UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FISICA BIOMEDICA

SEMESTRE 2025-2

PROYECTO INTEGRATIVO

“Aplicación de la ciencia de datos para la elaboración de un modelo de predicción de riesgo de colapso hospitalario a partir de datos disponibles”

Autor:

Luis Gerardo Pérez Martínez

Introducción a la Ciencia de Datos aplicada a escenarios médico – biológicos
13 de junio, 2025.

1. Introducción

No es sorpresa que uno de los principales problemas que enfrenta el sistema de salud en México es la insuficiencia presupuestaria. Durante años, la inversión en salud a estado por debajo de estándares recomendados por organismos internacionales como la Organización Mundial de la Salud (OMS), que, en cifras, se traduce en tan solo un presupuesto de apenas el 2.5%, muy por debajo del 6% del PIB que recomienda la OMS. Esto ha resultado en una infraestructura hospitalaria deficiente, particularmente en zonas rurales, donde los hospitales y clínicas suelen operar con equipos obsoletos, o en algunos casos, sin los insumos básicos necesarios para brindar atención. Además, la falta de personal medico es alarmante: el número de médicos y enfermeros por habitante esta muy por debajo del promedio de países de la Organización para la Cooperación y Desarrollo (OCDE), lo que agrava la incapacidad del sistema para atender a la población. La fragmentación del sistema de salud también es un obstáculo importante. Históricamente, la coexistencia de distintas instituciones como el IMSS, ISSTE y el reciente creado IMSS- Bienestar, han creado un modelo desigual y poco eficiente. Los ciudadanos enfrentan barreras burocráticas para acceder a los servicios y, en muchos casos, terminan recurriendo al sector privado, donde los costos suelen ser exorbitantes para la mayoría de la población. Estos problemas fueron fuertemente evidenciados durante la pandemia de COVID iniciada en el ano 2020, en donde se registró récords históricos en todo el país [1] con ocupaciones entre 90 y 100% en algunos hospitales de varios estados de la Republica. Según distintos estudios, 58% de todos los pacientes con COVID-19 que murieron, nunca llegaron a tener atención en el hospital y una gran proporción de los que llegaron a tener una cama en un hospital y fallecieron, no tuvieron acceso a cuidados intensivos. En conclusión, la pandemia de COVID llevo a exponer las profundas desigualdades en el acceso a la atención medica en el país. En los últimos anos, han surgido problemas que agravan la evidencia del sistema de salud en México y lo alejan cada vez mas a un sistema de primer nivel de atención medica con el de Dinamarca, por ejemplo, el aumento de incidencia de enfermedades crónicas no transmisibles (enfermedades cardiovasculares, diabetes e hipertensión), la tan estructura fragmentada del sistema de salud mexicano que se tiene actualmente en donde la calidad del servicio depende de la institución sea publica o privada y la región en la que se encuentre, falta de mantenimiento a la infraestructura hospitalario, dando casos como el ocurrido en Julio de 2023, donde en el Hospital General de Playa del Carmen en Quintana Roo, un paciente pediátrico falleció a causa de un fallo en el elevador derivado de problemas de infraestructura.

Otro aspecto critico es la corrupción y la mala gestión administrativa, como los escándalos de compras de medicamentos con sobreprecios en pleno ano 2025 y el gran problema de desabasto y distribución de fármacos que se vive a nivel nacional desde 2024. Estos problemas han orillado al **colapso hospitalario** en diversos hospitales del país. [2]

Entendemos al colapso hospitalario como la situación en la que un establecimiento de salud pierde la capacidad de atender adecuadamente la demanda de servicios médicos de manera oportuna para cada paciente. Las causas de la saturación son diversas e implican aspectos tanto externos como intrínsecos a la propia unidad. Pero los mas determinantes son propios de la dinámica hospitalaria, fundamentalmente la dificultad en adjudicación de cama para el ingreso y en su disponibilidad real. Esta saturación se asocia a un descenso de la mayoría de indicadores de la Salud. [1]

