RIO MAROCASO

Taller 5 – Selección de variables

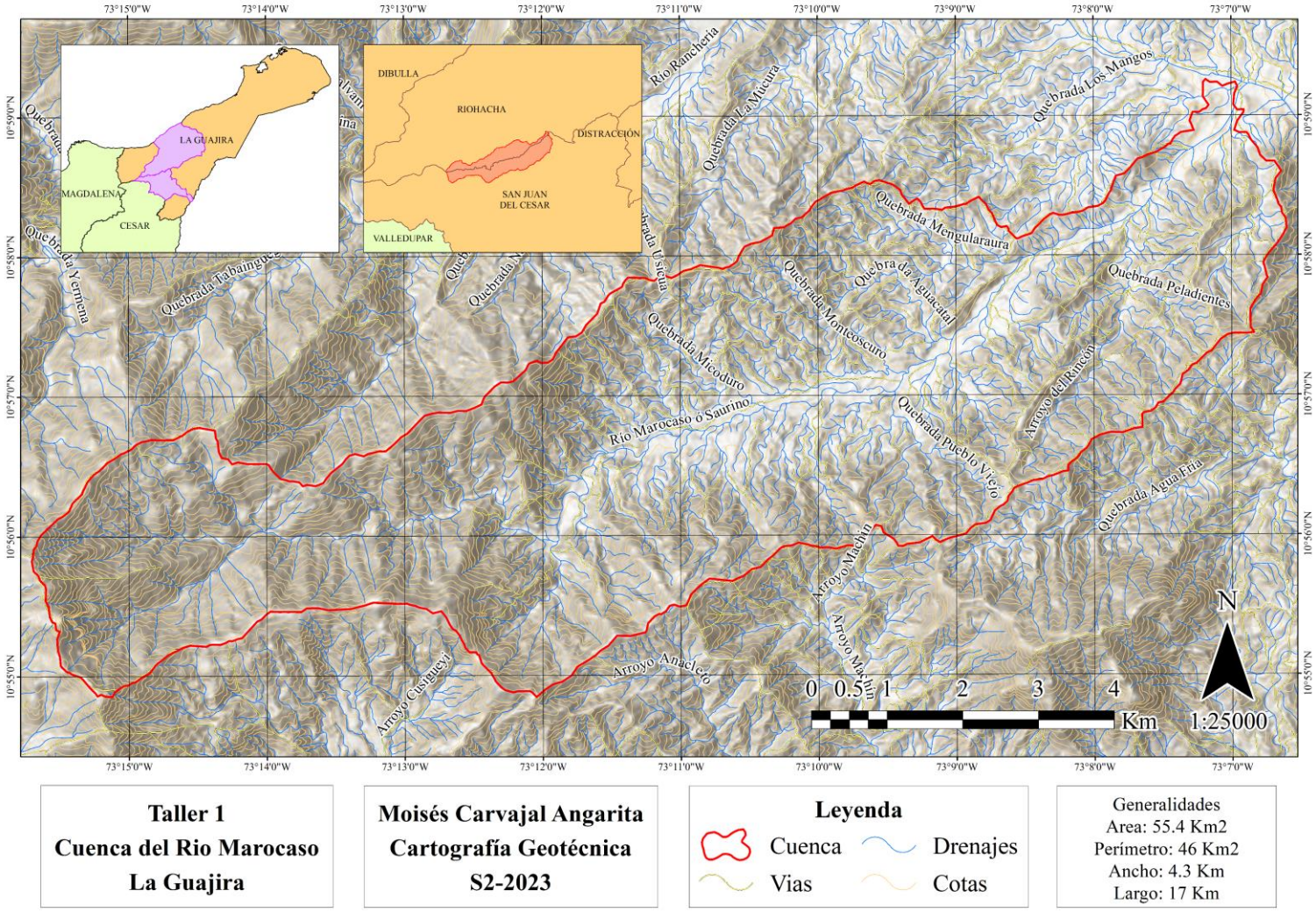
Moisés Carvajal Angarita

2023-S2

Generalidades

La cuenca del río Marocaso se encuentra en el Departamento de la Guajira a una altitud de aproximadamente 658 m.s.n.m., entre los municipios de Rihoacha y San Juan del Cesar.

Nace directamente desde las montañas de la Sierra Nevada de Santa Marta y es un afluente directo del Rio Ranchería.



