UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERIA
ESCUELA DE ESTUDIOS DE POSTGRADO
MAESTRÍA EN INGENIERIA PARA LA INDUSTRIA
CON ESPECIALIZACIÓN EN CIENCIAS DE LA COMPUTACIÓN
Nery Francisco Orellana Pineda
999013582

Proyecto No.1

Objetivo

El objetivo del proyecto es analizar la información hospitalaria del país de Guatemala, en especifico de los centros hospitalarios nacionales en su consulta externa. El análisis de patrones y asociaciones sen los datos de salud puede ofrecer valiosas perspectivas sobre el uso de servicios médicos.

A través de algoritmos de minería de datos, como Apriori o FPGrowth, es posible identificar reglas de asociación que revelan relaciones entre diferentes variables de atención médica. Este tipo de análisis es crucial para entender los patrones en el sistema de salud, facilitando la identificación de tendencias que pueden ser útiles en la toma de decisiones.

Metodología

Para el presente trabajo se utilizó la fuente de datos del Instituto Nacional de Estadística (INE) del año 2022, la fuente de datos era en especifico de los servicios de la consulta externa, se tomo como base la pestaña “3 sexo edad y tipo de consulta”. De los cuales fueron ignoradas las celdas Total, e Ignorado. Así mismo al tener el dataset se omitieron los valores de “Totales” para no afectar en los algoritmos.

Luego de ello se procedió a pasar a la data como factores, para un mejor análisis y se aplicó el algoritmo Apriori con un soporte de 0.1 y un nivel de confianza de 0.5, obteniendo un total de 62 reglas de asociación. Posteriormente se aplico el algoritmo de FPGrowth, de las cuales se obtuvieron un total de 133 reglas de asociación, de igual forma con un nivel de soporte de 0.1 y una confianza de 0.5.

Por ultimo se utilizo la el algorito de kmeans con 4 centros, para ello se volvió a carga la información a una nueva variable, esto para no afectar a los clusters, se convirtieron todos los valores a numéricos y se omitieron todos los valores nulos, se uso un seed de 32 para la reproducibilidad. Por último, se procedió a graficar los clusters para la visualización de la información.

Resultados obtenidos

Apriori:

Para el primer algoritmo se seleccionaron las siguientes reglas de asociación:

- {Reconsulta y emergencia=-} => {Sexo=Hombres}
- {Primera consulta y emergencia=11} => {Reconsulta=[113,9.4e+03)}
- {Sexo=Mujeres} => {Reconsulta=[1.08e+04,2.08e+04]}
- {Primera consulta=[7.91e+03,1.87e+04),Reconsulta y emergencia=-}=> {Reconsulta=[113,9.4e+03)}

La primera regla de asociación {Reconsulta y emergencia=-} => {Sexo=Hombres}, nos indica que una asociación entre la ausencia de datos en Reconsulta y emergencia y el género masculino. Esto podría significar que, en los casos en los que no se registran reconsultas y emergencias, el paciente suele ser hombre. esta regla resalta una tendencia de ausencia en registros de reconsulta y emergencia cuando el paciente es hombre.

La segunda regla de {Primera consulta y emergencia=11} => {Reconsulta=[113,9.4e+03)}, cuando hay exactamente 11 casos de "Primera consulta y emergencia", el número de reconsultas tiende a caer en un rango de 113 a 9,400. En otras palabras, esta asociación sugiere que, cuando se presenta este número específico de "Primera consulta y emergencia" (11), el número de reconsultas no suele ser muy alto y se mantiene dentro de ese intervalo.

Luego la regla {Sexo=Mujeres} => {Reconsulta=[1.08e+04,2.08e+04]}, Esta regla sugiere que cuando el paciente es una mujer, el número de reconsultas suele caer en un rango medio a alto, entre 10,800 y 20,800. Esto implica una asociación entre el género femenino y una mayor frecuencia de reconsultas dentro de este intervalo.