La utilidad de la Ciencia de Datos en la salud publica

Proponer soluciones para los problemas de salud pública actuales, requiere de métodos de estudio eficaces que nos lleven a análisis y conclusiones rigurosas, tal como lo hacen las recientes investigaciones enfocadas en el caso. La aplicación de modelos provenientes de la Ciencia de datos han sido una herramienta útil y sirve para orientar políticas sanitarias, optimizar la planificación de servicios y facilitar intervenciones clínicas en pacientes de alto riesgo. Un claro ejemplo es el modelo de inteligencia artificial Foresight diseñado por el NHS England, el cual utiliza el aprendizaje predictivo para anticipar eventos futuros. Este modelo ha sido entrenado con datos anónimos de egresos hospitalarios, tasas de vacunación contra la covid-19, para proyectar eventos clínicos como hospitalizaciones, infartos, etc. [3]

Antecedentes en México

En nuestro país, **el uso de la inteligencia artificial** para abordar los problemas del sistema de salud, han sido evaluados en diversos estudios por parte de distintas organizaciones publicas y privadas. En 2025, El articulo publicado por la Revista Científica Estelí titulado “Salud pública y su relación con el crecimiento urbano en la atención hospitalaria de la Ciudad de México, México” analizo la relación directa entre el crecimiento urbano y la deficiente atención hospitalaria utilizando datos vectoriales y datos de Sistema de Información de la Red IRAG. A través de una aproximación geoespacial y descriptiva, el estudio evidencio patrones de concentración poblacional y una desigualdad en la distribución de infraestructura hospitalaria demuestran esta relación. [6]

El actual trabajo propone seguir estudiante esta relación entre la creciente densidad poblacional y el desabasto hospitalario desde una perspectiva analítica y de modelado predictivo. Sin embargo, a diferencia del estudio anterior se busca construir un índice de riesgo hospitalario y aplicar técnicas de clasificación y predicción para la identificación de unidades hospitalarias con mayor probabilidad de colapso. [5]

Justificación

La Red IRAG es una plataforma creada en conjunto por la secretaria de Salud y la UNAM para transparentar la ocupación hospitalaria en tiempo real a quienes deseen consultarlo. Mientras IRAG informa ocupación actual, los modelos desarrollados también permiten anticipar escenarios de colapso e incluso en ausencia de datos actualizados. Esto plantea la capacidad de planeación anticipada, asignación de recursos correctos y una buena gestión de emergencias sanitarias.

2. Objetivos

Objetivo general:

- Desarrollar un modelo de predicción del riesgo de colapso hospitalario en la Ciudad de México del año 2023, utilizando variables disponibles en la base de recursos hospitalarios de la Secretaria de Salud.

Objetivos particulares:

- Explorar y analizar la base de recursos del 2023 de la secretaria de salud.
- Definir un umbral de colapso hospitalarios basado en la literatura o en criterios técnicos derivado en el análisis observado.
- Crear un índice de riesgo hospitalario para clasificación de riesgo de las unidades médicas.
- Entrenar y evaluar modelos de clasificación supervisada.

3. Metodología

En este apartado se detalla todos los pasos seguidos para la obtención del modelo. Se utilizo el lenguaje de programación Python en el entorno de desarrollo integrado Colab. Además, se uso las librerías de pandas, matplotlib, seaborn, Scikit-learn, numpy, etc. La estructura del análisis siguió metodologías típicas utilizadas en Ciencia de Datos, las cuales son el análisis exploratorio de los datos (EDA), el preprocesamiento, análisis estadísticos de la distribución de los datos y modelado supervisado. Durante la primera parte del EDA, se observó los diferentes tipos de datos, como datos numéricos (conteo de camas, consultorios, personal y equipos médicos), categóricos (jurisdicción, tipos de establecimiento y tipología), y de cadena (‘clues’ y ‘nombre del

establecimiento') como identificadores de cada unidad hospitalaria. Se observaron 427 unidades clínicas y hospitalarias para la CDMX durante 2023. También se realizo el tratado de la redundancia columnar (variables repeditas, errores o tratado ortografico) para algunas variables.