Generando el Data Frame y estadísticos

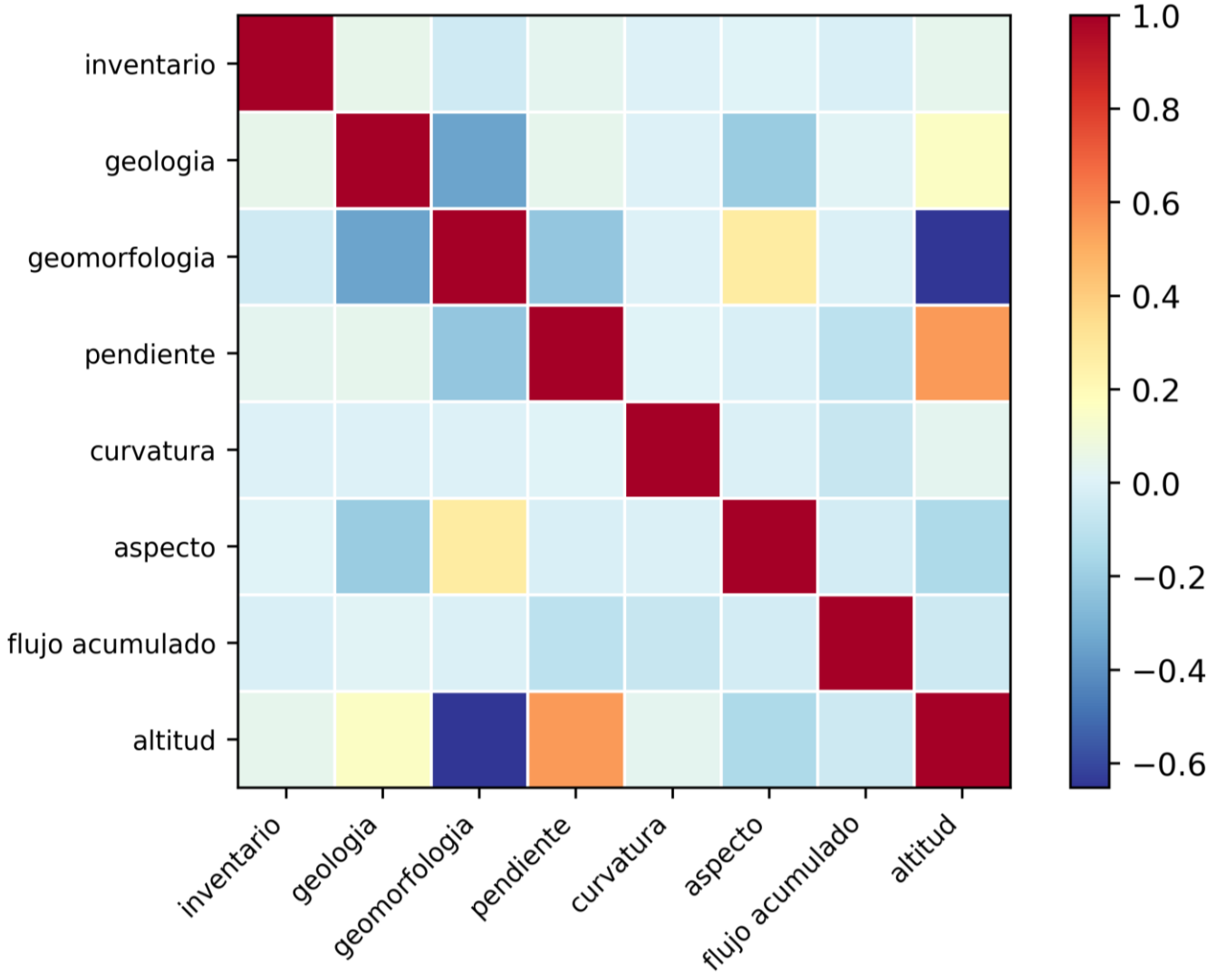
Lectura necesaria, para establecer equivalencia entre los datos.

	inventario	geologia	geomorfologia	pendiente	curvatura	aspecto	flujo acumulado	altitud
354614	0.0	1.0	1.0	14.880050	-0.64	70.201126	0.0	2562.0
354615	0.0	1.0	1.0	3.618883	-1.92	18.434948	0.0	2560.0
354616	0.0	1.0	1.0	15.793169	-0.64	278.130096	1.0	2562.0
354617	0.0	1.0	1.0	9.664409	7.68	273.366455	0.0	2568.0
354618	0.0	3.0	9.0	15.811913	2.56	132.137589	0.0	1817.0

Primeramente, se realiza un análisis exploratorio a partir de la generación de estadísticos para los datos contenidos en el Data Frame. Es evidente, que los estadísticos para variables categóricas como, inventario, geología y geomorfología no es lo ideal.

	count	mean	std	min	25%	50%	75%	max
inventario	354619.0	0.019514	0.138323	0.000000	0.000000	0.00000	0.000000	1.000000
geologia	354619.0	2.559995	0.801421	1.000000	3.000000	3.00000	3.000000	3.000000
geomorfologia	354619.0	5.984978	3.112072	1.000000	3.000000	5.00000	9.000000	10.000000
pendiente	354619.0	23.014332	11.077343	0.000000	14.281795	23.03548	31.304279	68.119446
curvatura	354619.0	-0.000430	1.582084	-48.020626	-0.640000	0.00000	0.640000	24.329609
aspecto	354619.0	170.882874	115.122215	-1.000000	78.070679	145.30484	293.198578	359.998260
flujo acumulado	354619.0	383.094238	6167.555664	0.000000	1.000000	4.00000	12.000000	215838.000000
altitud	354619.0	1206.386597	452.157532	567.000000	855.000000	1073.00000	1486.000000	2734.000000

Matriz de correlación



Se realiza la matriz de correlación para todas las variables.

Deja en evidencia que las correlaciones más altas están presentes con las mismas variables.

