Análisis de la Asistencia de Pacientes a Controles Odontológicos del Programa CERO: Enfoques Estadísticos y Modelos Predictivos

Autores: Fernanda Jahn, Rodrigo Fuenzalida, Alejandro Heredia, Gonzalo Barría
Profesor Corrector: Francklin Rivas Evcheverría

---

1. Resumen Ejecutivo

En los Centros de Salud Familiar del país se ejecuta el llamado 'Programa CERO'. El programa CERO es un programa orientado a la prevención de caries en niños de 0 a 9 años, que busca realizar controles preventivos en niños de esas edades, determinando el riesgo cariogénico y según aquello determinando la fecha aproximada del próximo control, el cual puede ser 6 o 12 meses después.

Este whitepaper presenta un análisis exhaustivo de la asistencia de pacientes a controles odontológicos del 'Programa CERO', utilizando datos provenientes de registros manuales y electrónicos de citas y ausencias, y datos calculados mediante geocoding.

A través de una combinación de inferencia estadística y machine learning, se busca identificar perfiles de pacientes propensos a la inasistencia, predecir patrones futuros y mejorar la programación de citas de dicho programa en los centros odontológicos.

La fase inicial se centra en el data wrangling y el análisis estadístico descriptivo e inferencial, estableciendo una base sólida para comprender las variables clave y sus interrelaciones.

Las fases subsiguientes implementan modelos de aprendizaje no supervisado, para dar paso a una posterior incorporación de redes neuronales y optimizar la precisión predictiva.

Se prestará especial atención a la segmentación de pacientes por edad (niños/adolescentes y adultos) para adaptar las estrategias de predicción y abordaje, reconociendo las particularidades de cada grupo. Este enfoque permitirá desarrollar soluciones innovadoras para optimizar los recursos y disminuir las inasistencias mejorando a su vez la cobertura de atención. El impacto esperado incluye una profundización de la población usuario para la mejora en la eficiencia de la gestión de citas y la optimización de recursos en el Programa CERO.

2. Objetivo del Estudio

El propósito fundamental de este análisis es identificar y caracterizar los perfiles de pacientes que presentan mayor riesgo de inasistencia a controles odontológicos, a fin de mejorar la planificación y eficiencia en la asignación de citas. En lugar de solo investigar las causas generales de inasistencia, el estudio se enfoca en crear perfiles basados en datos demográficos, historial de citas y otras variables relevantes. Este enfoque permitirá una comprensión más profunda de los factores subyacentes a la inasistencia y facilitará el diseño de intervenciones más efectivas.

En el corto plazo, el estudio se centrará en el análisis estadístico de los datos disponibles para comprender las variables significativas relacionadas con la inasistencia. A medida que los resultados se afiancen, se procederá con el desarrollo de modelos predictivos más complejos que puedan ofrecer recomendaciones personalizadas, permitiendo una gestión más efectiva de los pacientes.

El estudio también busca evaluar la viabilidad de integrar técnicas de deep learning para mejorar las predicciones y el uso de big data para manejar grandes volúmenes de información. Este análisis se alinea con las necesidades del Programa CERO y busca generar un impacto positivo en la salud bucal de la comunidad a través de labores de prevención y promoción.

3. Metodología

3.1. Preparación de los Datos

La calidad y la preparación de los datos son esenciales para el éxito de cualquier análisis. Junto con otras recomendaciones nos enfocaremos principalmente en el "Aseguramiento de la calidad estadística" del INE (2024). En este estudio, los datos provienen de una fuente principal llamada 'bajocontrol.csv'.

- `bajocontrol.csv`: Es un registro en base a una planilla Google Spreadsheets donde se registra, entre otras variables del paciente, el riesgo odontológico, la fecha (según el riesgo) de próximo control y si los pacientes asistieron o no a dicho control. La fecha de próximo control es un estimativo de cuándo el adulto responsable debe solicitar la hora de control, y la inasistencia se calcula cada vez que se abre la planilla, si la fecha del día es menor a la fecha fijada para el próximo control. El niño o niña puede haber quedado inasistente por haber solicitado una

cita y no haber asistido a ella, o por no haber solicitado la cita cuando correspondiese.

El dataset original incluye información crucial sobre los pacientes, como:

- *Datos personales y demográficos*: RUT (código único de identificación), sexo, etnia, discapacidad, entre otros.
- *Historial de control:* Fechas de nacimiento, últimos controles realizados, daño por caries, entre otros.
- *Inasistencias a Citas programadas:* Información sobre las fechas y horas de citas programadas a las que no asistieron.
- *Distancia al cesfam en metros:* Este campo es un campo calculado, pues luego de obtener las direcciones de todos los pacientes de 'bajocontrol.csv', se iteró por todas ellas utilizando las apis de geocoding de ArcGIS y Google Maps, parseada obteniendo solo las coordenadas. Luego de obtener las coordenadas, se realizó el cálculo de la distancia euclidiana al Cesfam con distance() de Shapely.

Anonimización y preparación de datos

El primer paso en la preparación de los datos incluye la limpieza y transformación de las variables. Se estandarizaron las fechas para poder realizar comparaciones coherentes, se eliminó información redundante y se ajustó los tipos de datos para asegurar que las variables estén listas para el análisis. En este proceso, se utilizó la librería dateparser para transformar algunas fechas que originalmente estaban en formato *string*.

Además, se eliminaron columnas irrelevantes para el análisis o que pudieran permitir la identificación de pacientes, tales como RUT, dígito verificador, nombre y columnas de índice, con el fin de resguardar la confidencialidad de los datos. También se generaron variables derivadas, como el redondeo de la edad en meses, lo que hizo innecesario conservar información sensible como la "Fecha de Nacimiento", la cual fue eliminada por su potencial para individualizar a los pacientes. Nuestro objetivo es seguir las recomendaciones de la "Guía de Formulación ética de proyectos de ciencia de datos" de Gobierno Digital y la UAI (2022) y de la "Guía de Buenas Prácticas de Privacidad y Seguridad de Datos en Salud" del CENS. (2024). Este proceso de preparación de datos garantiza la calidad y la integridad de la información utilizada en el análisis, mientras se asegura la protección a la vida privada de las personas.

Luego del proceso de cálculo de 'distancia_cesfam', se detectó la presencia de outliers que podían afectar los análisis, entonces se realizó la imputación de estos reemplazándolos por la mediana.

Dado que la cantidad de valores nulos (NaN) sin considerar la distancia al cesfam era baja y no representaba un porcentaje significativo del total, se optó por eliminarlos directamente sin aplicar métodos de imputación. Esta operación resultó en la eliminación de 72 filas del conjunto de datos.

El manejo de datos sensibles y la desanonimización fue realizada por uno de los integrantes del grupo, que es funcionario del Cesfam del cual se obtuvieron los datos. El resto del grupo nunca tuvo acceso a datos que permitieran individualizar a una persona.

Diccionario de Datos

Nombre de la variable	Descripción
SEXO	Sexo del paciente (Mujer / Hombre).
PUEBLO ORIGINARIO	Indica si pertenece a un pueblo originario (SI / NO).
MIGRANTES	Indica si el paciente es migrante (SI / NO).
USUARIO CON DISCAPACIDAD	Indica si el paciente presenta alguna discapacidad (SI / NO).
MEJOR NIÑEZ	Indica si pertenece al programa "Mejor Niñez" (SI / NO).
SENAME	Indica si pertenece al SENAME (SI / NO).
DAÑO POR CARIES	Número de dientes con daño por caries.
RIESGO	Clasificación de riesgo según evaluación odontológica (BAJO RIESGO / ALTO RIESGO).
MESES PARA PRÓXIMO CONTROL	Tiempo estimado (en meses) para el próximo control.
INASISTENTE	Indica si el paciente no solicitó una hora de atención o no asistió a su última cita programada (SI / NO APLICA). **Esta es nuestra variable objetivo.**
EDAD MESES	Edad del paciente expresada en meses.
MESES ULTIMO CONTROL	Tiempo transcurrido (en meses) desde el último control.
DISTANCIA AL CESFAM	Distancia euclidiana en metros al Cesfam.

3.2. Análisis Estadístico Inferencial

En esta etapa, se llevará a cabo un análisis exploratorio de los datos para comprender la distribución de las variables y detectar patrones potenciales en la asistencia a los controles odontológicos. Se realizará un análisis comparativo exhaustivo entre los grupos de edad para identificar diferencias significativas en los factores asociados a la inasistencia. Este análisis permitirá comprender mejor las necesidades y desafíos específicos de cada grupo. Las técnicas estadísticas inferenciales que se aplicarán incluyen:

- **1. Estadísticas descriptivas:** Se realizó un análisis de la media, mediana, moda, desviación estándar y distribución de variables clave como la edad, la cantidad de meses transcurridos desde el último control y el daño por caries. Además, se evaluaron correlaciones básicas entre las principales variables numéricas, lo que permitió identificar relaciones lineales preliminares relevantes. Este análisis proporciona una visión general de las características y posibles interrelaciones dentro de la población estudiada.

- **2. Pruebas de hipótesis:** Se emplearán pruebas de independencia Chi-cuadrado para evaluar la existencia de diferencias significativas en la distribución de las variables categóricas (como sexo o condición de discapacidad) entre los pacientes que asistieron a sus controles y aquellos que no lo hicieron. Asimismo, se aplicarán pruebas T-Student o análisis de varianza (ANOVA) para comparar las medias de las variables numéricas entre los distintos grupos. Estas pruebas permitirán identificar asociaciones estadísticamente significativas entre las variables estudiadas y la inasistencia, aportando evidencia robusta para el análisis de factores asociados.

3.2.1. Estadísticas Descriptivas de Conjunto de Datos

El conjunto de datos cuenta con **914 registros** (filas) y **14 variables** (columnas). Entre ellas, se identifican cinco variables numéricas de tipo *float64* y nueve variables categóricas de tipo *object*. Las variables categóricas representan atributos como sexo, pertenencia a programas sociales, condición de discapacidad, entre otros.

Las estadísticas descriptivas principales de las variables numéricas se presentan en la siguiente tabla:

	DAÑO POR CARIES	MESES PARA PRÓXIMO CONTROL	distancia_cesfam	EDAD MESES	MESES ULTIMO CONTROL
count	914.000000	914.000000	914.000000	914.000000	914.000000
mean	1.599562	7.299781	2169.329794	64.711160	9.342451
std	2.596957	3.376333	480.255915	36.889191	6.263108
min	0.000000	0.000000	236.267652	7.000000	-1.000000
25%	0.000000	6.000000	1937.982757	30.000000	4.000000
50%	0.000000	6.000000	2022.396997	65.000000	8.000000
75%	2.000000	12.000000	2425.585319	95.000000	14.000000
max	14.000000	12.000000	3681.308050	142.000000	22.000000

Distribución porcentual de variables categóricas

Al calcular la distribución porcentual de las variables categóricas se define lo siguiente:

- Sexo: La distribución está equilibrada, con 50.8% de hombres y 49.2% de mujeres.
- Mejor Niñez: Solo el 1.9% de los pacientes pertenece a este programa.
- SENAME: Un 2.0% de los pacientes pertenece o ha pertenecido al sistema.
- Migrantes: El 1.9% de los pacientes presenta condición de migrante.
- Discapacidad: El 1.4% presenta algún tipo de discapacidad.
- Riesgo clínico: La mayoría de los pacientes se encuentran clasificados como de alto riesgo (70.1%), frente al 20.9% con bajo riesgo.
- Inasistencia: La variable se distribuye casi equitativamente, con 52.1% de pacientes inasistentes y 47.9% de asistentes.

Esto refleja una muestra mayoritariamente compuesta por pacientes sin factores de vulnerabilidad declarados, pero con una alta concentración en la categoría de riesgo clínico alto.