En el preprocesamiento de los datos se manejaron las variables categóricas y numéricas por separado. Para las categóricas, se realizó el análisis y en caso de requerirse el tratado de los datos, estas fueron:

- Datos nulos
- Datos duplicados
- Visualización con gráficos

Parte de las visualizaciones en esta etapa revelaron las proporciones de clínicas y hospitales que tiene cada alcaldía. En donde, la alcaldía Iztapalapa posee la mayor cantidad de hospitales y clínicas de la CDMX.

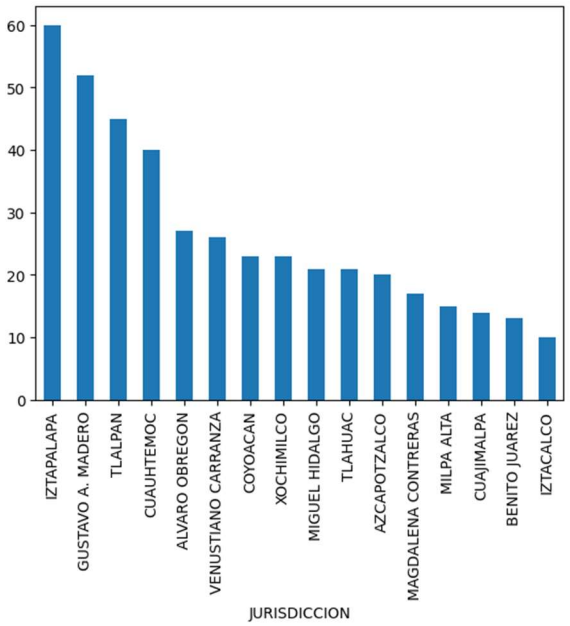


imagen 1. Distribución de clinicas y hospitales por alcaldia

El posterior análisis de los tipos de establecimientos que hay en la CDMX, revelo que únicamente hay unidades de consulta externa (CE) y hospitalizaciones (HO).

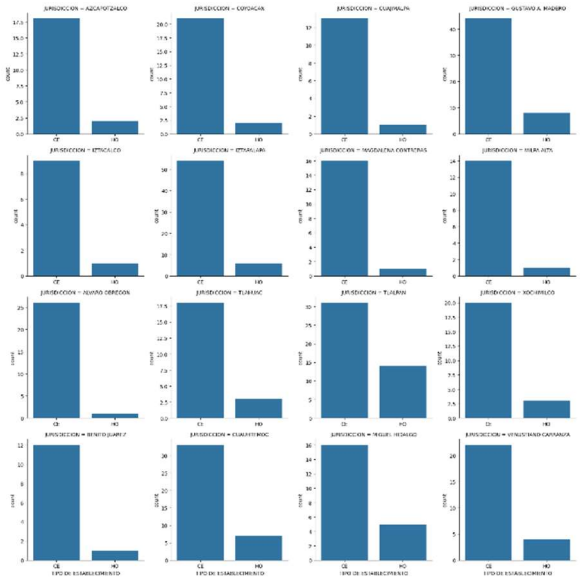


imagen 2. Distribucion de tipos de establecimientos por alcaldia

Analizando la información contenida en 'Tipología' , se notó que cada código en forma de letra representa un nivel de atención para cada unidad de CE y HO. La siguiente grafica muestra el tipo de tipología encontrada en la CDMX para cada CE y HO.