A pesar de que no logra ser tan evidente, se logra resaltar relaciones entre la pendiente con la altitud, así como la geomorfología con el flujo acumulado y la altitud.

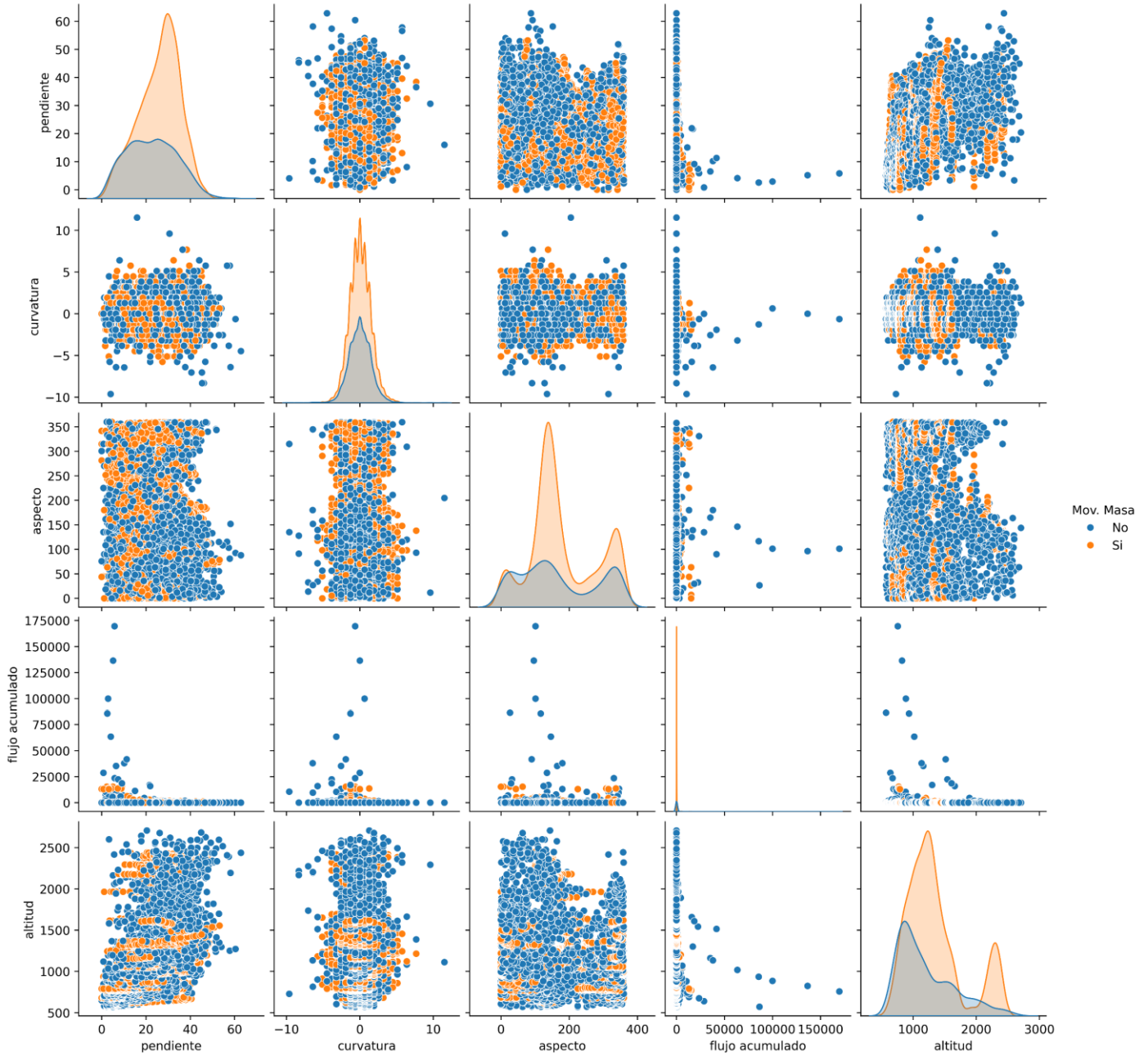
Sin embargo, es un acercamiento previo que nos permite visualizar que también existe una relación inversa entre algunas variables

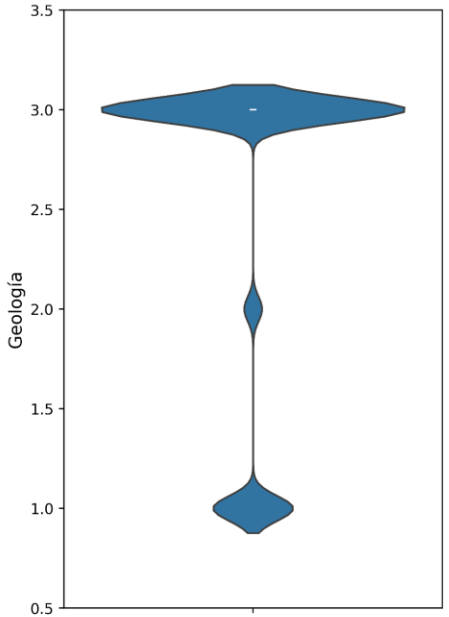
Relación de datos por densidad y movimientos en masa