Y la cuarta regla nos dice {Primera consulta=[7.91e+03,1.87e+04),Reconsulta y emergencia=-}=> {Reconsulta=[113,9.4e+03)}, Patrón de Seguimiento Moderado: Aunque el número de primeras consultas es alto (de 7,910 a 18,700), no se registran emergencias o reconsultas, y el volumen de reconsultas cae en un rango moderado. Esto podría indicar un grupo de pacientes que requieren cierto nivel de seguimiento, pero no de manera urgente.

FPGrowth

Las cuatro reglas de FPGrowth que se usaron para el estudio fueron las siguientes:

- {Sexo=Hombres, Reconsulta y emergencia=-}
- {Primera consulta=[1.87e+04,5.31e+04], Reconsulta=[1.29e+04,1.79e+04]}
- {Sexo=Mujeres, Primera consulta=[1.87e+04,5.31e+04]}
- {Sexo=Hombres, Primera consulta=[232,7.91e+03), Reconsulta=[9.4e+03,1.29e+04)}

La primera nos dice que cuando el paciente es hombre, es común que la columna de "Reconsulta y emergencia" esté vacía o sin datos. La asociación indica una relación entre ser hombre y no tener registro de reconsulta y emergencia.

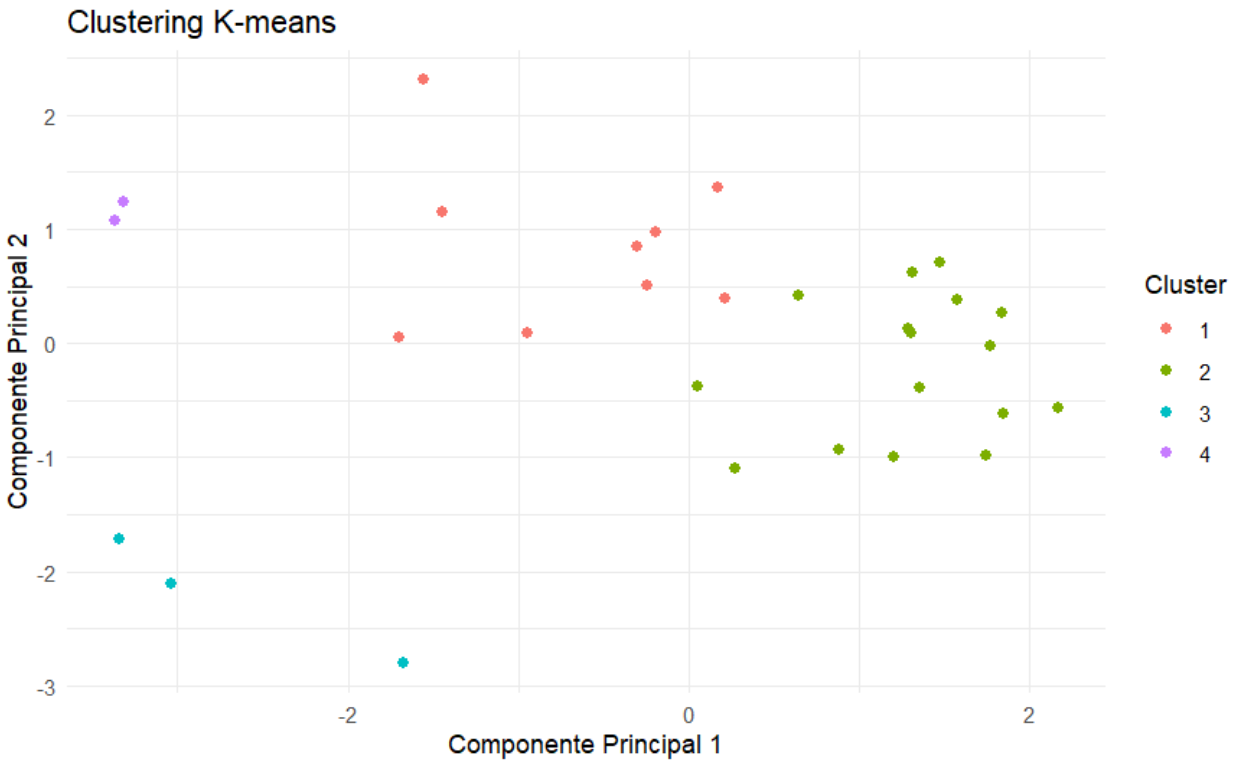
La segunda regla {Primera consulta=[1.87e+04,5.31e+04], Reconsulta=[1.29e+04,1.79e+04]}, cuando el número de primeras consultas se encuentra en el rango de 18,700 a 53,100, el número de reconsultas asociado también tiende a estar en un rango alto, de 12,900 a 17,900. En otras palabras, existe una asociación entre un volumen elevado de primeras consultas y una alta demanda de reconsultas.