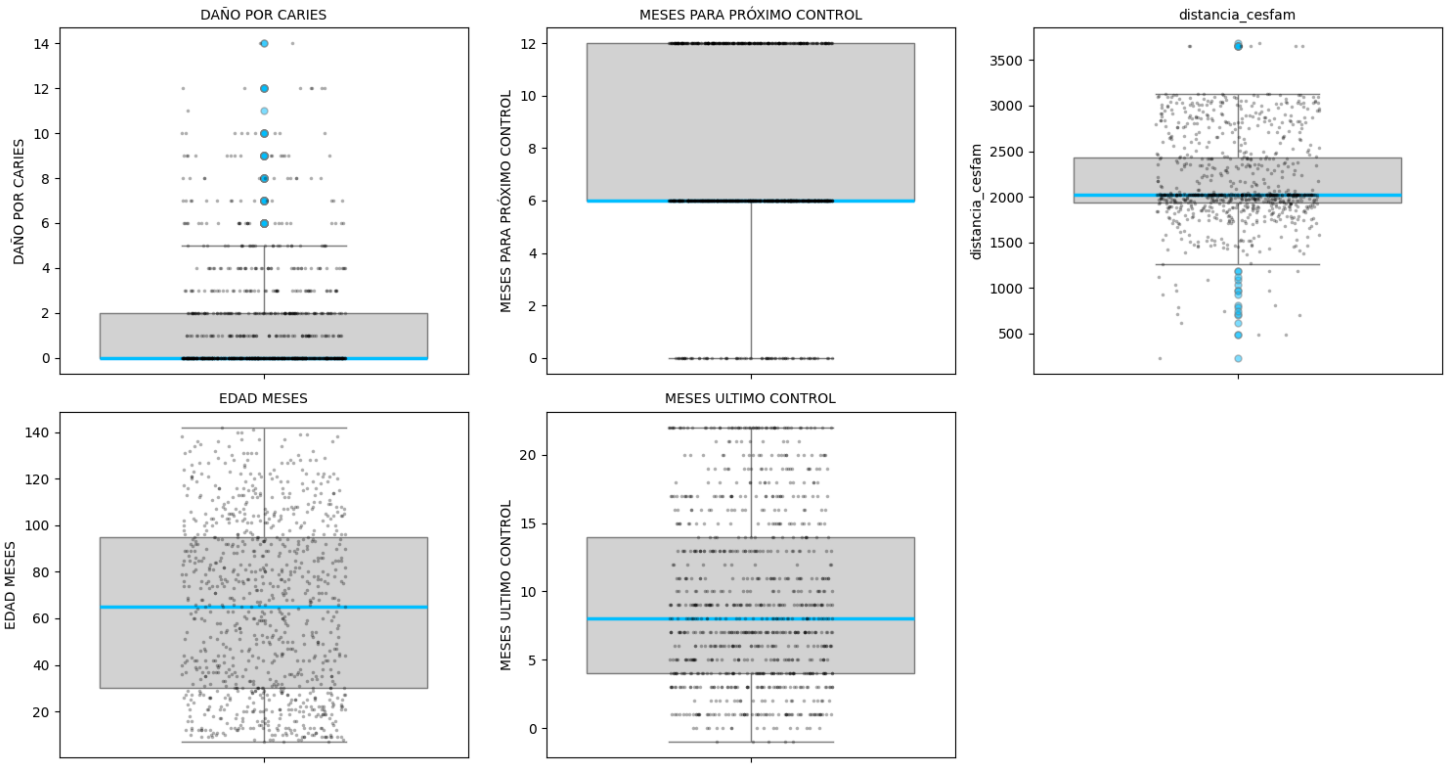
Boxplot de variables numéricas

Al graficar boxplots de las variables numéricas, podemos observar que en DAÑO POR CARIES la mediana está muy cercana al borde inferior de la caja, lo que sugiere una distribución sesgada. Esto indica que la mayoría de los pacientes presenta un daño bajo o nulo, concentrado especialmente en cero, mientras que los valores más altos

corresponden a una minoría con mayor severidad. Esto indica que el gráfico de distribución de esta variable tendrá una cola hacia la derecha.

Los valores detectados como outliers en los boxplots no se alejan significativamente del rango lógico o esperable de las variables. Por tanto, no es necesario tratarlos como valores atípicos.

Gráficos:



Análisis exploratorio de variables numéricas (Pairplot)

Para analizar visualmente las relaciones entre variables numéricas y su comportamiento según la variable INASISTENTE, se construyó un pairplot. Esta herramienta permite observar tanto las distribuciones univariadas como la dispersión bivariada entre variables, diferenciando los grupos por color.

Se identificaron patrones relevantes:

- MESES DESDE ULTIMO CONTROL: La distribución para el grupo de inasistentes está desplazada hacia la derecha, lo que indica que estos pacientes tienden a acumular

más tiempo sin control odontológico. Este es un efecto esperable y de menor significancia, pues la inasistencia depende directamente de esta variable.

- EDAD: El grupo de inasistentes presenta una distribución bimodal, con un segundo pico más pronunciado en edades mayores, lo que sugiere menor adherencia al seguimiento clínico en ese segmento etario.
- DISTANCIA AL CESFAM: La variable presenta una distribución muy similar entre inasistentes y asistentes. Esto sugiere que, por sí sola, no sería un buen predictor de inasistencia, ya que no se observan diferencias significativas entre los grupos.
- DAÑO POR CARIES: Aunque la mayoría de los valores se concentran en la zona baja, los inasistentes presentan valores más altos en promedio.

Estas observaciones sugieren que variables como **edad, meses desde el último control y daño por caries** podrían tener valor predictivo para clasificar a los pacientes según su adherencia al control.

Gráficos:

Análisis de correlación de variables

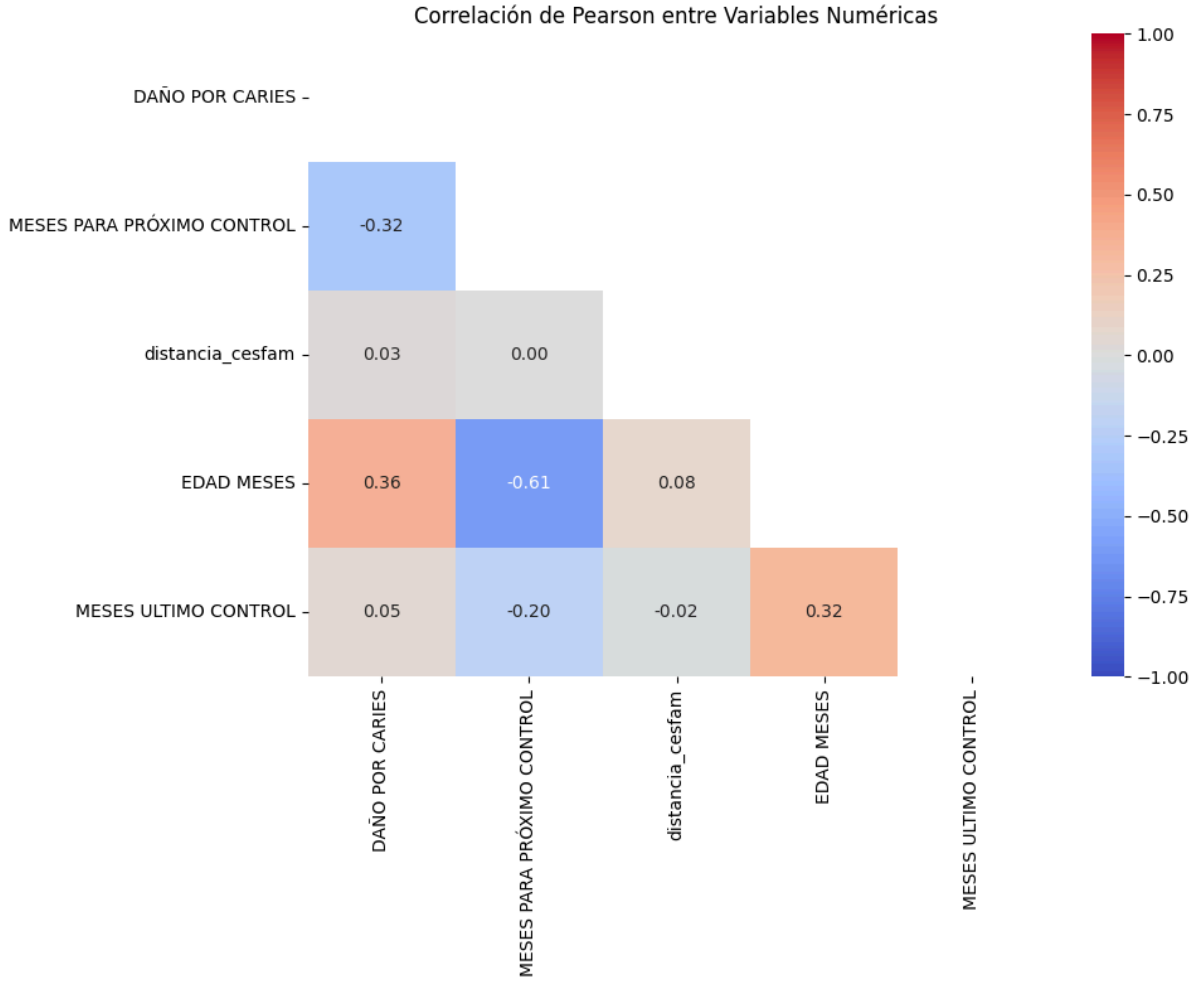
Correlación de Pearson

Correlación de Pearson (Variables Numéricas) mide relaciones lineales entre variables numéricas, con valores entre -1 y 1:

Resultados:

- EDAD MESES y DAÑO POR CARIES (r = 0.36):
    - Correlación positiva moderada Indica que a mayor edad, mayor daño por caries
    - Sugiere acumulación de caries con el paso del tiempo
    - Es consistente con la progresión natural de problemas dentales con el tiempo
- EDAD MESES y MESES PARA PRÓXIMO CONTROL (r = -0.61):
    - Correlación negativa fuerte
    - A mayor edad los niños tienen programados controles más frecuentes
    - Posible protocolo de atención más frecuente para niños mayores
- EDAD MESES y MESES ULTIMO CONTROL (r = 0.32):
    - Correlación positiva moderada
    - Niños mayores tienden a tener controles más espaciados
- DAÑO POR CARIES y MESES PARA PRÓXIMO CONTROL (r = -0.32):
    - Correlación negativa moderada
    - Sugiere un protocolo de seguimiento y refleja un enfoque clínico adecuado en casos graves
- DISTANCIA AL CESFAM:
    - En este dataset muestra correlación positiva débil con EDAD MESES (r = 0.08)
    - No muestra correlaciones significativas con otras variables numéricas
    - Sugiere que la distancia al centro de salud no afecta directamente variables clínicas

Gráficos:



Correlación de Spearman

La correlación de Spearman mide asociaciones por rangos, útil para variables ordinales o categóricas.