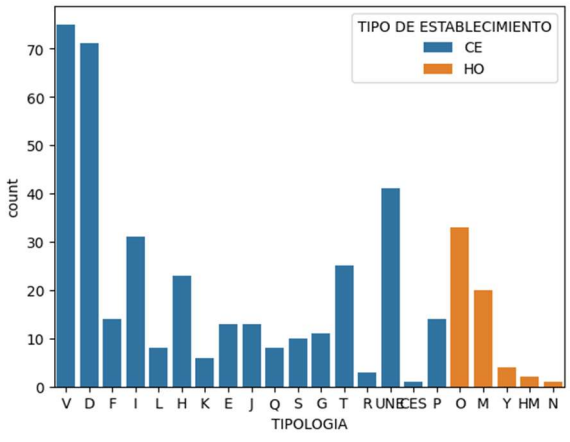


Imagen 3. Distribución de Tipología para la CDMX

La codificación de los códigos para las unidades HO se resumen en la siguiente tabla:

CODIFICACION DE 'TIPOLOGIA PARA HO'	
O	Hospital especializado
M	Hospital general
Y	Hospital psiquiátrico
HM	NA
N	Hospital integral

Tabla 1. tipología asignada para HO

Se realizo un análisis mas riguroso para determinar el nombre de los establecimientos. Sin embargo, se encontró que únicamente se tenían datos de centros de salud, clínicas y hospitales de la Secretaria de Salud (SSA) sin contar unidades del IMSS, IMSSTE y unidades privadas. Al notar esto, se determino que no seria posible modelar el riesgo de colapso hospitalario real. Además, debido también a la inmensa cantidad de personas atendidas en estas instituciones no es posible utilizar la población total por alcaldía, ya que se puede sobredimensionar la presión real del establecimiento de la SSA. Es así, que durante la metodología explicada, se propuso cambiar el enfoque metodológico sin perder de vista los objetivos planteados. De a partir de aquí se propuso trabajar con unidades de tipo HO, y dejando a CE para un modelo futuro. Y también se propuso cambiar la población total por alcaldía como fenómeno para definir umbrales por la afluencia anual y diaria de cada establecimiento. Para esto fue necesario recurrir a las bases de datos de urgencias y egresos hospitalarios de la SSA.

Seguidamente se realizó el tratamiento de datos numéricos en donde se analizo y corrigió lo siguiente:

- Datos nulos o faltantes
- Outliers o datos atípicos
- Agrupación de datos de conteos
- Pruebas estadísticas
- Transformaciones
- Mapa de correlación

Las variables numéricas presentaron el conteo total de cada tipo de médico, consultorio, camas disponibles, otro tipo de personal, enfermeros y equipo médico. Se realizaron variables agrupadas, que se consideraron mas importantes en el riesgo de colapso hospitalario:

1. Consultorios:
   - Generales
   - Especialidades
   - Salud mental
   - Urgencias
2. Camas:
   - De hospitalización
   - Criticas (recuperación, terapia intensiva e intermedia)
   - Camas de atención inmediata (urgencias)

- Camas neonatales

3. Médicos:
   - Generales
   - Urgenciólogos
   - Especialistas
   - Cirujanos (generales y especializados)

Con el objetivo de estimar el riesgo de colapso hospitalario se decidió construir tasas que relacionan la afluencia diaria de pacientes de urgencias y hospitalización con los recursos disponibles. Esta decisión fue por dos motivos principales:

1. Capturar la presión real sobre los recursos.
   Analizar ambos factores (cantidad de recursos físicos y afluencia) por separado no permitió identificar con claridad la sobrecarga relativa en cada unidad. La construcción de tasas tipo:

   Tasa = afluencia diaria / recurso disponible.

   Logra reflejar cuanta carga representa cada paciente sobre los recursos disponibles.

2. Problemas de multicolinealidad.
   En el análisis de correlación, se opto por reducir el numero de variables redundantes y que no muestran una fuerte relación entre ellas y el fenómeno, así conservando la información relevante en forma de relaciones funcionales entre la demanda/capacidad de respuesta. Las tasas construidas fueron las siguientes:

- Afluencia en urgencias/ camas de atención inmediata
- Afluencia en urgencias/médicos urgenciólogos
- Afluencia en urgencias/consultorios de urgencias
- Afluencia en urgencias/médicos cirujanos
- Afluencia en hospitalización/camas para hospitalización
- Afluencia en hospitalización/ médicos especialistas
- Afluencia en hospitalización/camas criticas
- Afluencia hospitalización/ camas generales
- Afluencia en hospitalización/médicos generales

