

Taller 5: Análisis exploratorio de variables

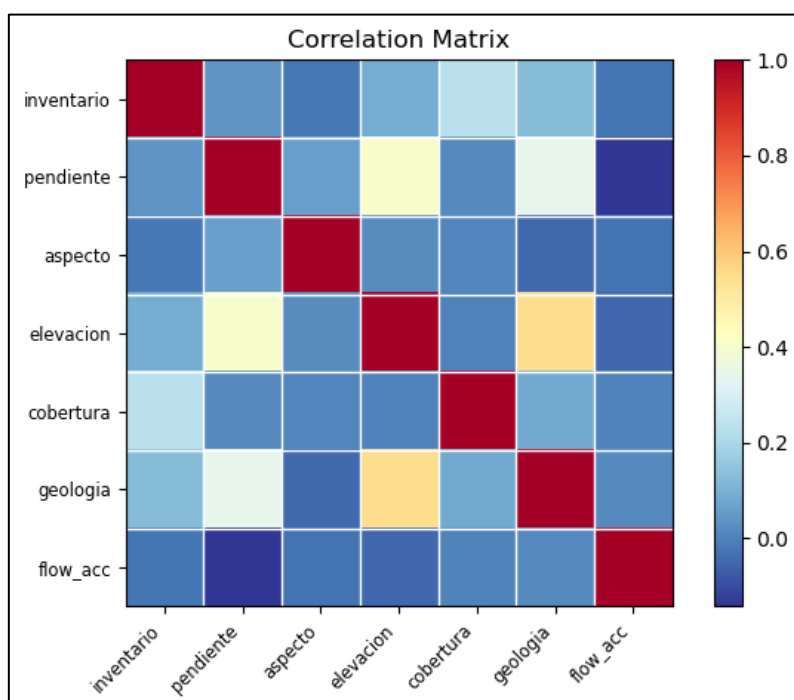
El objetivo principal de éste taller es seleccionar las variables con las cuales se van a trabajar los diferentes modelos (Heurístico, bivariado, multivariado, con base física, entre otros) y comparar la información que cada una de ellas nos ofrece, para verificar si realmente tenemos una base de datos lo suficientemente no correlacionable como para hacer un modelo adecuado. Si 2 variables comparten la misma información y se emplean ambas en los modelos, prácticamente no estaríamos modelando con variedad de información.

Se cargan entonces en el código de jupyter, los mapas de:

- Pendiente
- Aspecto
- Curvatura
- Flujo Acumulado
- Geología
- Elevación
- Cobertura
- Inventario MenM

