UNIVERSIDAD NACIONAL DEL SUR

1ER CUATRIMESTRE 2024

Identificación de Patrones en Indicadores de Salud de Pacientes Diabéticos mediante técnicas de Clustering y detección de comunidades sobre redes de pacientes

Minería de Texto y Aprendizaje Automático

PRESENTADO POR

Niveyro Ignacio

Razuc Gonzalo

Índice

Introducción

El objetivo de este proyecto es identificar patrones en los indicadores de salud de pacientes diabéticos mediante técnicas de minería de texto y aprendizaje automático. La diabetes es una enfermedad crónica que afecta a aproximadamente el 9% de la población mundial y tiene un impacto significativo en la salud pública.

En este proyecto, aplicamos técnicas de análisis no supervisado, específicamente el algoritmo k-means para clustering, y utilizamos la librería NetworkX para el estudio de redes complejas. Además, implementamos la similitud por coseno para encontrar subgrupos más pequeños con alta similitud en características específicas, revelando comunidades más detalladas que no son evidentes en el análisis global.

Estos métodos nos permitieron agrupar a los pacientes en clusters basados en sus indicadores de salud y explorar las relaciones entre diferentes factores de salud dentro de estos grupos. El análisis busca no solo identificar grupos de pacientes con características similares, sino también entender cómo diferentes factores de salud están interrelacionados y cómo pueden influir en el desarrollo y manejo de la diabetes.

Recursos Utilizados

Para este proyecto utilizamos diferentes tipos de recursos, los cuales son:

- **Dataset:** diabetes_012_health_indicators_BRFSS2015.csv

  Este dataset incluye una amplia gama de indicadores de salud relacionados con la diabetes, como el estado de diabetes (no-diabetes,pre-diabetes o diabetes), presión arterial alta, niveles de colesterol, índice de masa corporal (IMC), hábito de fumar, actividad física, entre otros.

- **Librerías de Python:**

  pandas: Utilizada para la manipulación y análisis de datos, facilitando la carga, limpieza y transformación del dataset.

  numpy: Empleada para realizar operaciones matemáticas y manejar arreglos de datos.

  scikit-learn: Utilizada para aplicar el algoritmo de clustering k-means y otras técnicas de aprendizaje automático.

  matplotlib: Utilizada para la creación de gráficos y visualizaciones de datos.

  seaborn: Librería basada en matplotlib para la visualización de datos estadísticos.

  NetworkX: Utilizada para la creación, manipulación y análisis de redes complejas, facilitando el estudio de las relaciones entre los diferentes indicadores de salud.

Todo el trabajo fue realizado utilizando Google Colab, el mismo puede encontrarse **aquí**

Desarrollo

1 - Limpieza y Normalización de Datos

El primer paso en nuestro análisis fue la limpieza y normalización de los datos. Esto es fundamental para asegurar que los resultados sean precisos y significativos.

Los pasos que seguimos fueron los siguientes:

Eliminación de valores nulos:
Eliminamos las filas con valores nulos para garantizar que nuestro análisis se realice sobre un conjunto de datos completo.

Normalización de los datos:
Normalizamos los datos para asegurar que todas las características tengan la misma escala, lo que es crucial y necesario para el algoritmo de clustering que queremos aplicar.