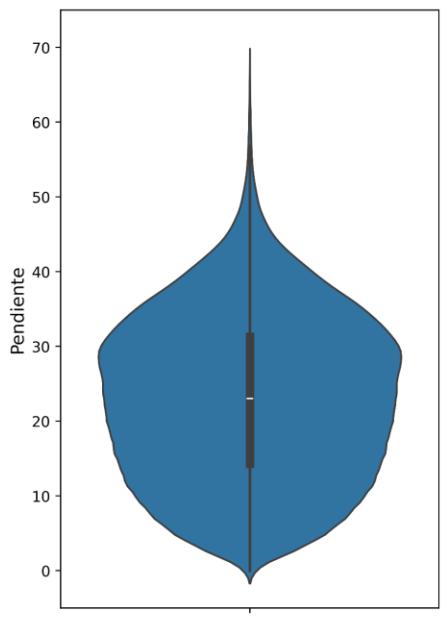
Se generó una matriz de comparación para todas la variables continuas presentes en el Data Frame, pero clasificando las celdas en función de si hay o no Movimientos en Masa.

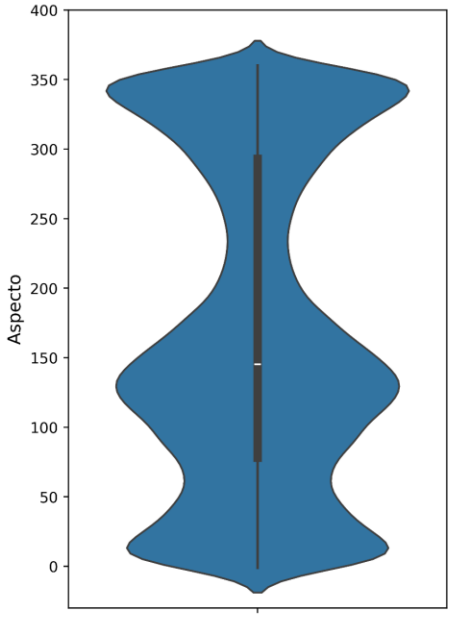
Cada gráfico individual en la diagonal es un histograma de densidad de Kernel.

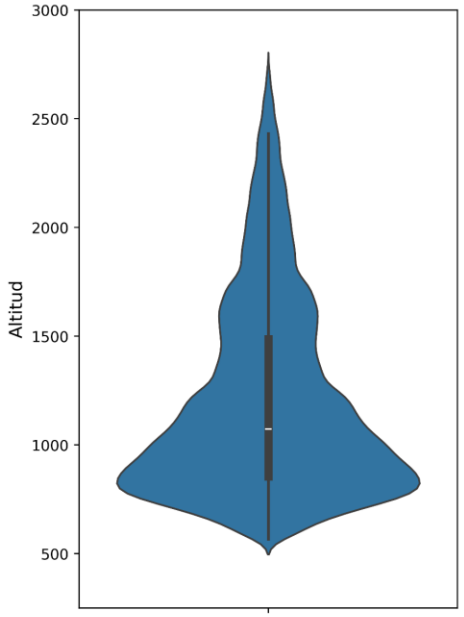
En esta matriz es un poco complejo la extracción de datos, por la falta de uniformidad que exponen. Lo que es evidente que en los gráficos de dispersión se logra observar que la gran mayoría de datos presentan una varianza considerable.