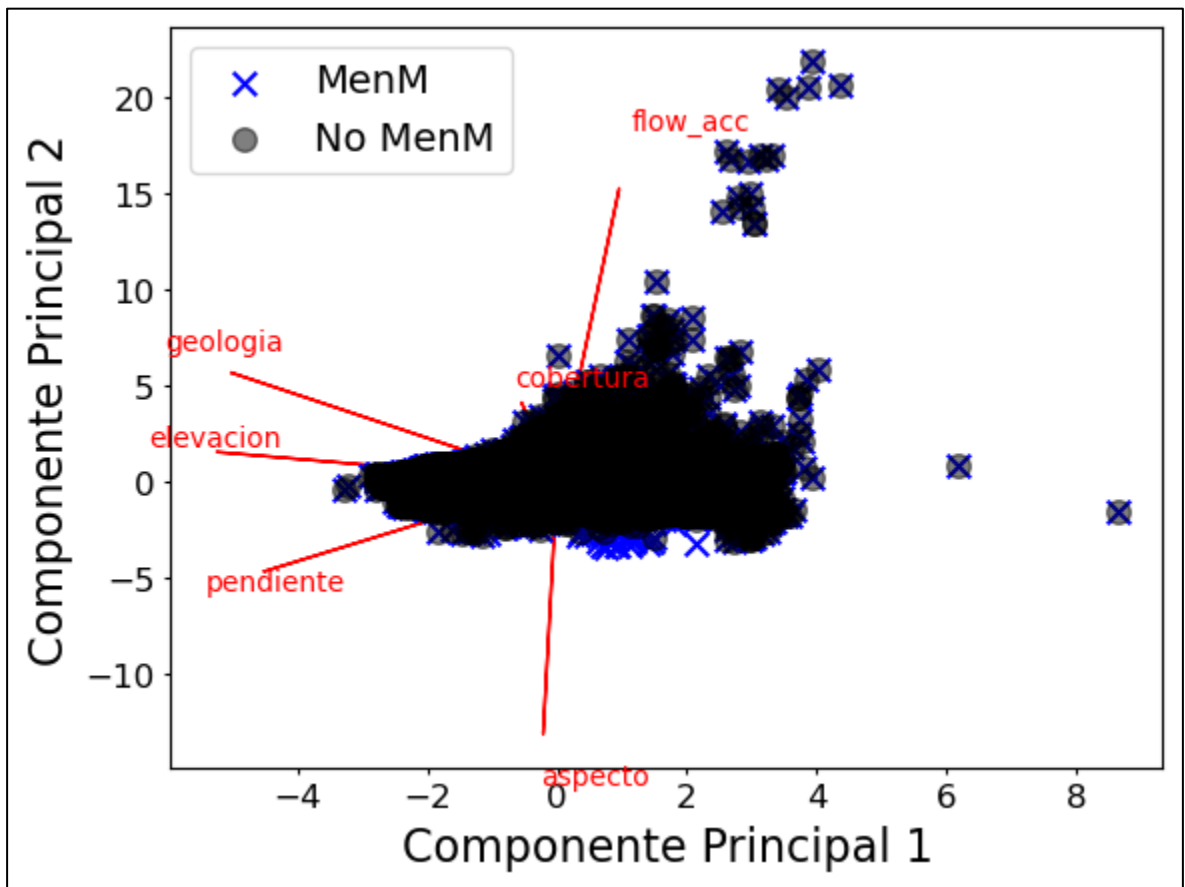
De éstos mapas, los que corresponden o son compatibles (por la longitud del vector) serían todos exceptuando el de curvatura que posee algunos inconvenientes (posiblemente en el momento de generarlo en la herramienta GIS) y por lo cual no puede incluirse en el DataFrame posteriormente para realizar las evaluaciones necesarias.

Se obtiene entonces, entre otros muchos resultados (diagrama de cajas, de campana, histogramas) la matriz de correlación tal que:



Según la matriz de correlación, tenemos que las variables que mas se correlacionan, son pendiente-elevación, pendiente-geología y elevación-geología, siendo mas alta, la correlación de las variables elevación-geología, por lo que el análisis exploratorio empieza a indicar que posiblemente ambas variables contengan información similar con base a las celdas que contienen movimientos en masa.

Finalmente se emplea el método PCA o de componentes principales para evaluar el comportamiento de cada clase y obtenemos el siguiente resultado:



Se puede comparar los comportamientos de cada variable, pudiendo obtener la siguiente información de la grafica:

1. Vemos que las variables de *geología*, *pendiente* y *elevación*, se comportan de una manera similar por lo que para el modelo podría elegirse solamente la elevación o se podría evaluar las variables *pendiente* y *geología* omitiendo la de *elevación*. En el caso de mi modelo se elegirán *geología* y *pendiente*.
2. La variable de *cobertura* no representa mayor información y se debe entrar a evaluar los datos de esa información, con el objetivo de incluir mayor cantidad o revisar su veracidad, por lo que ésta variable no se tendrá en cuenta para la evaluación de los modelos posteriores.
3. Las variables de *Flow_acc* y *aspecto* parecen mostrar cierta correlación inversa, pero se tendrán en cuenta para la evaluación posterior de los modelos.

Algunas conclusiones extra que se pueden extraer de la información obtenida de los histogramas y diagramas de caja principalmente es:

- La mayoría de los movimientos en masa que se tienen inventariados, se presentan en pendientes que varían de los 10 a los 30° de inclinación.
- La mayoría de los movimientos en masa que se tienen inventariados, se presentan en laderas con aspectos que varían desde los 75 a los 275, es decir, aquellas que comprenden sobre todo todo el hemisferio sur.