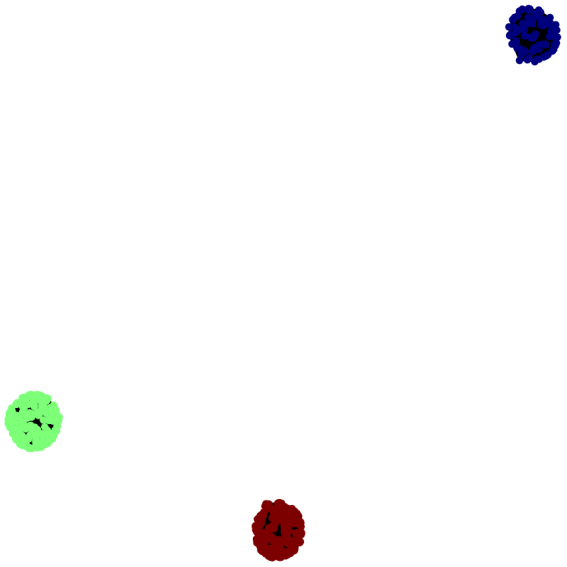
2 - Reducción del tamaño del dataset

Para facilitar el proceso de clustering y visualización, redujimos el tamaño del dataset a un subconjunto de 500 pacientes seleccionados aleatoriamente de los 5000 originales.
Redujimos el tamaño ya que si se utilizaban valores mucho más altos en la parte de creación del grafo, el proceso de ejecución demoraba mucho tiempo.

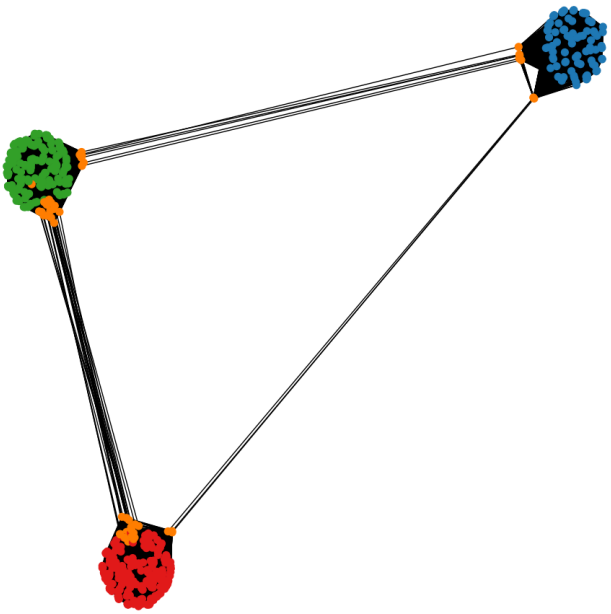
3 - Aplicación de Clustering (k-means)

Aplicamos el algoritmo k-means para agrupar los pacientes en clusters basados en sus indicadores de salud. Este algoritmo busca particionar los datos en k clusters, donde cada paciente pertenece al cluster con el centroide más cercano.

Visualización y Análisis de los resultados



Se puede observar que el algoritmo k-means encuentra 3 clusters bien diferenciados. Uno para los pacientes con diabetes, otro para los pacientes con pre-diabetes y otro para los que no poseen diabetes.



Usamos entonces la similitud por coseno para encontrar subgrupos (los nodos color naranja) más pequeños con alta similitud en características específicas, revelando comunidades más detalladas que no son evidentes

en el análisis global.

Estos nodos pueden representar grupos de pacientes que comparten características específicas de salud. Por ejemplo, podemos conectar pacientes con un buen nivel de salud, independientemente de si tienen diabetes, prediabetes o no tienen la enfermedad.

Análisis de la Salud General de los Pacientes:

