Tarea 1 Machine Learning

Felipe Carriel

22 de Septiembre 2025

1 Introducción

En esta tarea se realiza un análisis exploratorio de dos datasets, incluyendo la matriz de correlación para evaluar la redundancia entre descriptores y ejemplos, usando coeficientes de Pearson, Kendall y Spearman. Además, se aplica un análisis de componentes principales (PCA) para identificar la influencia de cada variable, la varianza explicada por componente y la varianza acumulada.

2 Análisis del Dataset 1: Maternal Health Risk

El dataframe utilizado contiene información clínica de los pacientes, incluyendo edad, presión arterial sistólica y diastólica, nivel de glucosa en sangre, temperatura corporal y frecuencia cardíaca. La variable de interés RiskLevel clasifica a los pacientes en tres categorías de riesgo (0 = bajo, 1 = medio, 2 = alto) mediante un mapeo numérico.

2.1 Matriz de Correlación

Se presentan los mapas de calor de las distintas correlaciones entre variables, para la matriz de Pearson se nota que normalizando los datos con minmax no cambian las correlaciones entre variables, si bien las distintas correlaciones pueden revelar diferentes patrones en este caso las correlaciones entre variables parece ser similar en todos los escenarios, se nota además que variables como SystolicBP y DiastolicBP presentan una alta correlación, y que la temperatura corporal presenta una correlación negativa con respecto al resto de variables.

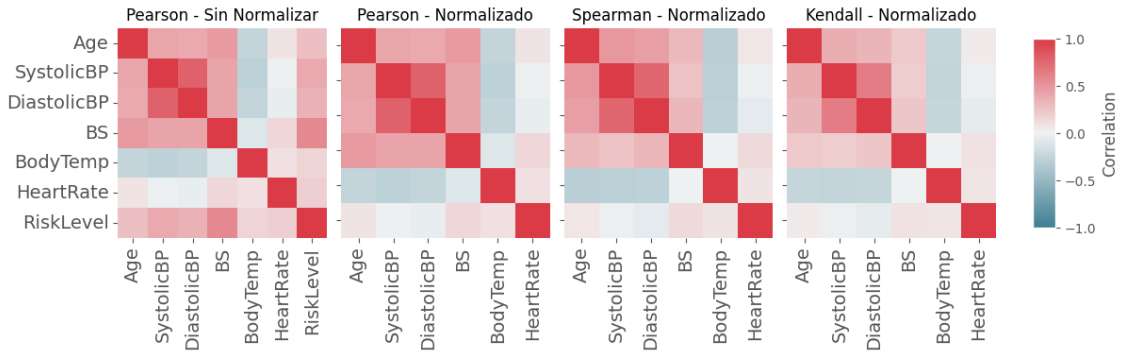

Figure 1: Correlación entre variables para datos sin normalizar y normalizados

2.2 Análisis de Componentes Principales (PCA)

Se muestran los mapas de calor de las 6 componentes principales para las características, algunas características comparten magnitudes en ciertas componentes, se muestra además como aumenta la varianza a medida que se agregan componentes llegando a 1, finalmente se muestra un grafico de 2 componentes, notando que la clase high_risk (en este caso 2:amarillo) se separa de las clases mid_risk y low_risk.

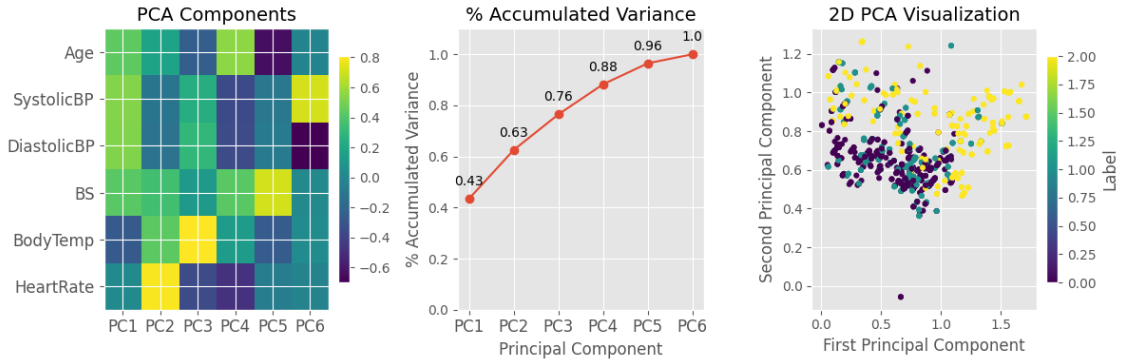

Figure 2: Mapa de calor de componentes principales, Varianza acumulada y visualizacion en 2D

3 Análisis del Dataset 2: COAD

El dataframe corresponde a un conjunto de datos de expresión génica, donde cada fila representa una muestra y cada columna, excepto target, representa un gen específico con su nivel de expresión cuantificado. La columna target indica la clase o condición asociada a cada muestra (por ejemplo, 0/1 para control/enfermo). En total, el dataset contiene 415 muestras y 25.151 genes, constituyendo la base para análisis estadísticos, correlacionales y modelado predictivo de expresión génica.