Resultados:

La correlación de Spearman mide asociaciones por rangos no necesariamente lineales, útil para variables ordinales o categóricas:

RIESGO e INASISTENTE (r = -0.36):

Correlación negativa moderada, pacientes de alto riesgo tienden a tener menos inasistencias (o viceversa) lo que sugiere mayor adherencia al tratamiento en casos de alto riesgo

MEJOR NIÑEZ e INASISTENTE (r = 0.08):

Correlación positiva débil y una ligera tendencia a mayor inasistencia en niños en el programa

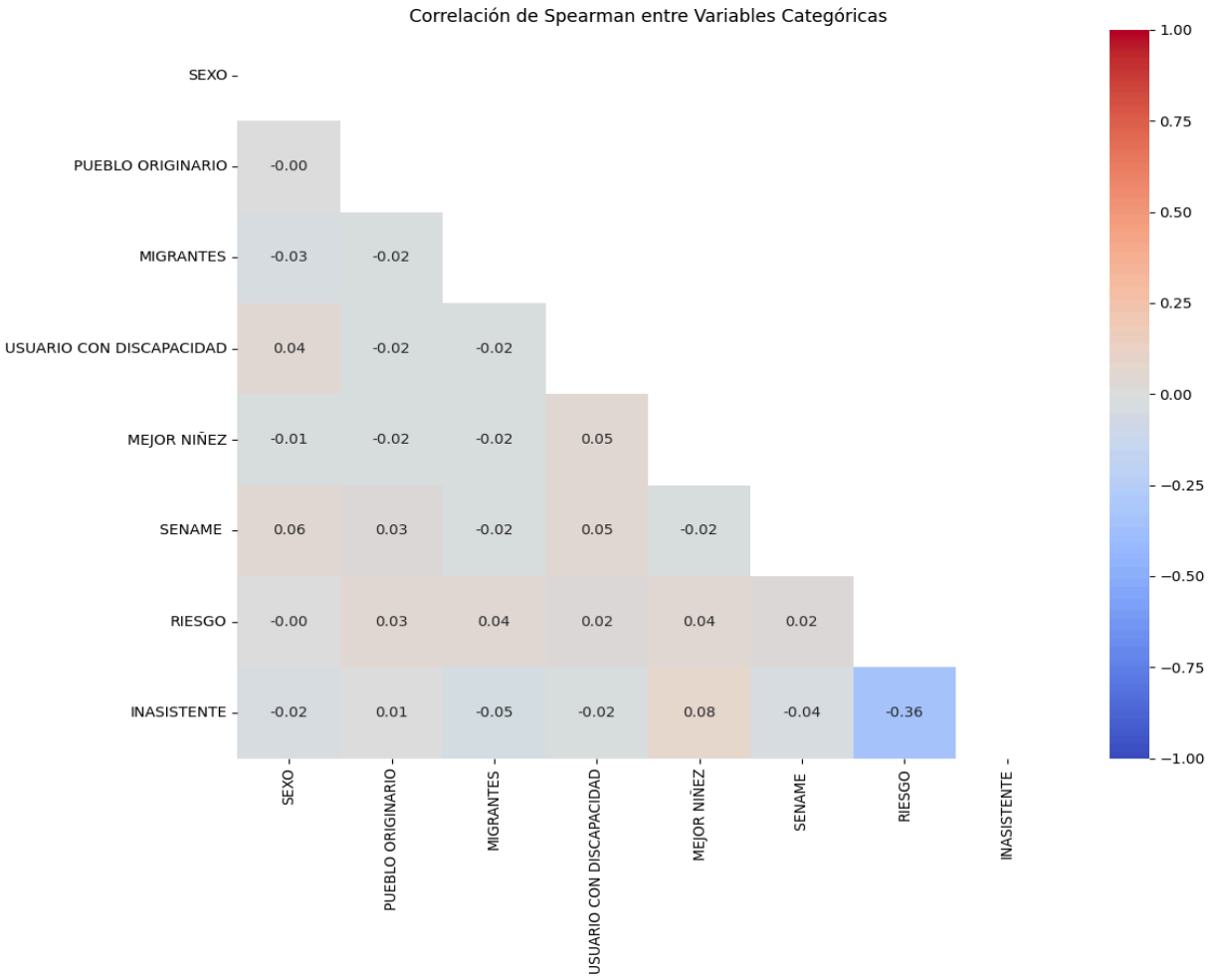
USUARIO CON DISCAPACIDAD y MEJOR NIÑEZ/SENAME (r ≈ 0.05):

Correlaciones positivas muy débiles sugiere cierta superposición entre estos programas sociales

Otras correlaciones:

La mayoría de las correlaciones entre variables categóricas son débiles (< 0.05) y no se observan patrones fuertes de asociación entre la mayoría de las variables categóricas

Gráficos:



Test Exacto de Fisher

El test de Fisher evalúa la asociación entre variables categóricas, siendo significativo si p < 0.05:

Resultados:

MEJOR NIÑEZ vs INASISTENTE (p = 0.00250, OR = 0.28):

Asociación estadísticamente significativa. Odds Ratio de 0.28 indica que los niños en el programa "MEJOR NIÑEZ" tienen aproximadamente 72% menos probabilidades de ser inasistentes, es un posible efecto positivo del programa en adherencia a controles

RIESGO vs INASISTENTE (p < 0.0001, OR = 0.18):

Asociación fuertemente significativa, pacientes de alto riesgo tienen 82% menos probabilidades de ser inasistentes, confirma la correlación observada en Spearman y sugiere buen funcionamiento del sistema de seguimiento para casos más severos

Otras comparaciones:

No se encontraron otras asociaciones significativas entre variables categóricas, esto sugiere independencia entre la mayoría de estos factores categóricos

Principales hallazgos de los análisis de correlación

1. **Factores de edad y progresión de caries**: La edad está moderadamente correlacionada con daño por caries (r = 0.36), evidenciando acumulación de patología dental con el tiempo, lo que refuerza la importancia de la prevención temprana.

2. **Efectividad del sistema de gestión de riesgo**: El nivel de riesgo muestra correlaciones fuertes con los protocolos de atención (r = 0.84 con frecuencia de controles) y es determinante en el manejo clínico. La fuerte asociación inversa entre riesgo e inasistencia (OR = 0.18) sugiere un sistema efectivo de seguimiento para casos críticos.

3. **Programas sociales y asistencia**: El programa "MEJOR NIÑEZ" muestra un efecto protector significativo contra inasistencias (OR = 0.28), demostrando el impacto positivo de intervenciones sociales en la adherencia al tratamiento dental.

4. **Perfil de pacientes inasistentes**: Se identifica un perfil claro de pacientes inasistentes caracterizado por mayor edad, más daño por caries y períodos más largos sin controles (r = 0.60 con MESES ULTIMO CONTROL). Este hallazgo permite focalizar estrategias de recuperación.

5. **Equidad en servicios dentales**: La ausencia de correlaciones significativas entre variables clínicas y factores como sexo, pertenencia a pueblos originarios o distancia al centro de salud, sugiere equidad en el acceso y calidad de atención dental.

6. **Población migrante**: Los migrantes tienden a ser pacientes más jóvenes (r = -0.08), lo que podría requerir adaptaciones en los protocolos preventivos para esta población específica.

7. **Impacto de SENAME**: Los niños vinculados a SENAME muestran patrones distintos de edad y control, con tendencia a ser más jóvenes y tener controles más recientes, evidenciando posiblemente un seguimiento más intensivo para esta población vulnerable.

Estos hallazgos proporcionan información valiosa para:

- Optimizar protocolos de seguimiento y prevención
- Fortalecer programas sociales que demuestran impacto positivo en adherencia
- Diseñar estrategias específicas para recuperación de pacientes inasistentes
- Mantener y mejorar la equidad en el acceso a servicios dentales

La toma de decisiones basada en estos datos permitiría una gestión más eficiente de recursos y mejores resultados en salud dental infantil.

3.2.2. Pruebas de Hipótesis

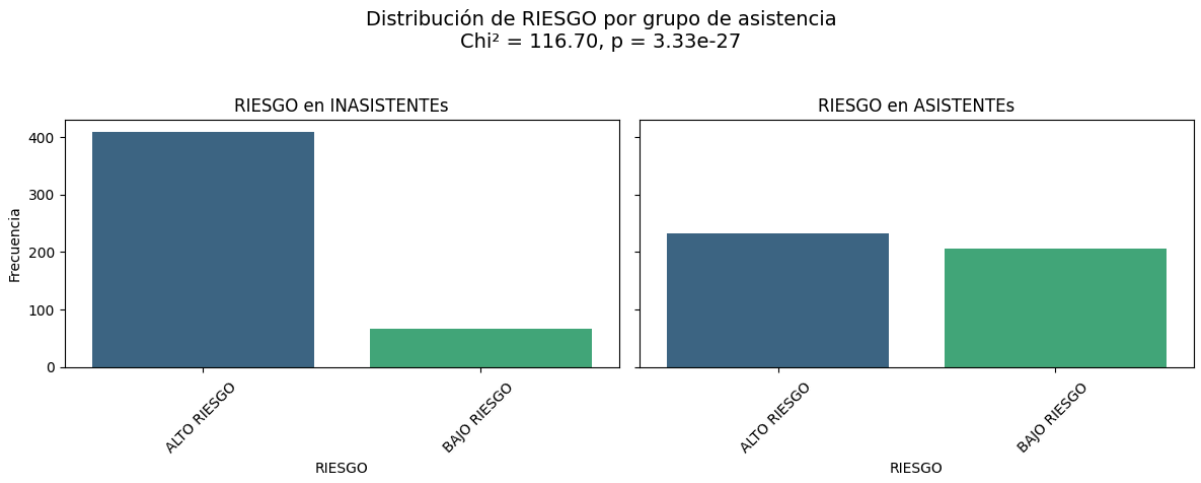
Chi cuadrado

Se llevaron a cabo pruebas de independencia Chi-cuadrado para evaluar la asociación entre las clases de la variable objetivo (INASISTENTE/ASISTENTE) y el resto de las variables categóricas. Únicamente se consideraron para su presentación y análisis aquellos resultados que alcanzaron significancia estadística ($p < 0.05$), los cuales fueron posteriormente graficados.

Resultados:

El análisis revela que los pacientes clasificados como de ALTO RIESGO presentan una mayor tendencia a la inasistencia a los controles odontológicos, mientras que aquellos identificados como de BAJO RIESGO muestran una mayor adherencia a las citas programadas. Este patrón sugiere que el nivel de riesgo, determinado según la aplicación de la Pauta Cero, podría estar actuando como un factor limitante para la asistencia, lo que evidencia la existencia de barreras adicionales precisamente en los grupos que más requieren seguimiento clínico.

Gráficos:



T-Student

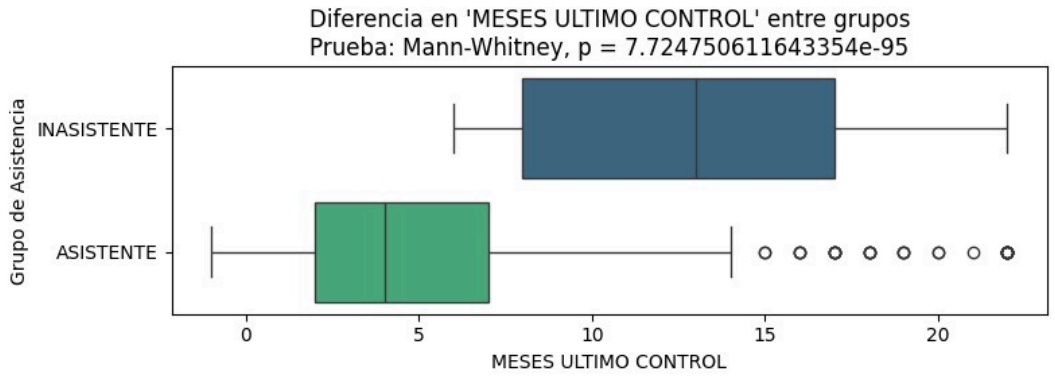
Realizamos una prueba T-Student comparando las clases de la variable objetivo (INASISTENTE/ASISTENTE). Previamente, se verifica la normalidad de los datos mediante el test de Shapiro-Wilk (considerando normalidad si p > 0.05) y la homogeneidad de varianzas mediante el test de Levene (p > 0.05). Si alguna de estas condiciones no se cumple, se utiliza la prueba no paramétrica U de Mann-Whitney.