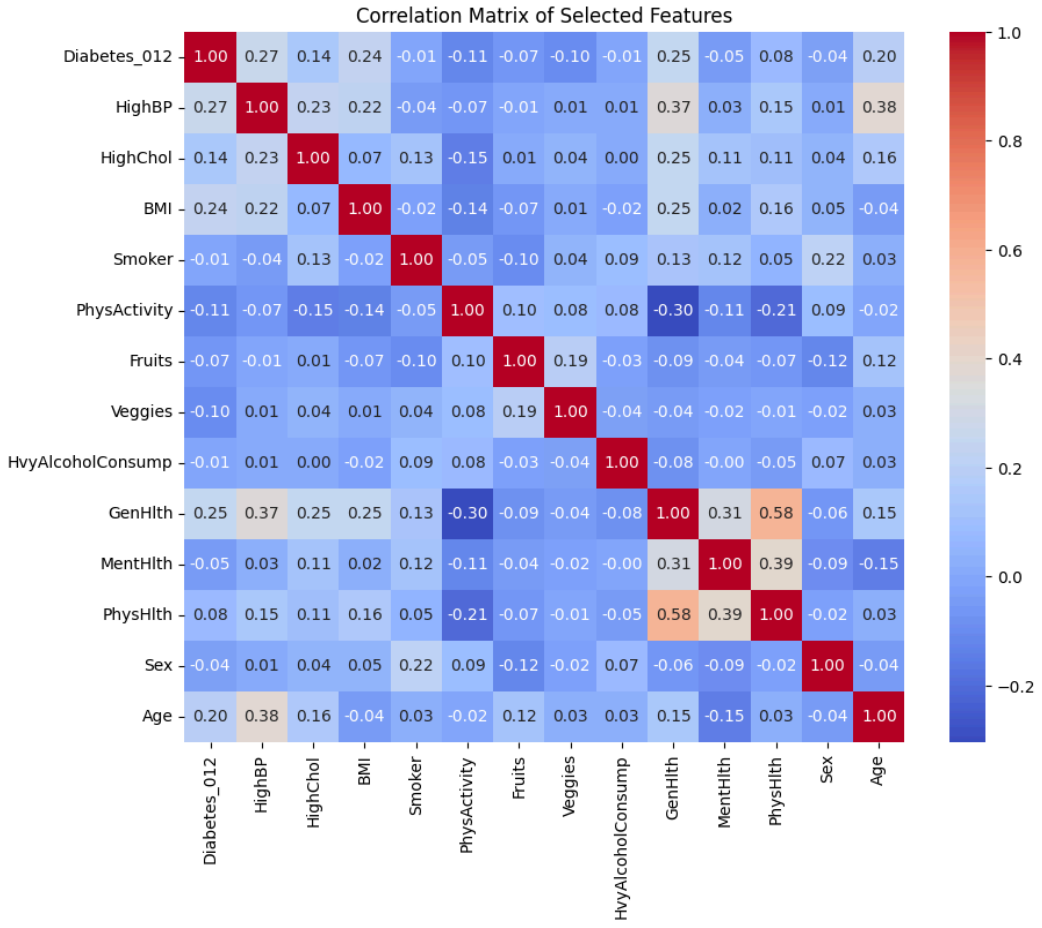
Exploramos la salud general de los pacientes en cada cluster para identificar diferencias clave.

Para ello, seleccionamos del dataset las características que creemos más importantes las cuales fueron:

Diabetes_012, HighBP, HighChol, BMI, Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, GenHlth, MentHlth, PhysHlth, Sex, Age

A partir de eso calculamos:

Matriz de Correlación



Una matriz de correlación es una herramienta estadística que muestra la relación entre diferentes variables en un conjunto de datos. Cada celda en la matriz representa el coeficiente de correlación entre dos variables, indicando la fuerza y la dirección de su relación.

Los valores de correlación pueden variar entre **-1 y 1**:

- **1** indica una correlación positiva perfecta, es decir, que a medida que una variable aumenta, la otra también lo hace de manera proporcional.
- **-1** indica una correlación negativa perfecta, es decir, que a medida que una variable aumenta,la otra disminuye de manera proporcional.
- **0** indica que no hay correlación entre las variables, es decir, los cambios de una variable no están directamente asociados con cambios en la otra.

Correlaciones notables

Diabetes_012 (0 = no diabetes, 1 = prediabetes, 2 = diabetes):

- **HighBP (0.27):** Personas con diabetes tienen más probabilidad de tener presión arterial alta.
- **BMI (0.24):** Correlación positiva con el índice de masa corporal, indicando que un mayor IMC está asociado con la diabetes.
- **GenHlth (0.25):** Correlación positiva con la salud general, indicando que las personas con diabetes tienden a calificar su salud general como peor.

HighBP (0 = no high BP, 1 = high BP):

- **Diabetes_012 (0.27):** Las personas con presión arterial alta tienen una mayor probabilidad de tener diabetes.
- **HighChol (0.23):** Correlación positiva con el colesterol alto.
- **GenHlth (0.37):** Correlación positiva con la salud general, indicando una peor salud general.
- **Age (0.38):** La presión arterial alta aumenta con la edad.