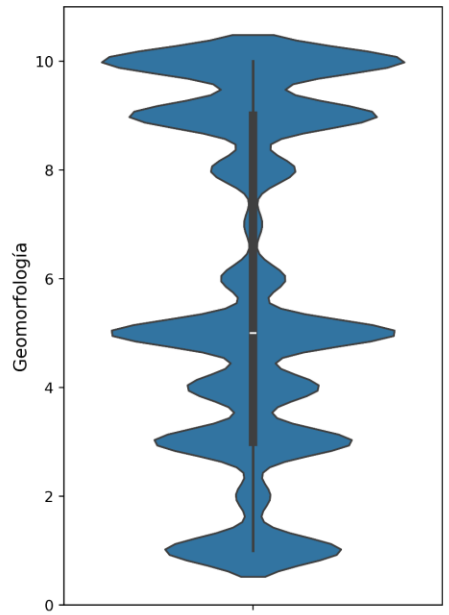


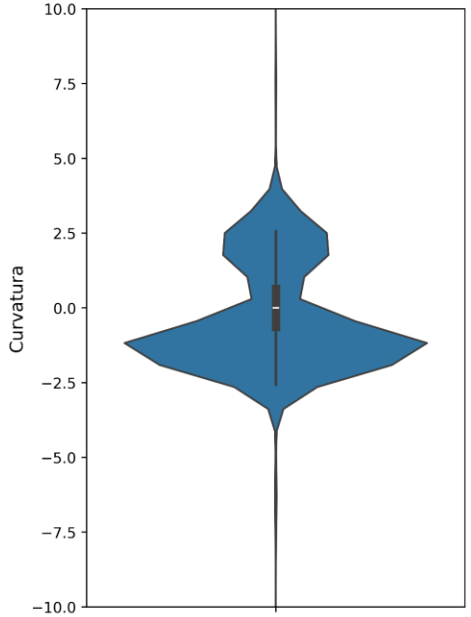


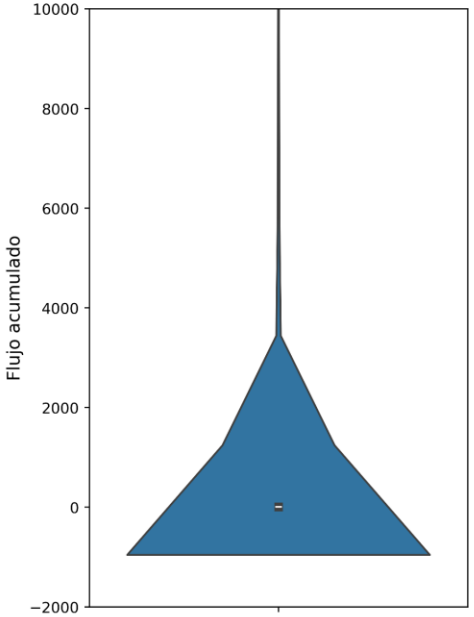








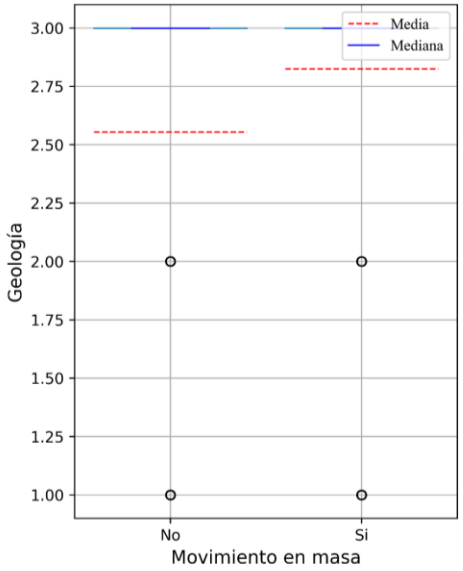


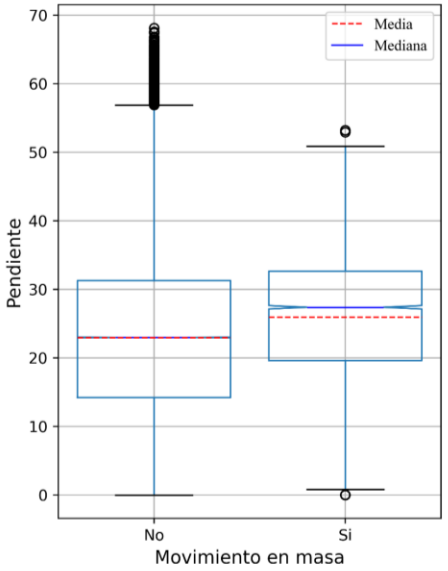


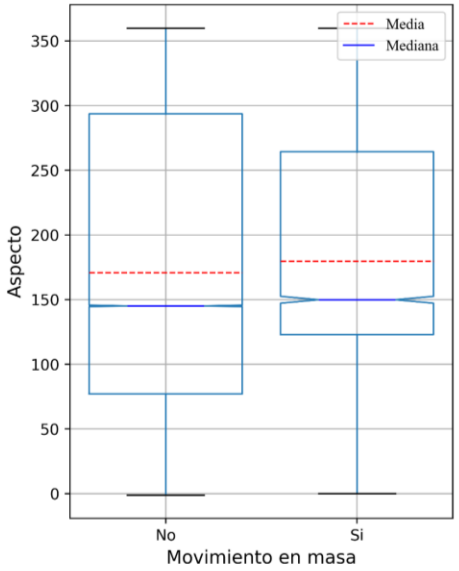
Gráficos de Violín

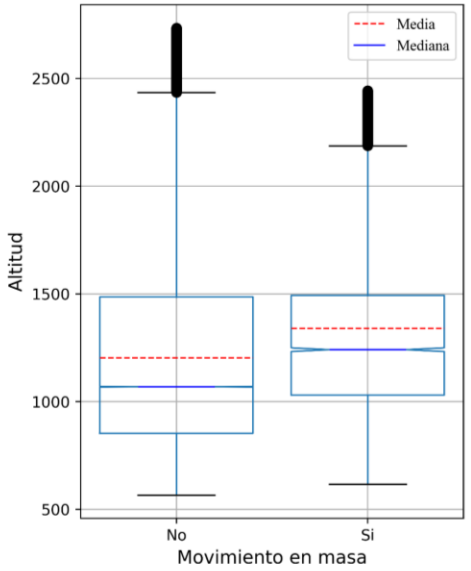
Los gráficos de violín permiten un acercamiento a la distribución de los datos introducidos. Son de utilidad para comprender tendencias de datos y máximos extremos.

Para las variables cualitativas, no se muestra de manera correcta, pues solo habrá datos en la reclasificación respectiva.

