| (5) | **0,82821** |  | (2, 5) | 0,64786 |  | (0, 3, 5) | 0,55500 |
| (0, 1) | **0,80393** |  | (3, 4) | 0,75643 |  | (0, 4, 5) | **0,90500** |
| (0, 2) | 0,66857 |  | (3, 5) | 0,64929 |  |  |  |
| (0, 3) | 0,55929 |  | (4, 5) | **0,84500** |  |  |  |
| (0, 4) | 0,71321 |  | (0, 1, 2) | **0,85357** |  |  |  |
| (0, 5) | 0,78607 |  | (0, 1, 3) | 0,63750 |  |  |  |
| (1, 2) | **0,82643** |  | (0, 1, 4) | 0,66536 |  |  |  |

## Agrupación por Tipo de Centro de Salud

1. CMM – Línea de Salud Digital, proyecto de estimación de demanda IRA, XXXX. [↑](#footnote-ref-1)
2. Departamento de Ingeniería Matemática and Centro de Modelamiento Matemático (CNRS IRL2807), Universidad de Chile, Santiago, Chile [↑](#footnote-ref-2)
3. Manual de Registro Sistema de Atención Diaria de Urgencia, Departamento de Estadística e Información de Salud (DEIS), Enero 2021 [↑](#footnote-ref-3)
4. <https://datos.gob.cl/dataset/atenciones-de-urgencia> [↑](#footnote-ref-4)
5. Manual de Registro Sistema de Atención Diaria de Urgencia, Departamento de Estadística e Información de Salud (DEIS), Enero 2021 [↑](#footnote-ref-5)