HighChol (0 = no high cholesterol, 1 = high cholesterol):

- **HighBP (0.23):** El colesterol alto y la presión arterial alta están correlacionados.

- **GenHlth (0.25):** El colesterol alto se asocia con una peor salud general.

BMI (Body Mass Index):

- **Diabetes_012 (0.24):** Mayor IMC está asociado con la diabetes.
- **HighBP (0.22):** Correlación positiva con la presión arterial alta.
- **GenHlth (0.25):** Un mayor IMC se asocia con una peor salud general.
- **PhysHlth (0.16):** Un mayor IMC está correlacionado con más días de problemas de salud física.

PhysActivity (0 = no, 1 = yes):

- **GenHlth (-0.30):** Mayor actividad física se asocia con una mejor salud general.
- **PhysHlth (-0.21):** Más actividad física se correlaciona con menos días de problemas de salud física.

GenHlth (Salud General: 1 = excelente, 2 = muy buena, 3 = buena, 4 = justa, 5 = pobre):

- **HighBP (0.37):** Peor salud general está asociada con presión arterial alta.
- **PhysActivity (-0.30):** Mejor salud general con más actividad física.
- **PhysHlth (0.58):** Fuerte correlación con días de problemas de salud física.
- **MentHlth (0.31):** Correlación positiva con días de problemas de salud mental.

MentHlth (Salud Mental: días con problemas de salud mental en los últimos 30 días):

- **GenHlth (0.31):** Peor salud general se asocia con más días de problemas de salud mental.
- **PhysHlth (0.39):** Más días de problemas de salud física se asocian con más días de problemas de salud mental.

PhysHlth (Salud Física: días con problemas de salud física en los últimos 30 días):

- **GenHlth (0.58):** Fuerte correlación con salud general, indicando que una peor salud física se refleja en una peor salud general.

- **MentHlth (0.39):** Problemas de salud física y mental están fuertemente correlacionados.

Por último, decidimos examinar la variación de la variable '**Diabetes012**', ya que esto podría revelar patrones interesantes sobre cómo los factores de salud y demográficos están asociados con diferentes estados de diabetes.

Para ello, procedemos a normalizar los datos utilizando las características seleccionadas. A continuación, calculamos la matriz de similitud por coseno y aplicamos el método de clustering jerárquico utilizando la matriz obtenida. Finalmente, calculamos los centroides de cada cluster resultante.

Los resultados fueron:

Cluster	Diabetes_012	HighBP	HighChol	BMI	Smoker	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	GenHlth	MentHlth	PhysHlth	Sex	Age	Count
0	-0.065609	-0.167171	-0.387657	-0.267810	-0.111051	0.199716	-0.305505	-1.382970	0.957033	-0.397426	-0.277757	-0.399759	0.072000	-0.053992	84
1	0.467122	0.391802	0.221499	0.324904	-0.069672	-0.378468	-0.058877	0.023471	-0.193247	0.658940	0.478140	0.632965	-0.223412	0.198114	206
2	-0.431981	-0.317470	-0.062217	-0.211591	0.112766	0.291373	0.179958	0.530164	-0.193247	-0.487418	-0.357930	-0.461004	0.190356	-0.172744	210

Cluster 0 (84 individuos):

1. **Baja Prevalencia de Diabetes y Prediabetes**: La puntuación en Diabetes_012 (-0.065609) es ligeramente negativa, indicando una menor prevalencia de prediabetes y diabetes.
2. **Baja Presión Arterial y Colesterol**: Las puntuaciones en HighBP (-0.167171) y HighChol (-0.387657) son negativas, indicando una menor prevalencia de hipertensión y colesterol alto.
3. **Bajo Índice de Masa Corporal (BMI)**: La puntuación en BMI (-0.267810) es baja, sugiriendo un índice de masa corporal más bajo.
4. **Bajo Consumo de Frutas y Vegetales**: Las puntuaciones en Fruits (-0.305505) y Veggies (-1.382970) son significativamente bajas, sugiriendo un bajo consumo de frutas y vegetales.
5. **Alto Consumo de Alcohol**: La puntuación en HvyAlcoholConsump (0.957033) es alta, indicando un alto consumo de alcohol.
6. **Salud General y Mental Mala**: Las puntuaciones en GenHlth (-0.397426) y MentHlth (-0.277757) son negativas, sugiriendo una percepción pobre de la salud general y más días con problemas de salud mental.

