ANÁLISIS EXPLORATORIO DE DATOS (EDA)

APLICACIÓN EN MUESTRAS DE DIABETES PROVENIENTES DE UN ESTUDIO METABOLÓMICO

CLAUDIA BALDERAS

Contenido

I. Introducción

De acuerdo a la Organización Mundial de la Salud, más de 300 millones de personas en el mundo tienen diabetes, siendo probable que se duplique este dato antes del 2030. El aumento de su incidencia nos lleva a buscar la comprensión de los efectos globales de la diabetes, para ampliar nuestro conocimiento de sus causas, su progresión y su tratamiento.

Una esperanza de futuro para el estudio de la diabetes y de distintas enfermedades es el análisis metabolómico, disciplina dedicada al estudio global de los metabolitos presentes en un sistema biológico.

En los últimos años las nuevas técnicas de análisis avanzado y la informática han facilitado significativamente una nueva generación en la investigación biomédica, incluyendo la biología de sistemas. Sin embargo, el tener que lidiar con un mayor conjunto de datos experimentales, hace necesario una considerable interacción entre los científicos, desarrolladores, usuarios, métodos estadísticos y expertos en bioinformática.

Los estudios metabolómicos generan grandes conjuntos de datos que suponen un gran desafío para el análisis. Cuando se lleva a cabo un estudio de este tipo se persiguen principalmente objetivos como visualizar diferencias globales, tendencias y relaciones entre las muestras y las variables, discernir qué metabolitos son responsables de esos cambios y la construcción de modelos para la predicción de los efectos biológicos en nuevas muestras.

El análisis de datos de manera manual es posible en muy pocos casos, es extremadamente lento y no puede recomendarse como estrategia. Por lo tanto, la inclusión de diferentes métodos exploratorios de datos y el análisis multivariante, podrían proporcionar una reducción de la dimensionalidad de los datos, lo que se reflejaría en una mejor representación final e interpretación.

II. Contexto de los datos de estudio

Los datos del estudio se extraen de un fichero con información de un análisis metabolómico del suero humano, realizado en 60 muestras de pacientes sanos, 60 muestras de pacientes con prediabetes y 60 muestras de pacientes con diabetes tipo 2, los cuales estaban sometidos a un estricto control metabólico.

III. Objetivos e hipótesis

Partiendo de la hipótesis de que los pacientes con diabetes tienen una huella metabólica distinta a los individuos sanos, independientemente que se encuentren o no bajo un estricto control metabólico. El objetivo principal de este trabajo se centrará en el Análisis exploratorio de datos (EDA) como herramienta útil para encontrar patrones que expliquen la variabilidad en los metabolitos de cada grupo de estudio y que permitan avanzar en el conocimiento de la diabetes.

Tomando en cuenta algunos parámetros dados en la matriz de datos como son: el índice de masa corporal (BMI), la hemoglobina glicosilada (HbA1c), el sexo (femenino, masculino) y su

estado de salud (sano, prediabetes, diabetes). Se formularon una serie de hipótesis a comprobar:

- El índice de BMI mayor y niveles altos de HbA1c en los grupos de prediabetes y diabetes, respecto a los sanos.
- El género podría influir en la enfermedad.
- Existen diferencias en el perfil metabólico entre individuos sanos y pacientes con prediabetes y diabetes, a pesar que estos últimos se encuentren bajo un estricto control metabólico.

IV. Desarrollo del Proyecto

A. Elección y obtención del set de datos.

La elección de la temática utilizada en esta memoria se basó en que es el área de trabajo en la que me desarrollo a nivel laboratorio y que me gustaría continuar como Data Scientist en un futuro. Debido a que estoy empezando a crear mi portfolio en Github, consideré como buena opción practicar con ese tipo de datos.

El conjunto de datos utilizados en este trabajo esta sacado de kaggle y se encuentra disponible en el siguiente enlace (https://www.kaggle.com/datasets/desertman/human-serum-metabolome-variability). Lleva por título "*Human Serum Metabolome Variability",* los datos se encuentran en formato de tipo Excel.

Los datos son parte de un estudio que ya está publicado como artículo de investigación, sin embargo el objetivo del artículo difiere de los presentados en esta memoria, debido a que el dataset estaba incompleto.

Artículo: Agueusop I, Musholt PB, Klaus B, Hightower K, Kannt A. Short-term variability of the human serum metabolome depending on nutritional and metabolic health status. Sci Rep. 2020 Oct 1;10(1):16310. doi: 10.1038/s41598-020-72914-7. PMID: 33004816; PMCID: PMC7530737.

B. Análisis del set de datos

El análisis de datos que se encuentra en el notebook, se dividió en las siguientes secciones:

Importaciones de librerías:

Numpy y pandas para la evaluación de los datos, a nivel de visualización se utilizó: Matplotlib, seaborn, plotly, y cimb_lite, siendo esta última una librería de un repositorio correspondiente a The Centre for Integrative Metabolomics and Computational Biology (https://github.com/orgs/CIMCB/repositories).

Exploración y análisis de los datos:

En esta sección extrajeron los datos correspondientes a las 3 hojas (“simple_metadata”, “data matrix” y “data_dictionary”) que contenía el dataset.

La primera hoja se denominó “sample_metadata”, la cual contenía datos con información general del estudio, estaba compuesta de 16 columnas y 181 filas. Se utilizaron las columnas BMI, sexo, hba1 y health_status, para estudiar la hipótesis 1 y 2 (información más detallada de los features, se encuentra en el notebook).

La hoja “Data_matrix” correspondia a los metabolitos encontrados, 181 filas (muestras) y 1486 variables (metabolitos). Estos datos se analizaron mediante un análisis de componentes principales (PCA) y con estadística clásica para sacar aquellos compuestos significativos solo en los grupos de diabetes y sanos. Los resultados nos permitieron evaluar la hipótesis 3.

Por último la hoja “data_dictionary”, compuesta con información de la identificación de los compuestos medidos, se utilizó únicamente a modo de visualización, para la elaboración de un gráfico de la clase de compuestos.

V. Conclusiones

Este análisis permitió aceptar dos de las hipótesis planteadas en un principio. Se confirmó la utilización de la HbA1c como marcador en el diagnóstico de la diabetes. Se confirmó en el grupo de diabetes que a pesar de estar controlados, tiene un perfil metabólico bastante marcado en comparación con la diabetes, originado por una serie de compuestos que nos podrían permitir avanzar en el estudio de la enfermedad.