01

Análisis parcial de datos genómicos crudos

Esta etapa fue propuesta para la metodología DSM-BCD para los datos de tipo genómico. Lo anterior, debido a que el EDA tradicional parte del análisis descriptivo, y en este caso los tipos de datos fueron obtenidos de diferentes fuentes médicas sin una estructura fija ni estándar.

02

Transformación de variables genómicas

En esta etapa, se procesaron las variables para que la información quedara estandarizada y así garantizar que los resultados generados por los modelos de ML fueran consistentes y veraces:

- Renombramiento de columnas
- Estandarización de datos genómicos
- Re-ajuste del tipo de variable

03

Tratamiento de datos ausentes

En esta etapa, se detectaron, imputaron y posteriormente se eliminaron las variables innecesarias identificadas en el análisis parcial de datos genómicos crudos, por medio de algoritmos.

- Detección
- Imputación
- Eliminación
- Consistencia

04

Análisis Descriptivo

En esta etapa, se realizó un análisis descriptivo para detectar cual era el comportamiento de las 41 variables seleccionadas para el entrenamiento de los modelos de ML. Dado lo anterior, se extrajeron las características más representativas relacionadas con las preguntas planteadas en el BCQM.

05

Correlación de Variables

En esta etapa, se utilizó el coeficiente de correlación de Spearman para determinar si existía una relación lineal o no lineal expresada en un rango de [−1, 1] de las 41 variables seleccionadas para el entrenamiento de los modelos de ML. Cabe resaltar, que se eligió este tipo de correlación debido a que no todas las variables numéricas del conjunto de datos se distribuían normalmente.