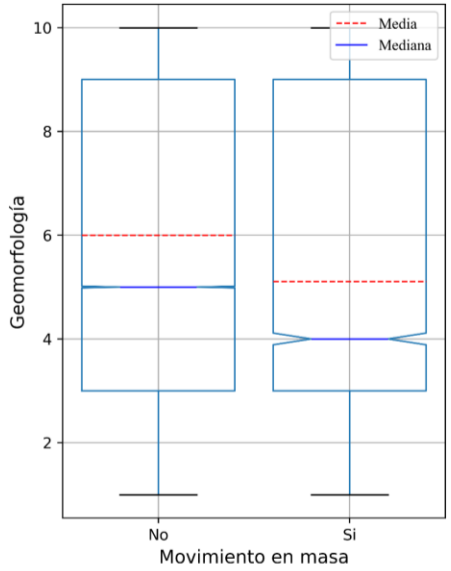


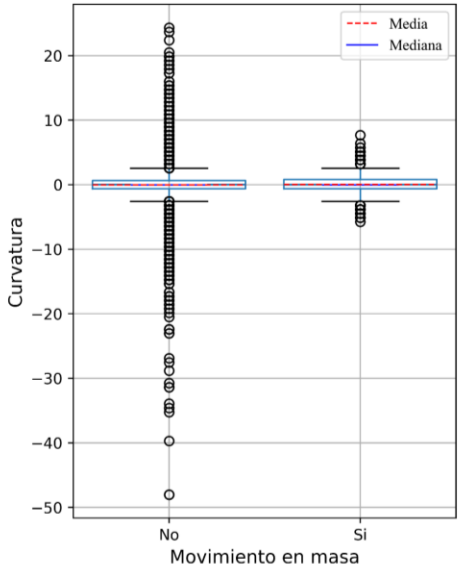


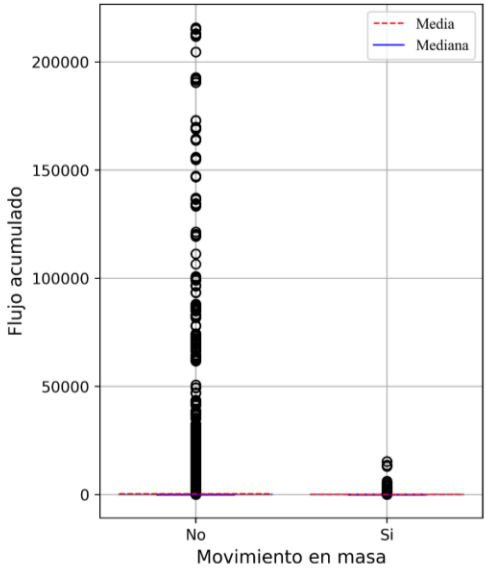




Diagramas de Caja

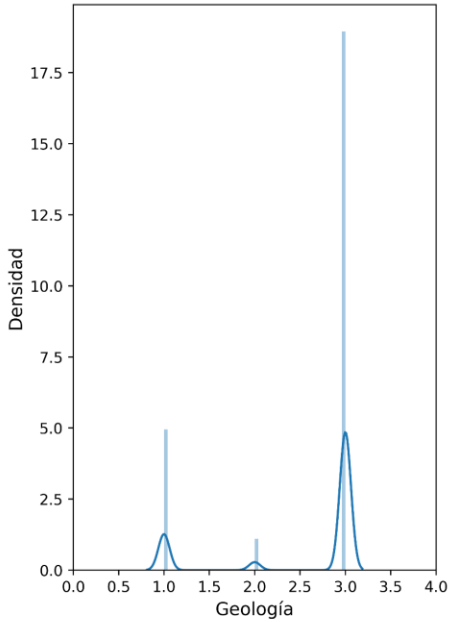


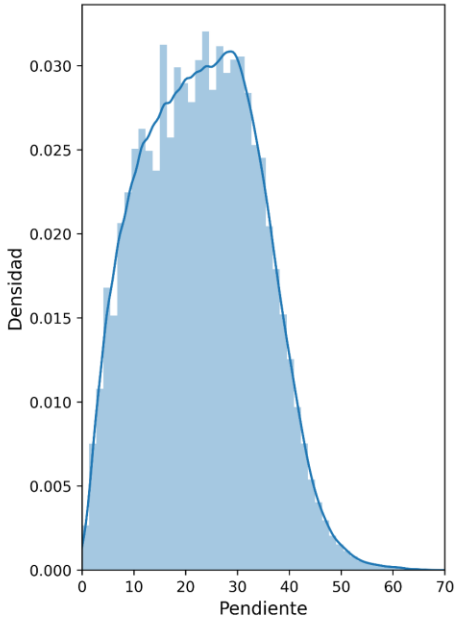


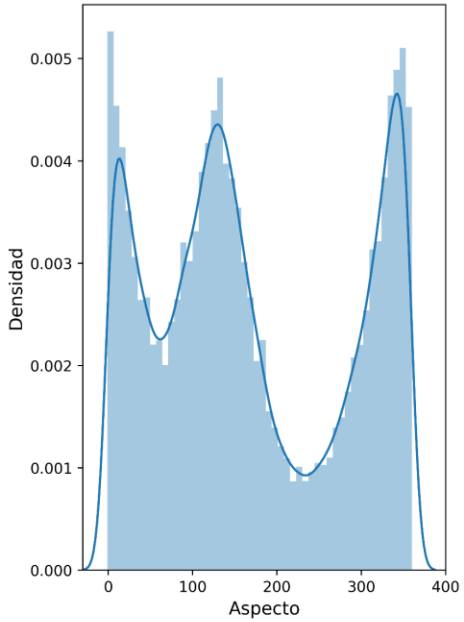