Cluster 1 (206 individuos):

1. **Alta Prevalencia de Diabetes y Prediabetes**: La puntuación en Diabetes_012 (0.467122) es significativamente mayor, indicando una alta prevalencia de prediabetes y diabetes.
2. **Alta Presión Arterial y Colesterol**: Las puntuaciones en HighBP (0.391802) y HighChol (0.221499) son significativamente mayores, indicando una mayor prevalencia de hipertensión y colesterol alto.
3. **Alto Índice de Masa Corporal (BMI)**: La puntuación en BMI (0.324904) es notablemente alta, sugiriendo un índice de masa corporal más alto.
4. **Bajo Nivel de Actividad Física**: La puntuación en PhysActivity (-0.378468) indica bajos niveles de actividad física.
5. **Problemas de Salud Mental y Física**: Las puntuaciones en MentHlth (0.478140) y PhysHlth (0.632965) son altas, indicando más días con problemas de salud mental y física.

Cluster 2 (210 individuos):

1. **Menor Prevalencia de Diabetes y Prediabetes**: La puntuación en Diabetes_012 (-0.431981) es significativamente menor, sugiriendo una menor prevalencia de prediabetes y diabetes.
2. **Alta Actividad Física**: La puntuación en PhysActivity (0.291373) es significativamente mayor, indicando altos niveles de actividad física.
3. **Buen Consumo de Frutas y Vegetales**: Las puntuaciones en Fruits (0.179958) y Veggies (0.530164) son significativamente altas, sugiriendo un buen consumo de frutas y vegetales.
4. **Consumo Moderado de Alcohol**: La puntuación en HvyAlcoholConsump (-0.193247) indica un consumo moderado de alcohol.
5. **Salud General Pobre**: La puntuación en GenHlth (-0.487418) es notablemente baja, sugiriendo una percepción pobre de la salud general.
6. **Problemas de Salud Mental y Física**: Las puntuaciones en MentHlth (-0.357930) y PhysHlth (-0.461004) son bajas, indicando menos días con problemas de salud mental y física.

Conclusiones

El uso de técnicas de minería de datos y aprendizaje automático, específicamente el algoritmo de clustering k-means y la similitud por coseno, permitió identificar patrones y subgrupos significativos dentro de la población de pacientes. Estos métodos revelaron clusters bien diferenciados que reflejan distintos estados de salud y hábitos de vida.

Los resultados obtenidos en resumen, arrojan los siguientes datos:

- **Cluster 0**: Baja prevalencia de diabetes y prediabetes, menor prevalencia de hipertensión y colesterol alto, bajo índice de masa corporal, bajo consumo de frutas y vegetales, alto consumo de alcohol, y una percepción pobre de la salud general y mental.
- **Cluster 1**: Alta prevalencia de diabetes y prediabetes, hipertensión y colesterol alto, alto índice de masa corporal, bajos niveles de actividad física, y más días con problemas de salud mental y física.
- **Cluster 2**: Menor prevalencia de diabetes y prediabetes, alta actividad física, buen consumo de frutas y vegetales, consumo moderado de alcohol, percepción pobre de la salud general, y menos problemas de salud mental y física.

Creemos que estos resultados pueden ayudar a los profesionales de la salud a identificar grupos de riesgo y desarrollar intervenciones más efectivas y personalizadas. Por ejemplo, los pacientes del Cluster 1 pueden beneficiarse de programas que aborden tanto la salud mental como la física, mientras que los del Cluster 0 podrían necesitar mayor educación sobre hábitos de vida saludables.