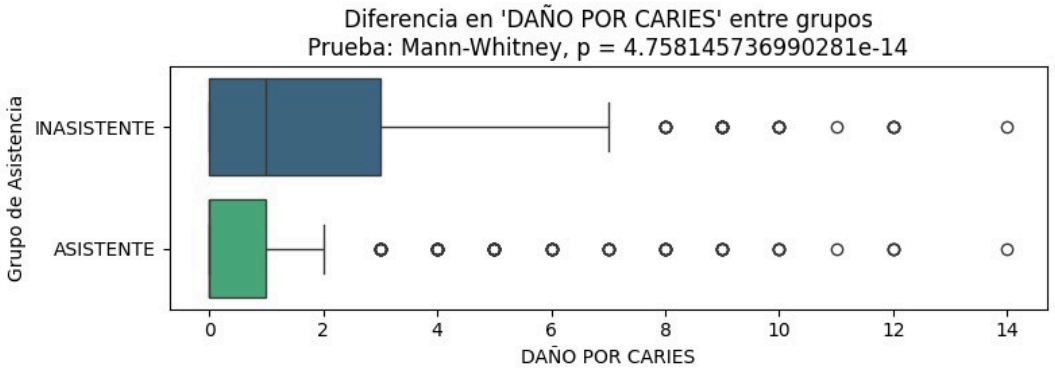
Resultados:

El análisis comparativo de la variable 'MESES ULTIMO CONTROL' evidencia una diferencia estadísticamente significativa, sin embargo, la inasistencia se determina precisamente según si han pasado más de 6 o 12 meses desde el último control, por lo que no tiene mayor relevancia.

Asimismo, los pacientes clasificados como INASISTENTES presentan un daño por caries significativamente superior en comparación con los ASISTENTES, lo que sugiere que la falta de asistencia a los controles está vinculada a peores resultados en salud bucal, reflejados en un mayor número de piezas dentales afectadas por caries.

Gráficos:





Análisis Cruzado (Correlación Point-Biserial)

Este análisis evalúa correlaciones entre variables categóricas (binarias) y numéricas:

Resultados: Asociaciones significativas

1. **MIGRANTES vs EDAD MESES** (r = -0.08, p = 0.0145):
    - Correlación negativa débil pero significativa
    - Los migrantes tienden a ser ligeramente más jóvenes en la muestra
    - Posible reflejo de patrones demográficos en la población migrante

2. **MEJOR NIÑEZ vs MESES ÚLTIMO CONTROL** (r = 0.13, p = 0.0001):
    - Correlación positiva significativa
    - Niños en este programa tienen intervalos más largos desde su último control
    - Posible indicio de menor acceso previo a servicios dentales

3. **SENAME vs Variables Numéricas**:
    - Correlación significativa negativa con EDAD MESES (r = -0.07, p = 0.045)
    - Correlación significativa negativa con MESES ULTIMO CONTROL (r = -0.08, p = 0.017)
    - Usuarios de SENAME tienden a ser más jóvenes y con controles más recientes
    - Sugiere atención focalizada para esta población vulnerable

4. **RIESGO vs Variables Numéricas**:
   Correlaciones altamente significativas con cuatro variables numéricas:
    - EDAD MESES (r = -0.45, p < 0.0001): Mayor riesgo en niños mayores
    - DAÑO POR CARIES (r = -0.38, p < 0.0001): Mayor riesgo asociado a más caries
    - MESES PARA PRÓXIMO CONTROL (r = 0.84, p < 0.0001): Correlación muy fuerte; alto riesgo implica controles más frecuentes
    - MESES ÚLTIMO CONTROL (r = -0.13, p = 0.0001): Alto riesgo asociado a controles más recientes
    - La variable RIESGO está profundamente integrada con indicadores clínicos clave

5. **INASISTENTE vs Variables Numéricas**:
    - Correlaciones significativas con todas las variables clínicas:
        - EDAD MESES (r = 0.15, p < 0.0001): Mayor inasistencia en niños mayores
        - DAÑO POR CARIES (r = 0.17, p < 0.0001): Más inasistencia asociada a más caries
        - MESES PARA PRÓXIMO CONTROL (r = -0.14, p < 0.0001): Inasistentes programados para controles más frecuentes
        - MESES ÚLTIMO CONTROL (r = 0.60, p < 0.0001): Correlación fuerte; inasistentes tienen intervalos mucho más largos desde último control

- Perfil claro de pacientes inasistentes: mayor edad, más daño dental y mayor tiempo desde último control

Variables Sin Asociaciones Significativas:

SEXO, **PUEBLO ORIGINARIO** y **distancia_cesfam**: No muestran correlaciones significativas con ninguna variable

3.3. Modelos Predictivos

En la siguiente etapa del análisis, se implementarán modelos de aprendizaje no supervisado para comprender las variables significativas relacionadas con la inasistencia. El objetivo es identificar patrones complejos y no lineales que puedan estar presentes en los datos y que no hayan sido capturados por el análisis estadístico tradicional.

Se considerará el siguiente enfoque:

1. **Clusterización (K-Means, DBSCAN, K-Medoids):** Los métodos no supervisados permitirán segmentar a los pacientes en grupos con características similares, lo que puede ayudar a identificar perfiles de alto riesgo y diseñar intervenciones personalizadas.
2. **Validación:** Para asegurar la robustez de los modelos, se utilizará validación con métricas como la Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index.

La implementación de estos modelos permitirá anticipar la inasistencia y optimizar la gestión de citas, contribuyendo a una mejor cobertura y eficiencia del Programa CERO.

3.3.1 Clusterización

K-means

La aplicación del algoritmo K-means permitió segmentar la población en grupos homogéneos en función de variables numéricas clave como edad en meses, daño por caries, meses desde el último control y distancia al CESFAM.

Para la identificación de grupos del conjunto de datos, primero revisamos el Silhouette Coefficient para resolver la cantidad óptima de clusters.

Al analizar los diagramas de silhouette para valores de k entre 2 y 5, se observa que k = 3 ofrece el mejor equilibrio entre cohesión interna y diferenciación entre grupos.

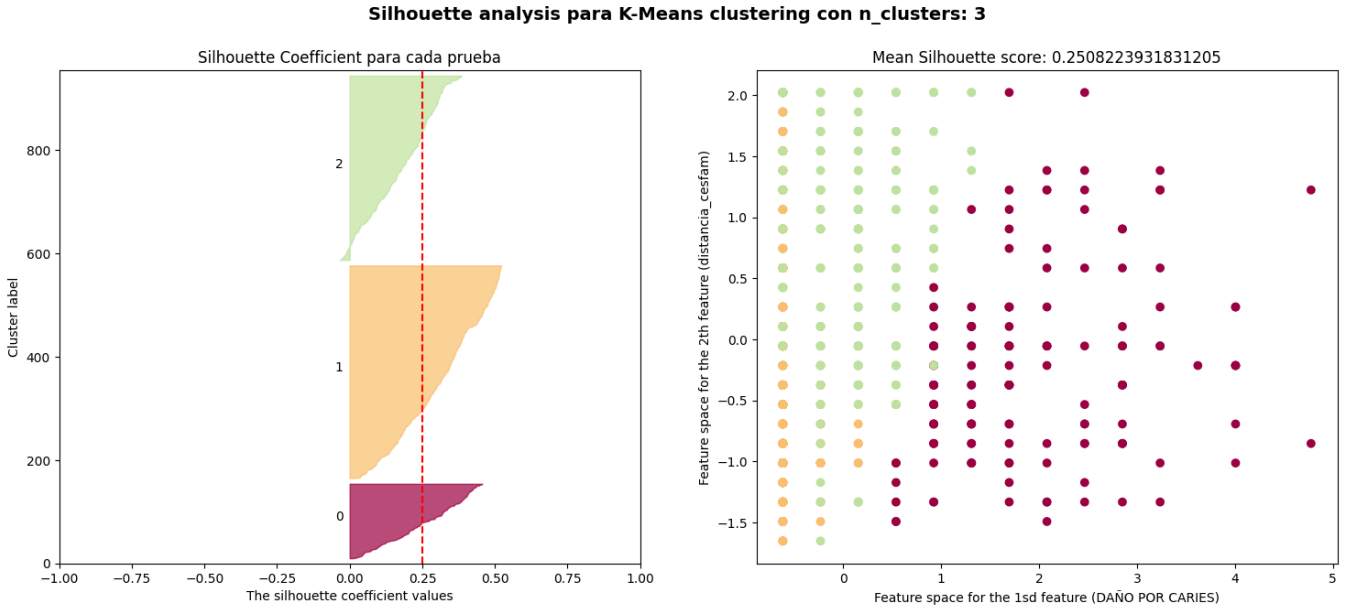
Aunque k = 2 presenta el valor promedio de silhouette más alto (0.288), su segmentación es demasiado general y agrupa puntos con valores cercanos a cero o negativos, lo que indica menor calidad en la asignación.

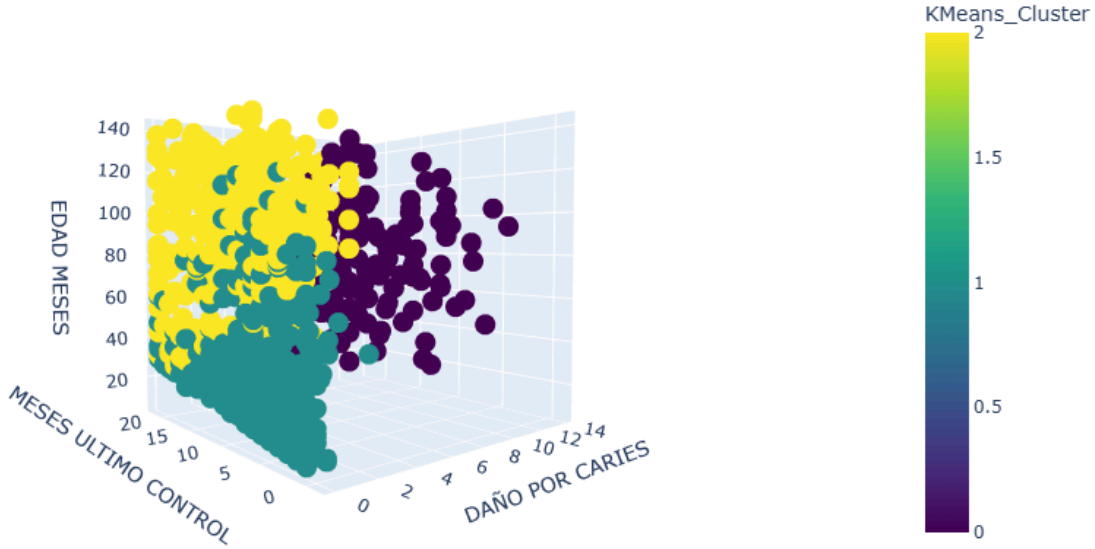
En cambio, con k = 3, el silhouette promedio se mantiene alto (0.278), y los tres clusters muestran una estructura más clara y coherente, con menos puntos mal asignados y formas de grupo más definidas.

Los resultados para k = 4 y k = 5 muestran una caída en cohesión y la aparición de grupos con estructuras débiles. Por ello, se utilizará k = 3 para continuar el análisis de segmentación.

Generamos el modelo de K-means y lo ajustamos con los datos del dataset normalizados.