La variable de riesgo hospitalario fue definida a partir de la clasificación de las diversas tasas construidas que reflejan la presión sobre los recursos hospitalarios. El criterio utilizado para las tasas fue el siguiente:

Valor de la tasa	etiqueta	Nivel de riesgo (índice)
Menor a 0.7	1	Bajo

Entre 0.7 – 1	2	Moderado
Mayor a 1	3	Alto

Tabla 3. Criterio para definir el umbral por tasas

Seguidamente, se definió una variable compuesta (riesgo global) considerando el numero total de tasas en niveles altos y de riesgo moderado en cada unidad: El criterio fue el siguiente:

- Riesgo alto (3): si una unidad tiene al menos 6 tasas en niveles críticos
- Resigo moderado (2): si hay al menos 5 tasas (criticas y moderadas) en una unidad
- Riesgo bajo (1): para los demás casos.

Esta lógica se basa en un enfoque aditivo, donde la acumulación de presiones en múltiples recursos indica una mayor probabilidad de disfunción hospitalaria. Si bien no existe una regla universal para determinar estos umbrales, se optó por valores que permitieran una distribución balanceada de clases, interpretabilidad clínica, y que pudieran ser útiles como insumo para modelos de predicción y priorización operativa.

Por ultimo se divido los datos en entrenamiento y validación para la prueba de modelos. Se utilizaron 2 algoritmos supervisados: árbol de decisión y K-vecinos cercanos (KNN) calculando sus métricas y comparándolas para concluir sus rendimientos.

3. Resultados

A continuación, se muestra los resultados para los modelos utilizados:

RESULTADOS PARA: RIESGO GLOBAL

	precision	recall	f1-score	support
1	0.33	1.00	0.50	2
2	0.78	0.70	0.74	10
3	1.00	0.57	0.73	7
accuracy			0.68	19
macro avg	0.70	0.76	0.65	19
weighted avg	0.81	0.68	0.71	19

Imagen 4. Métricas para el algoritmo árbol de decisión.

muestran que el 68% del total de predicción fueron correctas. Las demás métricas se mantienen en valores aceptables (macro avg y weighted avg) y comportamientos adecuados para un conjunto de tres clases. El modelo funciona mejor para la clase 2 (riesgo moderado) siento la mas común en la base de datos. Notamos que tiene mas dificultades de precisión en cuanto a clases menos frecuentes como el caso de la 1 (riesgo bajo), sin embargo, el modelo identifico correctamente a los 2 establecimientos que en verdad fueron de riesgo bajo. Para la clase 2 (riesgo moderado) el modelo acertó en un 78% para los verdaderos positivos y en cuanto a la clase 3 aunque identifica con mucha precisión, su efectividad del 57% sobre los verdaderos positivos hace que no podamos identificar todos los hospitales en una situación de riesgo alto de colapso. Este comportamiento puede deberse al desequilibrio entre clases y también al hecho de la definición que se escogió para el riesgo hospitalario global de cada unidad.

RESULTADOS PARA: RIESGO GLOBAL - KNN

	precision	recall	f1-score	support
1	0.33	0.50	0.40	2
2	0.58	0.70	0.64	10
3	0.50	0.29	0.36	7
accuracy			0.53	19
macro avg	0.47	0.50	0.47	19
weighted avg	0.53	0.53	0.51	19

imagen 5 Métricas para KNN

predicciones correctas. Sin embargo, mostro un recall del 29 % para las unidades de riesgo alto, lo que indica que identifica pocos casos críticos correctamente. Además, se muestra limitaciones importantes en detectar casos de riesgo bajo, y en general tener métricas generales menos eficientes. Esto sugiere que el modelo KNN no distingue con suficiente claridad entre las distintas clases de riesgo, probablemente por la cercanía en el espacio multivariado de algunas tasas. La siguiente grafica muestra la superioridad del modelo de Árbol de decisión frente al KNN.