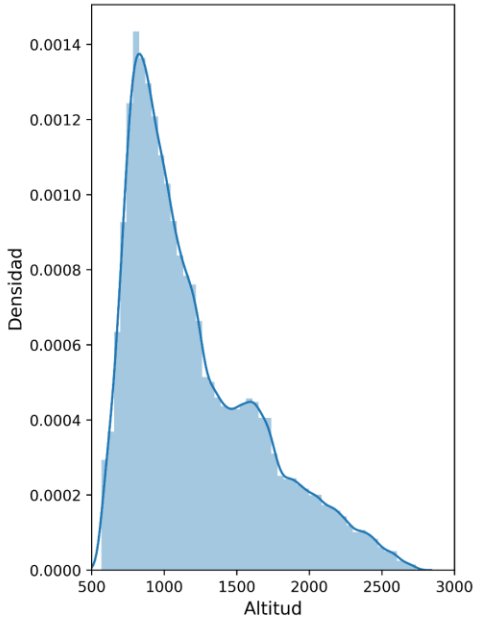
Los gráficos de caja permiten visualizar datos importantes como la media, mediana, los cuartiles y los valores extremos.

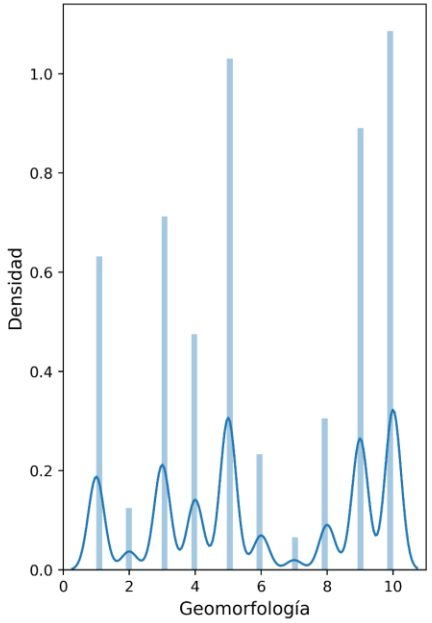
También permite entender datos atípicos que se pueden tener en la base de datos.