Gráficos:



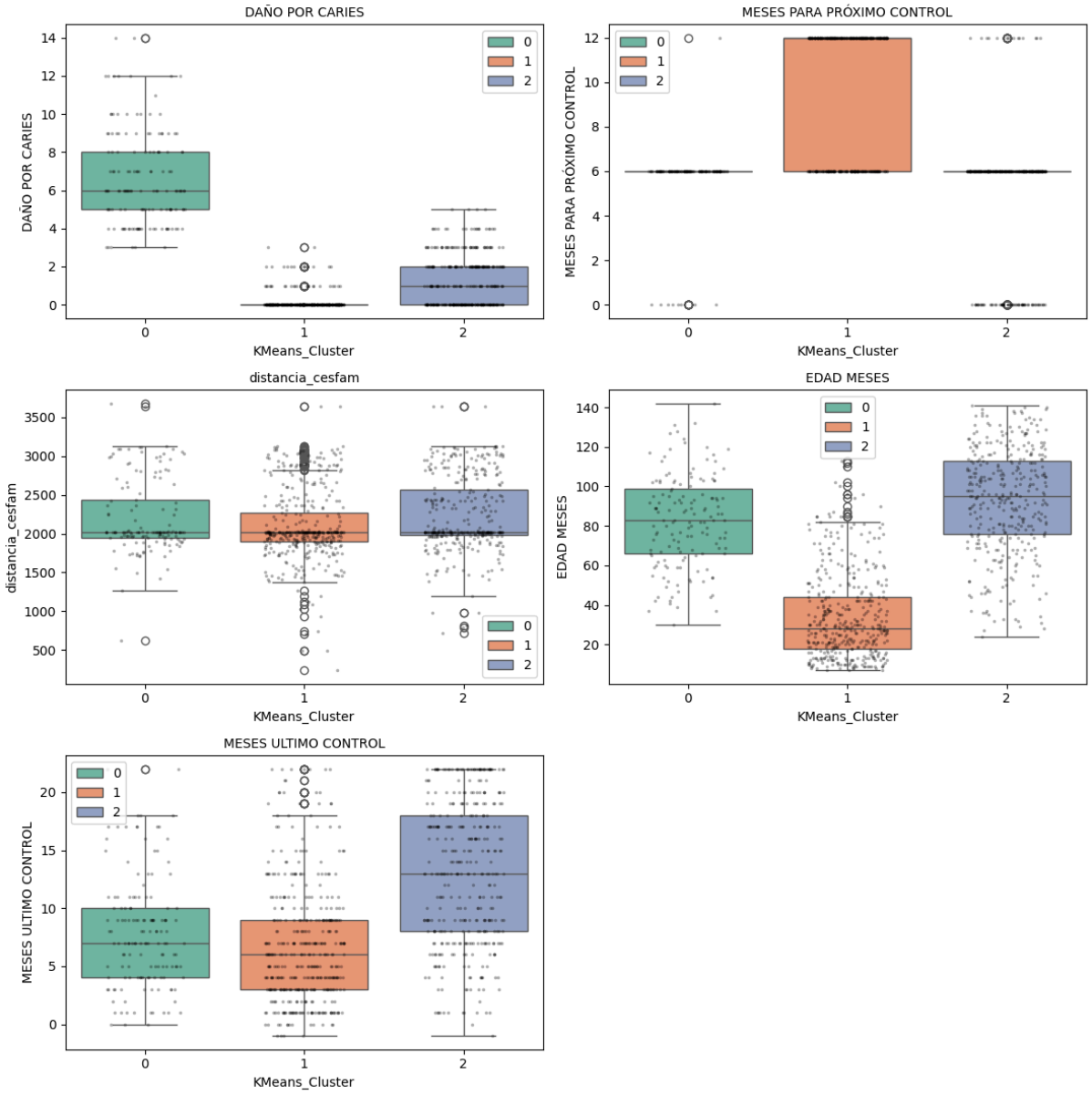


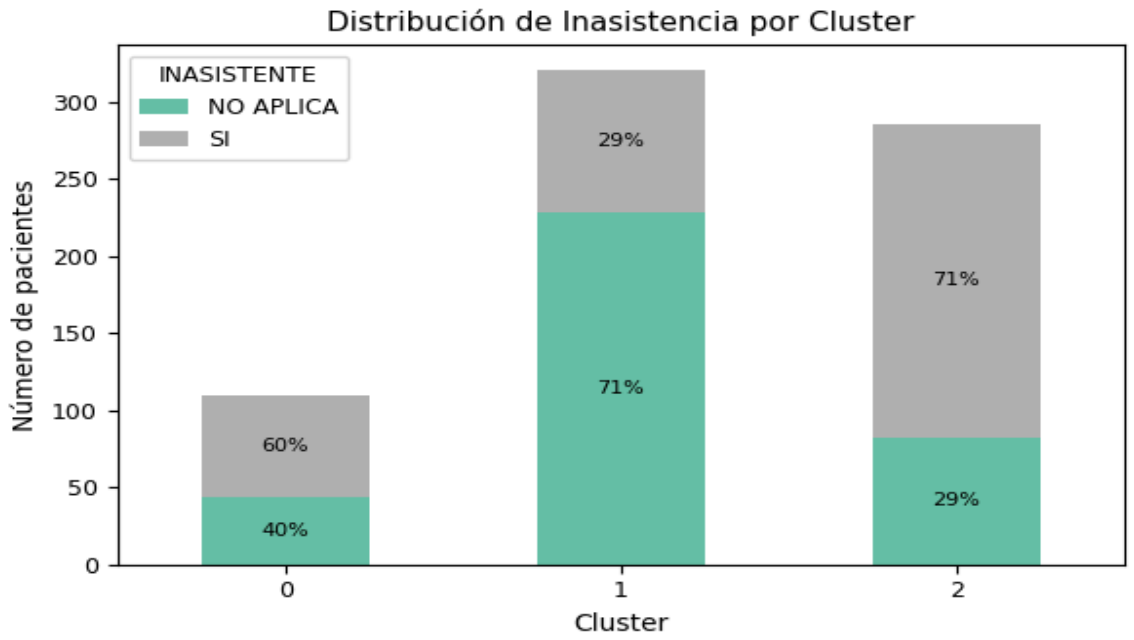
Descripción de los clusters

- Cluster 0:
    - Pacientes de edad intermedia, con mayor daño por caries y una leve tendencia a la inasistencia. Este grupo concentra la mayor necesidad de intervención clínica y labores de rescate mayores.
- Cluster 1:
    - Pacientes de menor edad, con bajo daño por caries y controles relativamente recientes. Este grupo mostró la mayor adherencia a los controles y un bajo riesgo clínico.
- Cluster 2:
    - Pacientes mayores dentro del rango etario, con daño por caries moderado y mayor cantidad de meses desde su último control. Se observa una proporción significativa de inasistentes, posiblemente por una menor percepción de riesgo o barreras culturales.

	DAÑO POR CARIES			MESES PARA PRÓXIMO CONTROL			distancia_cesfam			EDAD MESES			MESES ULTIMO CONTROL			Cantidad de Puntos
	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	
KMeans_Cluster																
0	3.0	14.0	6.60	0.0	12.0	5.71	619.07	3681.31	2202.56	30.0	142.0	82.32	0.0	22.0	7.88	145
1	0.0	3.0	0.13	6.0	12.0	9.64	236.27	3648.22	2096.78	7.0	113.0	34.16	-1.0	22.0	6.74	412
2	0.0	5.0	1.27	0.0	12.0	5.24	719.77	3648.22	2239.56	24.0	141.0	92.82	-1.0	22.0	12.94	357

Gráficos:


DAÑO POR CARIES
DAÑO POR CARIES
KMeans_Cluster
MESES PARA PRÓXIMO CONTROL
MESES PARA PRÓXIMO CONTROL
KMeans_Cluster
distancia_cesfam
distancia_cesfam
KMeans_Cluster
EDAD MESES
EDAD MESES
KMeans_Cluster
MESES ULTIMO CONTROL
MESES ULTIMO CONTROL
KMeans_Cluster
0
1
2



DBSCAN

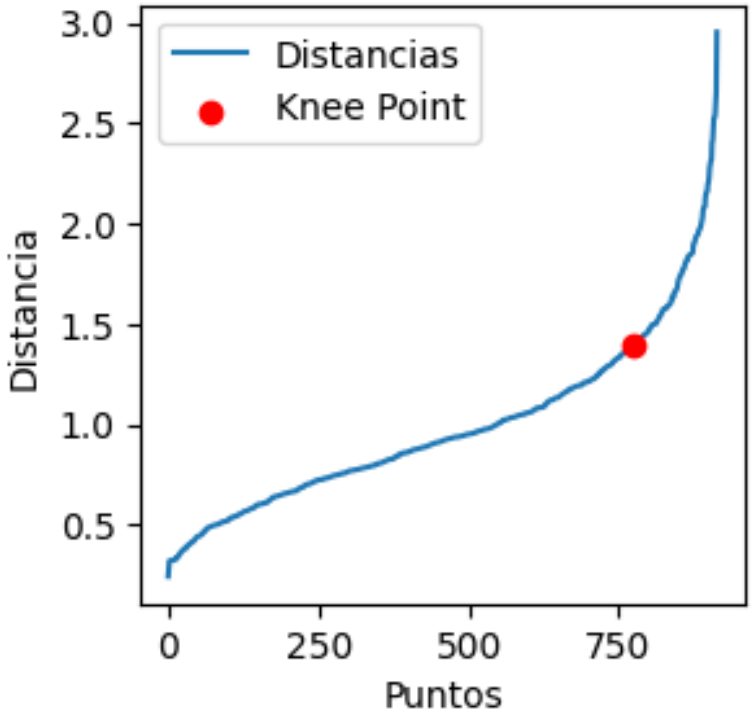
El algoritmo **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise), adecuado para identificar grupos de densidad variable y detectar outliers, reveló una estructura diferente en la población.

Previo a definir el modelo, se utilizó el método de **KneeLocator** para estimar el valor óptimo del parámetro `eps`, obteniendo un punto de quiebre (knee point) en **1.39**. Para el parámetro `min_samples`, se consideró la recomendación de utilizar aproximadamente **el doble del número de dimensiones** del conjunto de datos. En este caso, al trabajar con cinco dimensiones, se probó inicialmente con `min_samples = 10`.

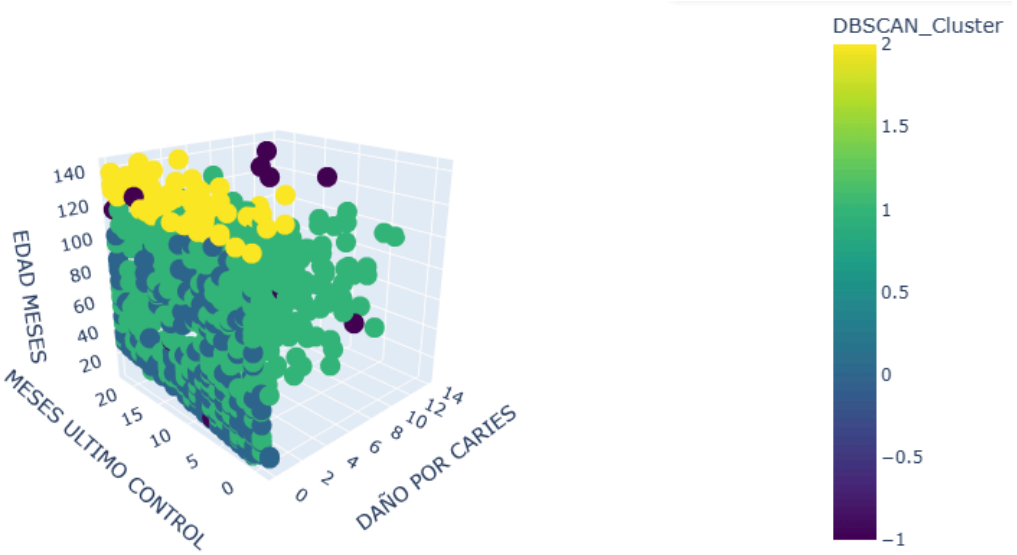
A partir de esos valores iniciales, se realizaron iteraciones ajustando `eps` y `min_samples`, observando cómo influían en la cantidad y forma de los clusters. Finalmente, se seleccionó `eps = 1.53` y `min_samples = 10`, ya que estos parámetros ofrecieron:

- Una **mejor separación** entre grupos
- Una cantidad **aceptable de outliers**
- Una estructura coherente con la distribución tridimensional de las variables analizadas