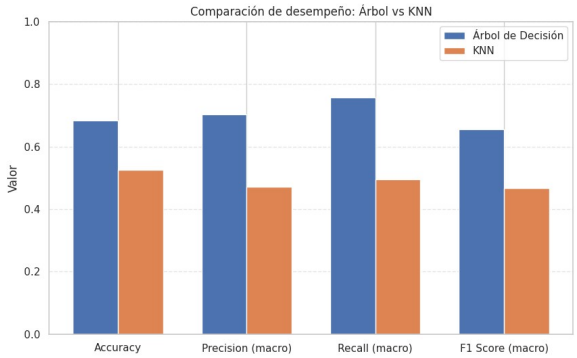


imagen 6. Comparación de árbol de decisión y KNN

decisión, revela claramente ciertos errores de clasificación anteriormente comentados, especialmente entre niveles moderado – bajo y alto – bajo/moderado. Estos errores tienen un gran impacto en las decisiones correctas frente a un riesgo real de colapso y la mejora de infraestructura interna. Estos errores pueden ser atribuidos a la propia naturaliza de los datos.

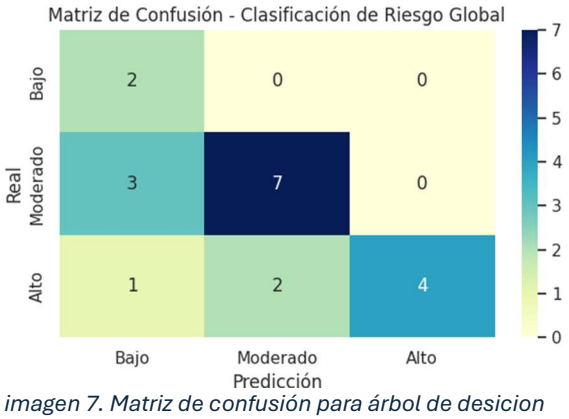


imagen 7. Matriz de confusión para árbol de desicion

En la metodología se propuso realizar grupos para cada variable. El análisis de correlación mostro grupos con alta correlación que obligo a reducir dimensiones y trabajar con indicadores (tasas) seleccionados.

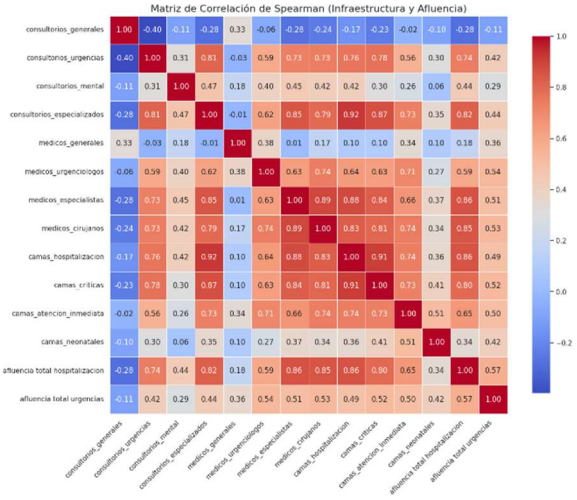


Imagen 8. Matriz de correlación para las variables agrupadas

Se analizo la distribución de las tasas obtenidas, teniéndose distribuciones no normales y asimétricas debido a que provienen de distribuciones no normales y asimétricas positivas, sin embargo, no se opto por un proceso de transformación como estandarización o aplicar logaritmo