3.1 Correlación de Genes

Como el tamaño la cantidad de características son bastante mayor que en el caso anterior se muestran las correlaciones mas fuertes entre los Genes y la clase target, es posible notar que las correlaciones no cambian al escalar los datos y vemos ademas que distintas correlaciones encuentran distintos patrones, se muestran las correlaciones más fuertes postivias y negativas.

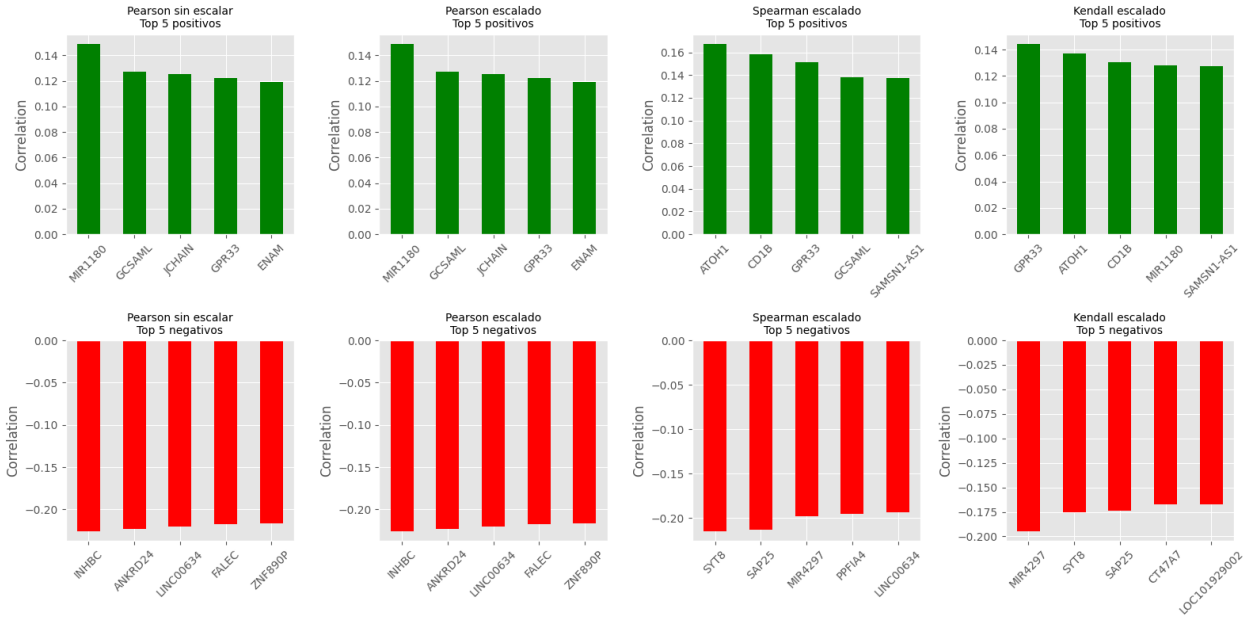

Figure 3: Correlaciones de genes con la variable objetivo

3.2 Análisis de Componentes Principales (PCA)

Se analizan las componentes principales de los genes, se muestran las primeras 10 componentes, su mapa de calor y varianza acumulada, se muestra la visualización 2D de componentes entre genes y es posible apreciar que no hay separación evidente debido a la complejidad de los datos.

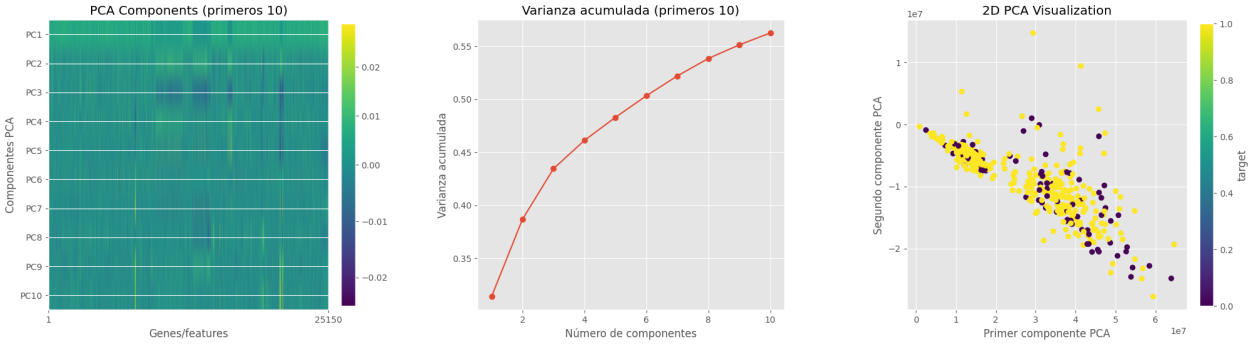

Figure 4: Enter Caption

4 Conclusiones

El análisis exploratorio mediante correlaciones y PCA permitió identificar patrones y relaciones entre variables en ambos datasets. Las correlaciones mostraron redundancia en algunas variables del dataset clínico y distintas asociaciones en el dataset de expresión génica. El PCA evidenció que, mientras en el dataset clínico es posible distinguir clases, en el de expresión génica la alta dimensionalidad dificulta la separación. En general, correlaciones y PCA son herramientas útiles para entender la estructura de los datos y orientar análisis posteriores.