La visualización 3D con las variables **Edad en meses**, **Meses desde el último control** y **Daño por caries** permite apreciar la distribución de los clusters encontrados y los posibles outliers detectados por el algoritmo.




```
### Entrenamos el algoritmo DBSCAN
### Probamos desde min_samples 10 y distintos valores de eps desde el que nos entrega el KneeLocator
dbscan = DBSCAN(eps=1.53, min_samples=10)
wp_df_cluster['DBSCAN_Cluster'] = dbscan.fit_predict(wp_df_scaled)
```

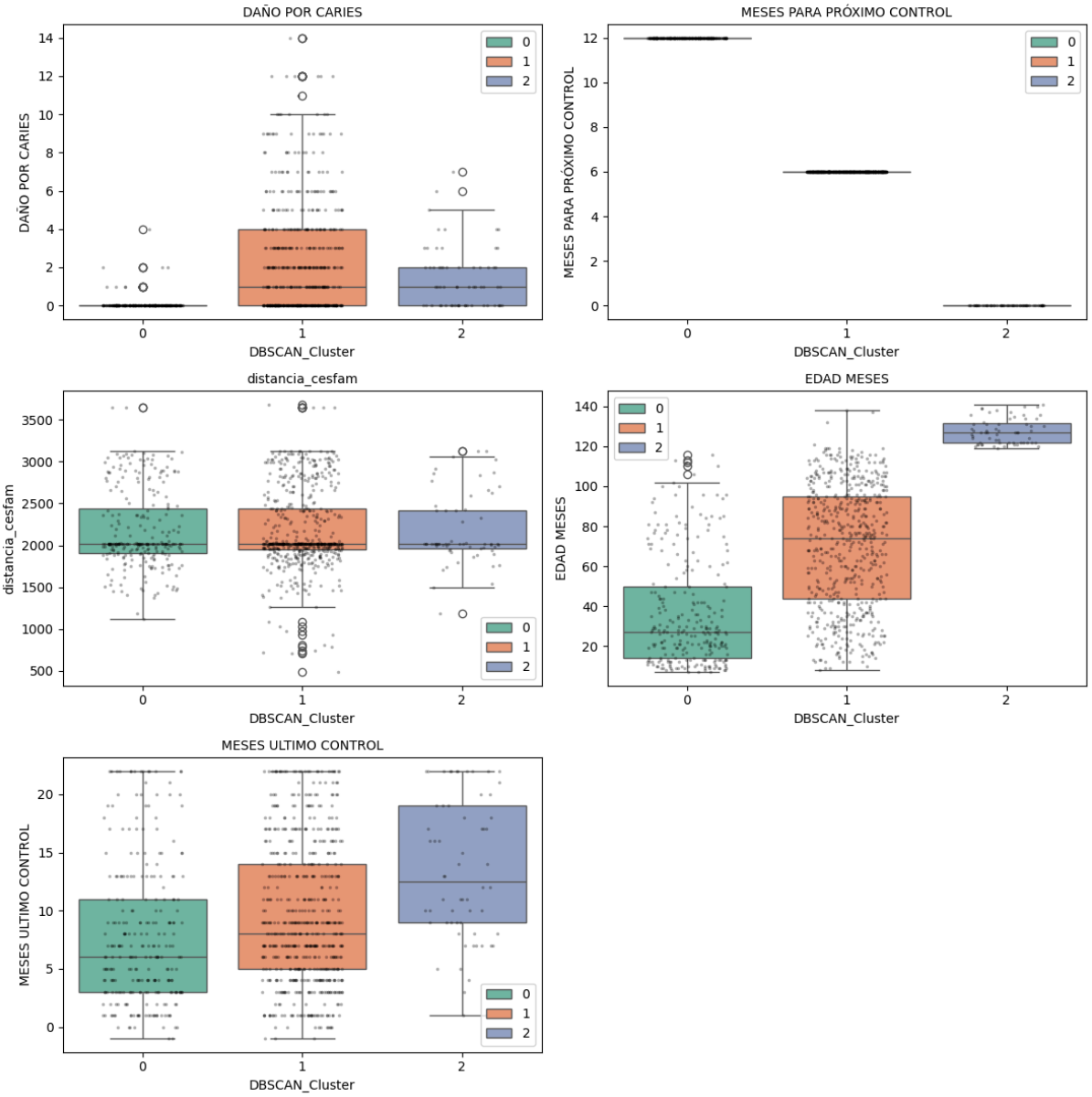


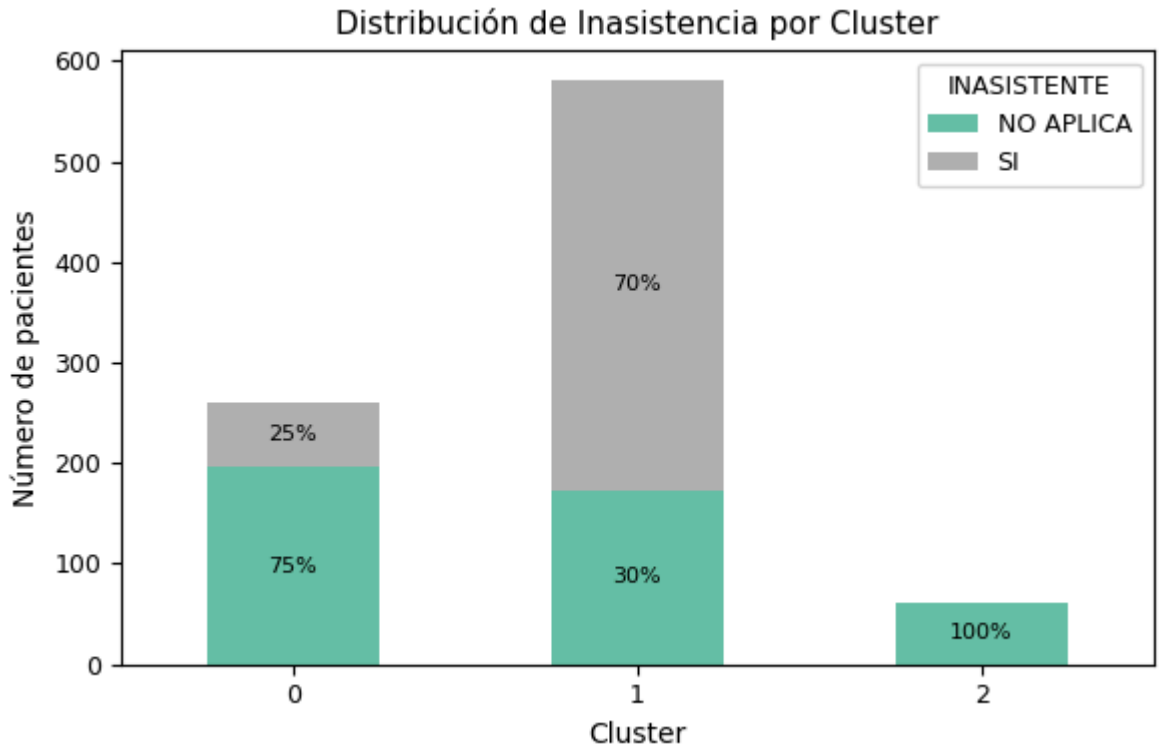
Descripción de los clusters

- Cluster principal (Cluster 1):
    - Agrupa a la mayoría de los pacientes con características clínicas y demográficas promedio, es decir, niños de edad media, bajo a moderado daño por caries y controles recientes. Este grupo presenta una **alta tasa de inasistencia a la cita odontológica (70%)**, a pesar de contar con buena accesibilidad y riesgo clínico elevado. Esto sugiere que existen barreras no clínicas que afectan la adherencia, como desmotivación, dificultades familiares u otros factores psicosociales.
- Subgrupos de densidad baja (Clusters 0 y 2):
    - DBSCAN identificó subgrupos periféricos compuestos por pacientes con valores extremos, como aquellos con daño por caries muy elevado o controles extremadamente espaciados. Estos subgrupos corresponden a casos atípicos que podrían requerir estrategias de seguimiento individualizado.
    - **Cluster 0:** Compuesto principalmente por **niños pequeños con daño por caries prácticamente nulo y controle**s relativamente recientes. A pesar de su perfil clínico favorable, este grupo presenta una **adherencia intermedia (75% asistencia)**. Esto puede estar influenciado por la dependencia de cuidadores para asistir a los controles.
    - **Cluster 2**: Representa a pacientes de mayor edad, con bajo daño por caries y controles más espaciados en el tiempo. Sin embargo, presentan una **adherencia perfecta (100% asistencia)**. Esto sugiere un grupo autónomo y estable
- Outliers:
    - El método detectó un pequeño número de pacientes que no se agrupan con el resto, caracterizados por combinaciones inusuales de edad, daño por caries y distancia al CESFAM. Estos pacientes pueden representar situaciones sociales o clínicas excepcionales.

	DAÑO POR CARIES			MESES PARA PRÓXIMO CONTROL			distancia_cesfam			EDAD MESES			MESES ULTIMO CONTROL			Cantidad de Puntos
	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	
DBSCAN_Cluster																
-1	0.0	12.0	5.30	0.0	12.0	5.4	236.27	3095.04	1519.75	14.0	142.0	97.80	1.0	22.0	12.60	10
0	0.0	4.0	0.06	12.0	12.0	12.0	1123.30	3648.22	2175.34	7.0	116.0	36.48	-1.0	22.0	7.94	261
1	0.0	14.0	2.24	6.0	6.0	6.0	487.01	3681.31	2177.55	8.0	138.0	70.10	-1.0	22.0	9.49	581
2	0.0	7.0	1.47	0.0	0.0	0.0	1189.50	3126.64	2171.76	119.0	141.0	127.77	1.0	22.0	13.37	62

Gráficos:





Comparación entre Clusterings K-Means y DBSCAN

	Métrica	K-Means	DBSCAN
0	Silhouette Score	0.250822	0.219701
1	Davies-Bouldin Index	1.482943	1.437789
2	Calinski-Harabasz Index	290.147986	209.225307

Al evaluar el desempeño de ambos algoritmos mediante métricas de validación interna, se observan resultados similares en términos generales, aunque con ligeras ventajas para K-Means:

- **Silhouette Score**: Ambos métodos presentan una estructura de agrupamiento débil pero existente, lo que indica que los datos tienen cierta organización natural, aunque no muy marcada. K-Means obtiene un puntaje levemente superior, lo que sugiere una cohesión algo mejor entre los puntos de cada cluster.
- **Davies-Bouldin Index**: Ambos algoritmos muestran valores similares, pero DBSCAN obtiene un índice levemente menor, lo que sugiere una separación ligeramente más clara entre los clusters, con menos superposición interna.

- **Calinski-Harabasz Index**: K-Means muestra un valor más alto, lo que indica una mejor relación entre la dispersión inter-cluster y la cohesión interna. Esto respalda la validez de la estructura de agrupamiento generada por este método.

Aunque ambos algoritmos capturan cierta estructura en los datos, K-Means muestra una performance superior según las tres métricas evaluadas. Sin embargo, DBSCAN aporta valor al identificar outliers y estructuras de densidad variable que K-Means no detecta.

Clustering con K-Medoids

Para realizar el análisis de agrupamiento, se utilizó el algoritmo K-Medoids, una variante más robusta de K-Means. Ambos métodos buscan agrupar observaciones similares, pero mientras K-Means utiliza promedios como centros de los clusters (centroides), K-Medoids selecciona observaciones reales del conjunto de datos como centros (medoides), lo que lo hace menos sensible a valores extremos.