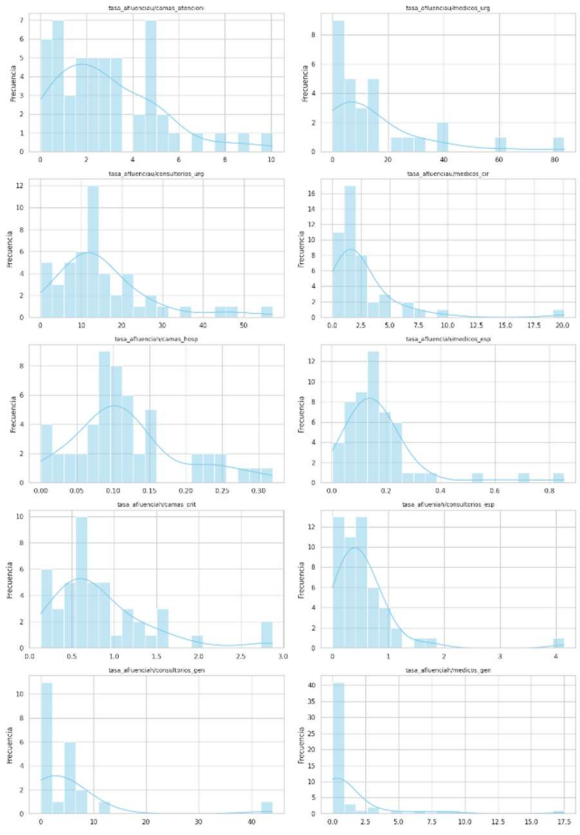


Imagen 9. Gráficos de distribucion de las 10 tasas creadas

debido al diseño y construcción de los umbrales con los valores de las tasas.

Las variables de afluencia provenientes de bases de urgencia y egresos hospitalarios también relevaron este comportamiento.

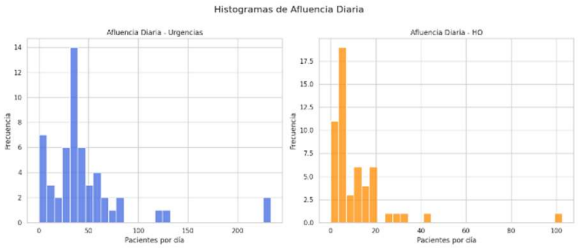


Imagen 10. Graficos de Dsitribucion de la afluencia urgencias/hospitalización

En cuanto a los distintos grupos pertenecientes al total de variables por consultorios, camas y personal se obtuvo el mismo comportamiento en cuanto a valores altamente atípicos, distribución de los datos y pruebas estadísticas.

Las variables de la distribución de camas revelo datos atípicos tanto en camas de especialización, de atención inmediata que involucra camas de urgencias, camas de atención critica y neonatales. Sin embargo, se realizó una exploración e investigación a las unidades que presentaron este tipo de datos y se encontró que pertenecen a hospitales con alta capacidad de afluencia y estructura interna por lo cual evidencio la capacidad real hospitalaria. Esta explicación concluyo no imputar o eliminar los outliers lo cual pudo haber contribuido fuertemente a los resultados.

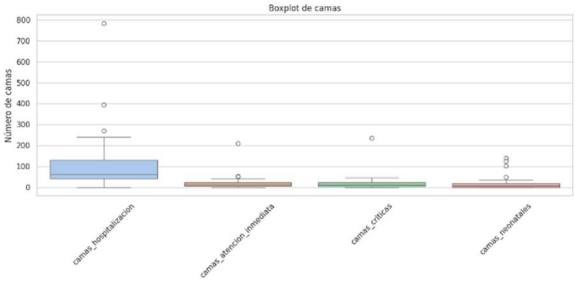
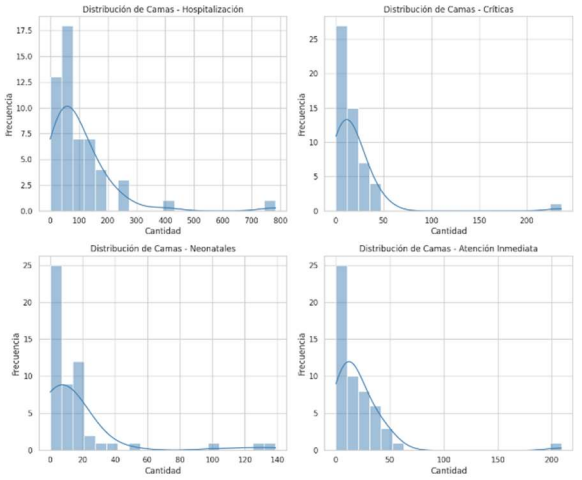