{Sexo=Mujeres, Primera consulta=[1.87e+04,5.31e+04]}, que es la tercera regla nos dice que cuando el paciente es mujer, el número de primeras consultas tiende a estar en un rango elevado, de 18,700 a 53,100. Esto implica una asociación entre el género femenino y una alta demanda de primeras consultas dentro del sistema de salud.

Y la última regla de FPGrowth analizada es la de {Sexo=Hombres, Primera consulta=[232,7.91e+03), Reconsulta=[9.4e+03,1.29e+04)}, cuando el paciente es hombre y el número de primeras consultas está en un rango bajo a moderado (232 a 7,910), entonces el número de reconsultas asociadas suele encontrarse en un rango moderado a alto (9,400 a 12,900). Esto sugiere que, a pesar de tener un volumen de primeras consultas bajo a moderado, la demanda de reconsultas para estos pacientes masculinos es relativamente alta.

KMeans

Al implementar el algoritmo de KMeans con un centro 4 clusters se obtuvo la siguiente grafica



La gráfica muestra los puntos distribuidos en dos componentes principales (Componente Principal 1 y Componente Principal 2). Estos componentes son combinaciones de las variables originales que simplifican la visualización al reducir la dimensionalidad, permitiendo observar la agrupación en dos dimensiones.

Los clusters que están dispersos en un área mayor, como el Cluster 1 (rojo), pueden representar grupos de datos con mayor variabilidad interna. En contraste, clusters más compactos, como el Cluster 2 (verde), representan grupos de datos más homogéneos.

La gráfica del clustering muestra que los datos se dividen en cuatro grupos distintos, con ciertas variaciones internas en cada grupo. La separación entre algunos clusters indica diferencias marcadas entre ellos, mientras que los clusters cercanos comparten algunas similitudes.

Conclusiones

- Las reglas de asociación indican una tendencia clara de diferencias de género en la demanda de reconsultas. Las mujeres tienden a presentar una mayor frecuencia de reconsultas en rangos altos (10,800 a 20,800), mientras que los hombres presentan una demanda menor o moderada de reconsultas, especialmente cuando no hay registros de emergencia. Este hallazgo sugiere que las mujeres podrían estar buscando más seguimiento médico o requieren atención continua en mayor medida que los hombres.
- Una de las reglas de asociación muestra que cuando el volumen de primeras consultas es alto (entre 18,700 y 53,100), la cantidad de reconsultas tiende a ser moderada a alta (12,900 a 17,900). Esto puede indicar que ciertos grupos de pacientes, después de su primera consulta, requieren una cantidad significativa de seguimiento. Este patrón sugiere la importancia de prever una demanda constante de reconsultas en áreas de atención inicial intensiva, lo cual es crucial para la planificación de recursos en hospitales.
- El clustering K-means reveló cuatro grupos distintos de pacientes (o casos), cada uno con características únicas en términos de sus componentes principales (que representan variables combinadas). Estos clusters podrían representar perfiles específicos de atención, como pacientes de bajo seguimiento, alto seguimiento, consultas emergentes y consultas de control. La separación entre clusters indica diferencias en las necesidades de atención, lo que permite segmentar a los pacientes y asignar recursos de forma más eficiente.

Bibliografía

- Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Nguyen, T., & Li, X. (2020). Applications of data mining techniques in healthcare data. En *Health Information Science* (pp. 246-257). Springer.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), 651-666.