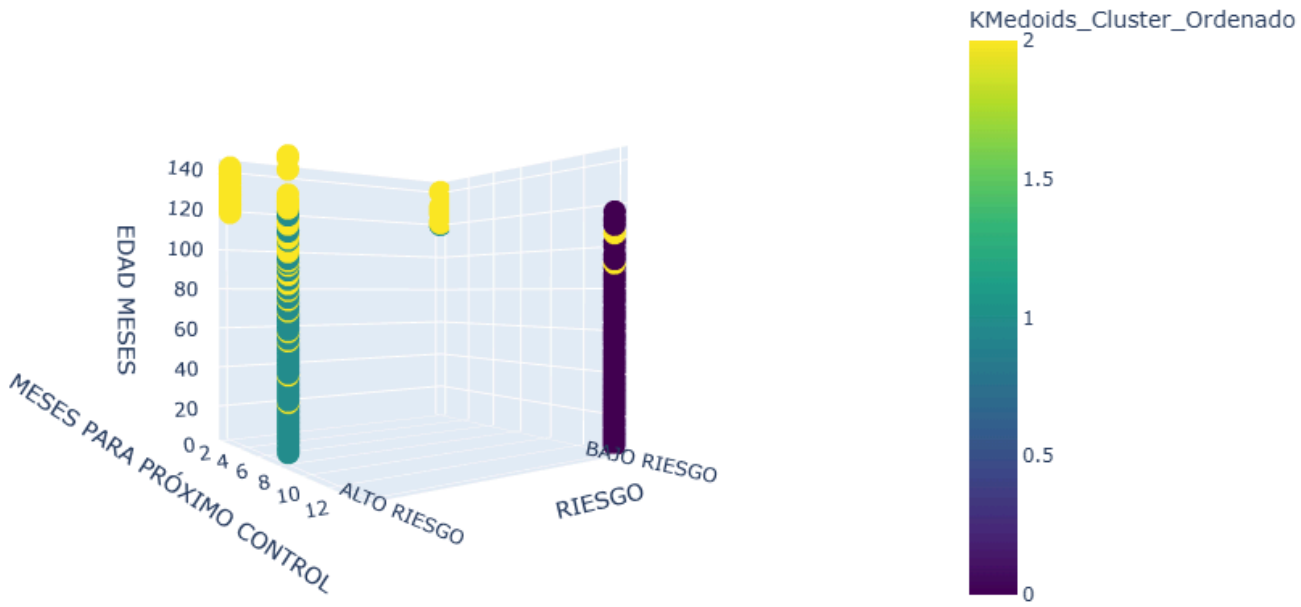
Se eligió este enfoque porque permite combinar variables numéricas (previamente normalizadas) y categóricas binarias (codificadas como 0 y 1) en una misma matriz, conservando información relevante de ambos tipos. Esto permitió identificar perfiles diferenciados de pacientes sin perder riqueza en los datos.

Para la definición del modelo se eligió la misma cantidad de clusters óptimos determinada previamente con K-Means, considerando la similitud entre ambos métodos y el objetivo de mantener una cantidad manejable de grupos al momento de caracterizar a los pacientes.

```
### Elegimos índices iniciales aleatorios como medoids
iniciales = np.random.choice(range(len(data_kmedoids)), size=3, replace=False)

### Definimos métrica euclideana
metric = distance_metric(type_metric.EUCLIDEAN)

### Creamos el modelo y lo entrenamos
kmedoids_instance = kmedoids(data_kmedoids, iniciales.tolist(), metric=metric)
kmedoids_instance.process()
```

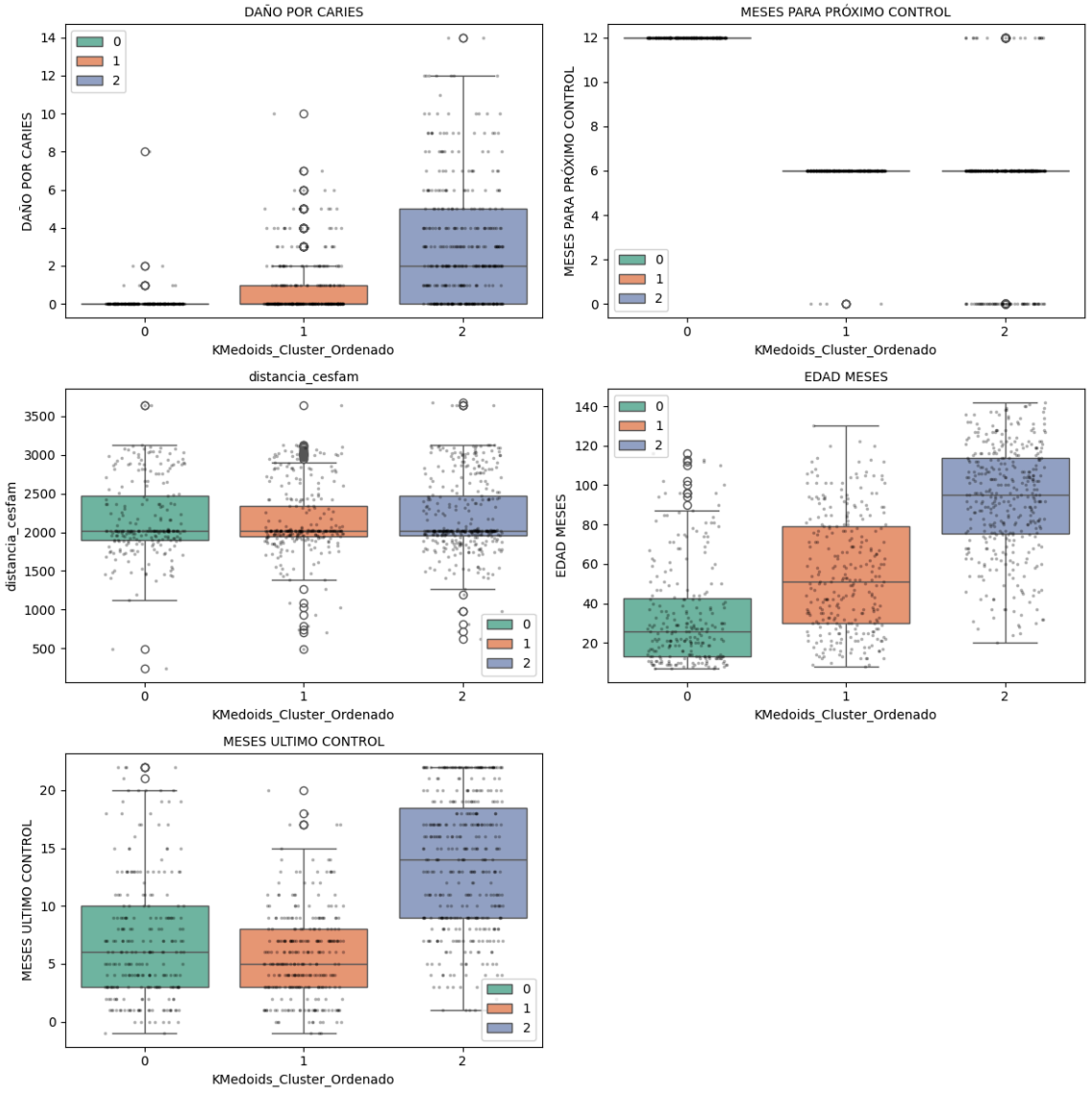


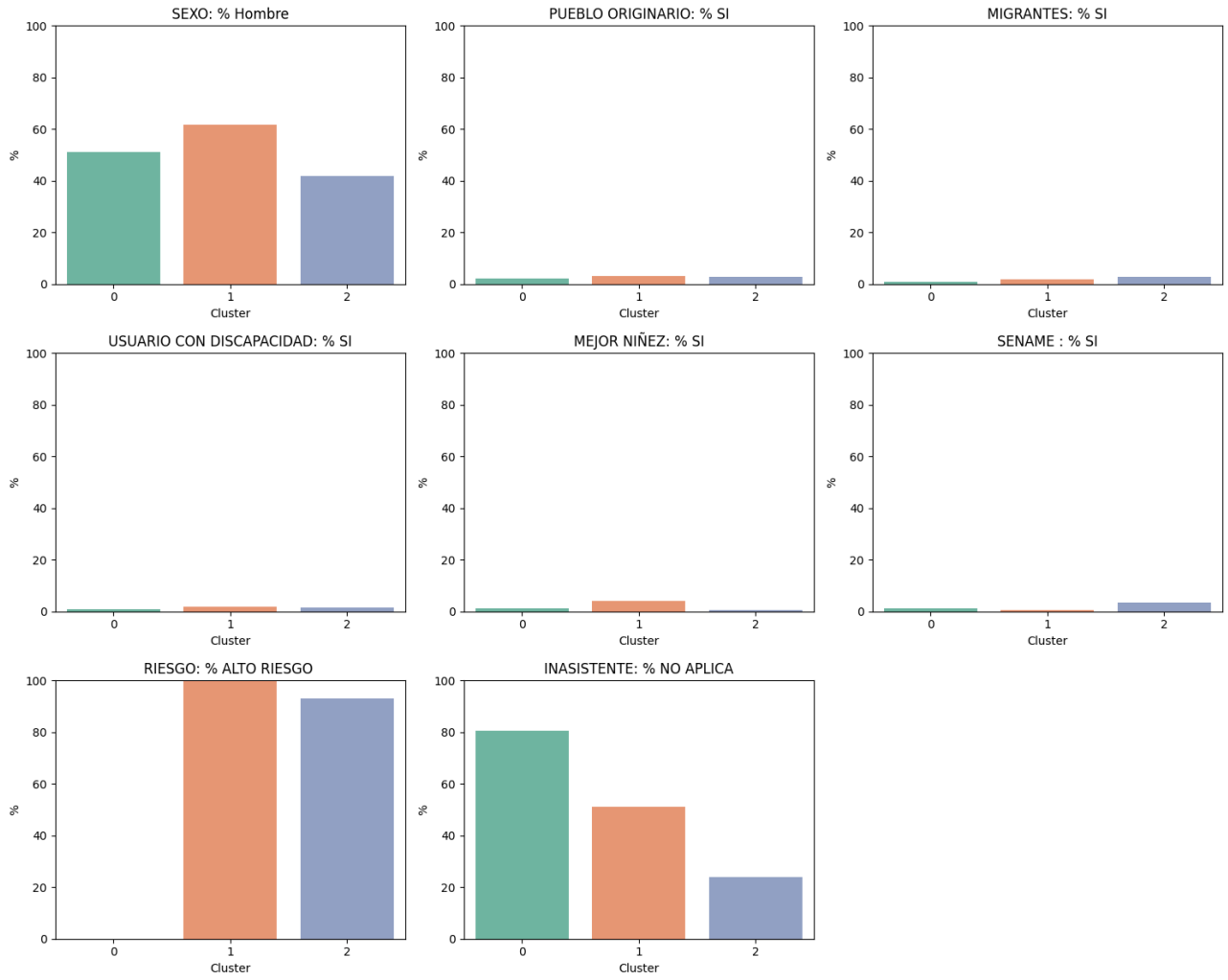
Descripción de los clusters

- Cluster 0:
    - Agrupa principalmente a niños pequeños (edad promedio de 33 meses), con daño por caries bajo y controles recientes. Presenta la **mayor tasa de asistencia a controles odontológicos (80%)**, lo que sugiere un buen compromiso por parte de cuidadores y un perfil preventivo
- Custer 1
    - Corresponde a un grupo de edad media, con daño por caries moderado y controles recientes. Se observa una **asistencia media (51%)**, a pesar de que casi la totalidad de los pacientes están clasificados como de alto riesgo.
- Cluster 2:
    - Reúne a los pacientes de mayor edad, con el mayor daño por caries y los controles más espaciados. Es el grupo con **menor adherencia a la atención (76% de inasistencia)**, siendo la tasa de inasistencia más alta de todos los grupos. Este cluster representa el caso más crítico, tanto por riesgo clínico como por comportamiento asistencial.