Imagen 11. Boxplots para camas hospitalarias, atención inmediata, criticas y neonatales

Se utilizaron pruebas de Shapiro Wells y de Fisher debido a la cantidad de datos disponibles en la base, las pruebas arrojaron valores menores que el valor critico (0.05), con lo cual se rechaza la hipótesis nula y se concluye que los datos no presentan normalidad en los datos. La prueba de Fisher revelo un valor de 4.91, asumiendo una distribución asimétrica sesgada a la derecha.




```
Shapiro-Wilk: Estadístico = 0.5135, p-valor = 0.0000
Asimetría (skewness) = 4.9197
Distribución sesgada a la derecha.
```

Imagen 12. Pruebas estadisticas para comprobar normalidad y simetría para las variables de camas

La forma de la distribución nos ayudo a diagnosticar la calidad estadística de las variables y guiar decisiones posteriores en cuanto a la selección de modelos.

Una de las principales consideraciones metodológicas del modelo fue la necesidad de estimar la afluencia diaria promedio de pacientes, ya que los datos disponibles correspondían al total anual. Ante la ausencia de registros diarios o mensuales, se optó por dividir el total anual entre 365 días, lo cual permitió generar tasas ajustadas de atención por recurso disponible (como egresos/camas, urgencias/médicos, etc.). Si bien esta media no capta la variabilidad estacional o regional, ofrece una aproximación razonable para homogeneizar la comparación entre unidades hospitalarias.

4. Conclusiones.

Desde una perspectiva medica y de políticas de salud pública, el uso de modelos predictivos como el arbol de decisión permite identificar hospitales con alta probabilidad de saturación, lo que puede derivar en demoras críticas, deficiencia de atención y posibles retrasos para tratamientos de los pacientes.

En términos presupuestales, esta clasificación predictiva permite orientar mejor las inversiones a distintas áreas de la institución y establecer medidas para evitar el colapso hospitalario.

Aunque esta metodología presenta limitaciones inherentes a la disponibilidad de datos, permite avanzar en la construcción de herramientas de predicción con enfoque práctico, ajustado a la realidad de los sistemas de salud públicos.

Referencias:

[1] Veme Digital. (s.f.). *El tejemaneje del desabasto de medicamentos*. https://www.veme.digital/post/el-tejemaneje-del-desabasto-de-medicamentos

[2] Organización Editorial Mexicana. (2024). *Crisis del sistema de salud pública: un reto pendiente*. El Sol de Acapulco. https://oem.com.mx/elsoldeacapulco/analisis/crisis-del-sistema-de-salud-publica-un-reto-pendiente-13345798

[3] México, ¿Cómo Vamos? (2023). *Los retos de nuestro sistema de salud*. Animal Político. https://mexicocomovamos.mx/animal-politico/2023/09/los-retos-de-nuestro-sistema-de-salud/

[4] López Miranda, R. J. (2022). *Desabasto de medicamentos y sus implicaciones para el sistema público de salud en América Latina: una visión crítica*. Revista FAREM-Estelí, Universidad Nacional Autónoma de Nicaragua. https://www.camjol.info/index.php/FAREM/article/download/19981/24417?inline=1

[5] Rivera González, O. D. (2025). Salud pública y su relación con el crecimiento urbano en la atención hospitalaria de la Ciudad de México, México. *Revista Científica Estelí, 13*(52), 52–66. https://revistas.unan.edu.ni/index.php/Cientifica/article/view/4892

[6] Escobar, A., Roldán, L., & Pérez, M. (2015). *Evaluación de la carga de trabajo asistencial en un servicio de urgencias hospitalario*. Emergencias, 27(2), 113–120. https://revistaemergencias.org/wp-content/uploads/2023/08/Emergencias-2015_27_2_113-120-120.pdf