	DAÑO POR CARIES			MESES PARA PRÓXIMO CONTROL			distancia_cesfam			EDAD MESES			MESES ULTIMO CONTROL			Cantidad de Puntos
	min	max	mean	min	max	mean	min	max	mean	min	max	mean	min	max	mean	
KMedoids_Cluster_Ordenado																
0	0.0	8.0	0.07	12.0	12.0	12.00	236.27	3648.22	2169.60	7.0	116.0	33.58	-1.0	22.0	7.06	246
1	0.0	10.0	0.90	0.0	6.0	5.92	487.01	3648.22	2134.62	8.0	130.0	55.21	-1.0	20.0	5.73	293
2	0.0	14.0	3.15	0.0	12.0	5.30	619.07	3681.31	2196.27	20.0	142.0	92.55	1.0	22.0	13.66	375

KMedoids_Cluster_Ordenado	SEXO (% Hombre)	PUEBLO ORIGINARIO (% SI)	MIGRANTES (% SI)	USUARIO CON DISCAPACIDAD (% SI)	MEJOR NIÑEZ (% SI)	SENAME (% SI)	RIESGO (% ALTO RIESGO)	INASISTENTE (% NO APLICA)
0	51.22	2.03	0.81	0.81	1.22	1.22	0.00	80.49
1	61.77	3.07	1.71	1.71	4.10	0.68	99.66	51.19
2	41.87	2.67	2.67	1.60	0.53	3.47	93.07	24.00





Comparación de los tres métodos de Clusterización

	Métrica	K-Means	DBSCAN	K-Medoids
0	Silhouette Score	0.250822	0.219701	0.187376
1	Davies-Bouldin Index	1.482943	1.437789	1.812412
2	Calinski-Harabasz Index	290.147986	209.225307	242.184284

Los resultados muestran diferencias en la calidad del agrupamiento según el enfoque:

- **Silhouette Score**: Esta métrica mide qué tan bien definidos están los clusters. **K-Means obtuvo el mejor puntaje**, indicando una separación más clara entre grupos. DBSCAN y K-Medoids mostraron valores menores, lo que sugiere una estructura de clusters más difusa.

- **Davies-Bouldin Index**: Aquí, **valores más bajos son mejores**. El mejor resultado lo obtuvo **DBSCAN**, seguido por K-Means. K-Medoids presentó el índice más alto, lo que indica mayor superposición entre clusters en este caso.

- **Calinski-Harabasz Index**: Esta métrica evalúa la relación entre la dispersión entre clusters y la cohesión interna. **K-Means nuevamente destacó con el valor más alto**, indicando una mejor estructura de agrupamiento. K-Medoids tuvo un resultado intermedio, y DBSCAN el más bajo.

4. Resultados finales y discusión

Síntesis de Hallazgos Clave

El análisis exhaustivo de los datos del Programa CERO reveló patrones críticos en la inasistencia a controles odontológicos pediátricos.

Los pacientes clasificados como ALTO RIESGO presentaron una tasa de inasistencia significativamente mayor (χ^2 $p < 0.05$) en comparación con los de BAJO RIESGO, lo que sugiere que la percepción o el impacto real del riesgo clínico actúa como barrera para la adherencia.

Además, se identificó que el tiempo transcurrido desde el último control (MESES ULTIMO CONTROL) mostró una diferencia estadísticamente relevante (t-test $p < 0.01$) entre inasistentes y asistentes, indicando que la falta de seguimiento continuo se asocia con una mayor probabilidad de abandonar el programa.

Se aplicaron tres métodos de clusterización para agrupar a los pacientes según sus características clínicas, demográficas y conductuales: K-Means, K-Medoids y DBSCAN. De ellos, K-Means mostró el mejor desempeño cuantitativo según las métricas internas de validación (Silhouette Score, Davies-Bouldin y Calinski-Harabasz), mientras que K-Medoids permitió incluir variables categóricas relevantes para una caracterización más informativa.

Los resultados revelaron perfiles diferenciados de pacientes, destacando grupos con mayor riesgo clínico, menor adherencia a los controles o características que podrían requerir estrategias específicas. En particular, se identificaron clusters con alta inasistencia y daño por caries elevado, lo que representa un desafío para la continuidad de cuidado y el control efectivo de la salud bucal infantil.

La capacidad de segmentar a los pacientes mediante estas técnicas ofrece una herramienta útil para orientar decisiones clínicas y organizacionales. Permite identificar pacientes que podrían beneficiarse de intervenciones más intensivas, acompañamiento educativo o estrategias comunitarias. Además, esta segmentación puede facilitar la planificación eficiente de horas de atención, priorizando aquellos grupos con mayor necesidad o riesgo.

Congruencia con estudios previos de no-show

El estado del arte en predicción de inasistencia a controles odontológicos ("no show") confirma los hallazgos identificados en el Programa CERO. La literatura actual destaca que los modelos predictivos basados en machine learning (Random Forest, Gradient Boosting) alcanzan precisiones de 71-75% (AUC), siendo el historial previo de inasistencias y el perfil de riesgo los predictores más robustos. Estudios recientes demuestran que las intervenciones diferenciadas según nivel de riesgo, como las propuestas para pacientes de ALTO RIESGO, han logrado reducciones de hasta 46% en tasas de inasistencia en contextos similares al chileno.

Recomendaciones Clínicas y de Gestión

Los resultados evidencian que la inasistencia no es un fenómeno aleatorio, sino que sigue patrones estructurados vinculados a variables clínicas y demográficas. La clasificación de riesgo emerge como un predictor crítico, lo que sugiere la necesidad de intervenciones diferenciadas:

- Para pacientes de ALTO RIESGO: Implementar recordatorios proactivos multicanal (SMS, correo electrónico) y asignar citas en horarios flexibles para cuidadores.
- Para pacientes con DAÑO POR CARIES elevado: Establecer protocolos de seguimiento intensivo, con intervalos entre controles ajustados dinámicamente según la progresión de la condición.

La falta de correlación significativa entre la DISTANCIA AL CESFAM y la inasistencia ($r \approx 0$) contradice hipótesis iniciales, indicando que las barreras geográficas no son determinantes en este contexto. Esto orienta a priorizar soluciones centradas en factores socio-clínicos más que logísticos.

Limitaciones y Consideraciones Éticas

El estudio se basó en datos de un único CESFAM, lo que limita la generalización de resultados. Además, la categorización binaria de "INASISTENTE" podría no capturar matices como el retraso ocasional frente al abandono total. Se respetaron rigurosamente los protocolos de anonimización y ética en el manejo de datos sensibles, alineados con las guías del CENS (2024).

Recomendaciones

- Intervenciones Piloto: Diseñar programas de incentivos comunitarios (ej.: certificados de reconocimiento para familias con adherencia perfecta) para grupos de alto riesgo.
- Monitoreo Continuo: Implementar un tablero de control en tiempo real que alerte sobre pacientes con probabilidad >70% de inasistencia, basado en modelos predictivos.

Posibles Líneas de Trabajo Futuro

- Modelos Predictivos Avanzados: Integrar técnicas de machine learning (como XGBoost o redes neuronales) para predecir inasistencia con base en patrones multifactoriales, incluyendo variables conductuales no analizadas en este estudio.
- Análisis cualitativo de barreras psicosociales: Desarrollar un estudio mixto que explore factores emocionales y de percepción (como ansiedad dental infantil o creencias parentales sobre salud oral) mediante entrevistas semiestructuradas para complementar los hallazgos cuantitativos y diseñar intervenciones culturalmente sensibles.
- Implementación de modelos de series temporales: Aplicar técnicas para analizar patrones estacionales y tendencias temporales en las inasistencias, identificando períodos críticos del año o intervalos horarios con mayor probabilidad de ausentismo para optimizar la programación.
- Estratificación dinámica del riesgo: Desarrollar un algoritmo que ajuste la clasificación de riesgo en tiempo real incorporando variables contextuales (como cambios estacionales, eventos escolares o factores socioeconómicos emergentes) para optimizar la programación de citas y recursos asistenciales.

Conclusión Final

Este estudio no solo identifica los factores críticos asociados a la inasistencia en el Programa CERO, sino que propone un marco accionable para transformar datos en estrategias preventivas. Al priorizar a los pacientes de ALTO RIESGO y optimizar la frecuencia de controles mediante modelos basados en evidencia, se podría incrementar la eficiencia operativa y la salud de la población. La integración de aprendizaje profundo en fases posteriores representa un salto cuantitativo hacia la odontología preventiva personalizada, asegurando que ningún niño quede fuera del sistema por barreras prevenibles.

---

5. Bibliografía

- Orientaciones Técnicas Programa Cero: https://drive.google.com/file/d/19mbZF2TYgbT_pPg---2HldwYhC99TeVv/view
- Aseguramiento de la calidad estadística (2024), Instituto Nacional de Estadísticas: https://www.ine.gob.cl/calidad-estadistica/aseguramiento-de-la-calidad-estadistica
- Formulación ética de proyectos de ciencia de datos (2022), División de Gobierno Digital y Universidad Adolfo Ibáñez: https://digital.gob.cl/transformacion-digital/estandares-y-guias/guia-formulacion-etica-de-proyectos-de-ciencia-de-datos/
- Guía Introductoria de Buenas Prácticas de Privacidad y Seguridad de Datos en Salud (2024), Centro Nacional en Sistemas de Información de Salud: https://cens.cl/guia-introductoria-de-buenas-practicas-de-privacidad-y-seguridad-de-datos-en-salud/
- Factors associated with regular dental attendance by aged adults: A systematic review https://onlinelibrary.wiley.com/doi/pdf/10.1111/ger.12661
- Estado de Salud Oral y Asistencia al Control Odontológico en Escolares de 12 Años, Comuna de Penco, Región del Biobío https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-381X2013000300009

- Factors affecting children's adherence to regular dental attendance: A systematic review https://www.sciencedirect.com/science/article/abs/pii/S0002817714601914
- Three factors predicting irregular versus regular dental attendance: A model fitting to empirical data https://pubmed.ncbi.nlm.nih.gov/6942958/

  https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0528.1980.tb01320.x
- Factors associated with regular dental attendance by aged adults: A systematic review https://onlinelibrary.wiley.com/doi/full/10.1111/ger.12661
- Frecuencia de asistencia a la consulta odontológica en el control prenatal y factores asociados en un hospital público de Bogotá, Colombia, 2011-2012 http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0034-74342016000400288
- Predicting no-shows for dental appointments https://peerj.com/articles/cs-1147/
- Automated Machine Learning in Dentistry: A Narrative Review of Applications, Challenges, and Future Directions https://www.mdpi.com/2075-4418/15/3/273
- Exploring Factors Associated With Missed Dental Appointments: A Machine Learning Analysis of Electronic Dental Records https://pmc.ncbi.nlm.nih.gov/articles/PMC10656597/
- The utilization of AI in healthcare to predict no-shows for dental appointments: A case study conducted in Saudi Arabia https://www.sciencedirect.com/science/article/pii/S2352914824000285
- Solano, P., Marconi, C., & Cartes-Velásquez, R. (2022). Predicting no-shows for dental appointments. PeerJ Computer Science, 8. https://peerj.com/articles/cs-1147/
- Alabdulkarim, M., Xu, J., & Wang, P. (2022). Exploring Factors Associated With Missed Dental Appointments: A Machine Learning Analysis of Electronic Dental Records. Journal of Medical Internet Research, 24(11). https://pmc.ncbi.nlm.nih.gov/articles/PMC10656597/
- Schuurs, A.H.B., Duivenvoorden, H.J., & Thoden van Velzen, S.K. (1980). Three factors predicting irregular versus regular dental attendance: A model fitting to empirical data. Community Dentistry and Oral Epidemiology, 8(8):413-418. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0528.1980.tb01320.x
- Vera, F., Marconi, C., & Weintraub, G.Y. (2023). Predicting no-show appointments in a pediatric hospital in Chile using machine learning. Health Care Management Science, 26:468-483. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10257628/