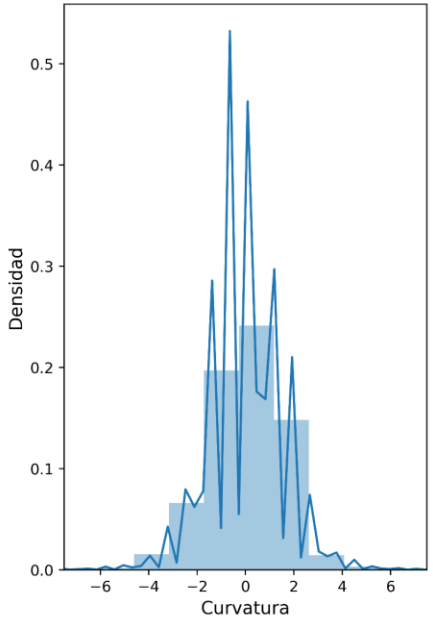


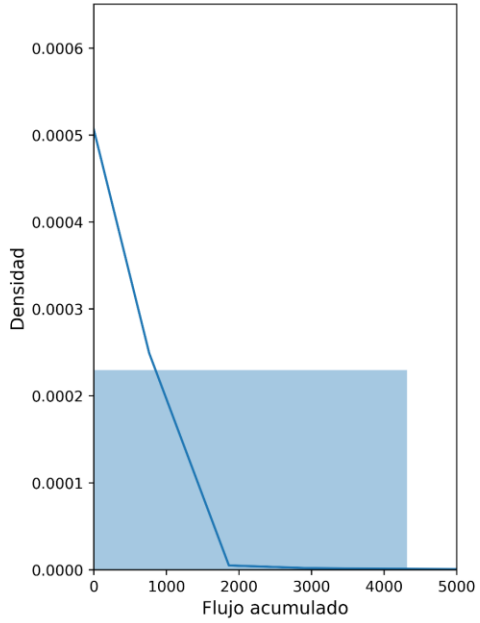








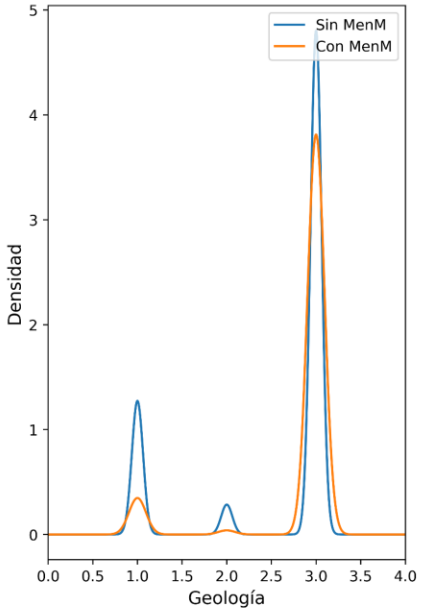


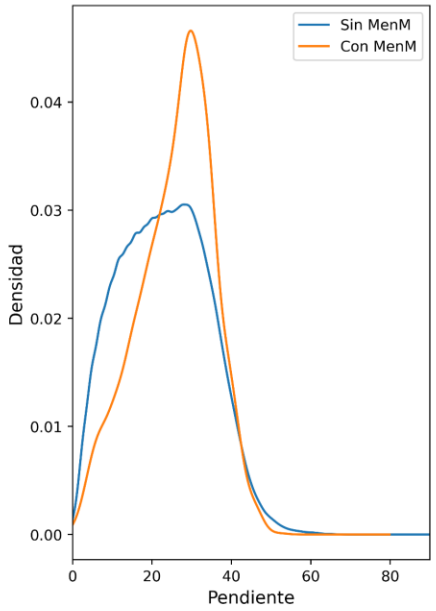


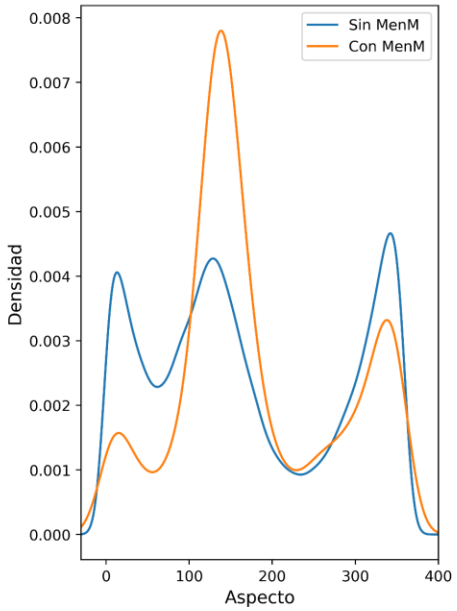
Histogramas de densidad

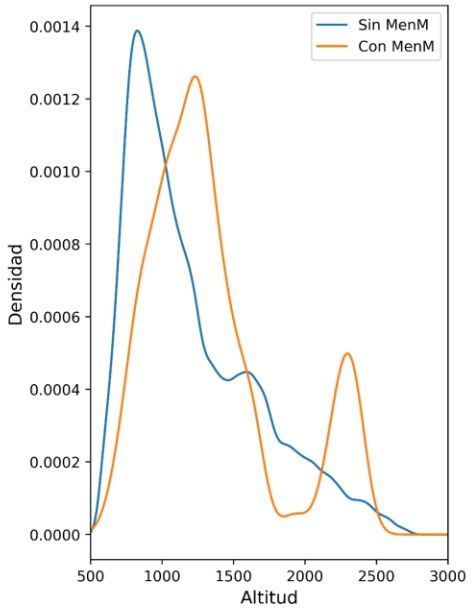
Los histogramas de densidad permiten entender cómo se distribuyen los datos a lo largo de tos los valores posibles para una variable, lo que será importante para lograr dar peso a las clases posteriormente.

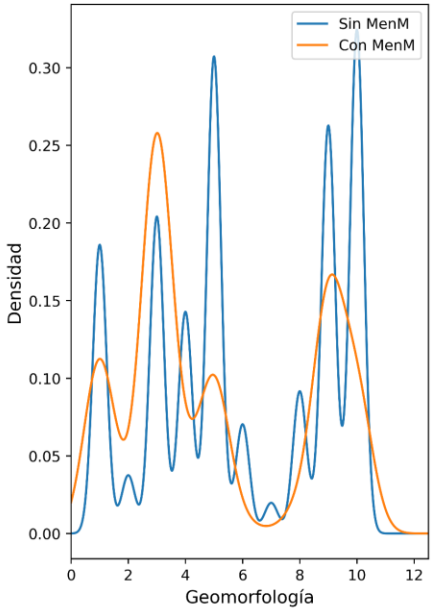
Histogramas de densidad en función de Mov. en M.

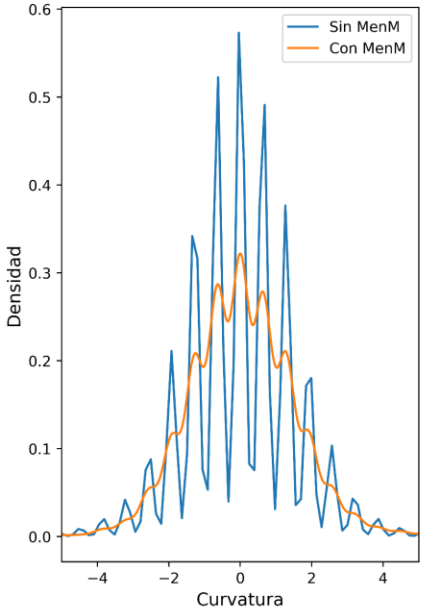


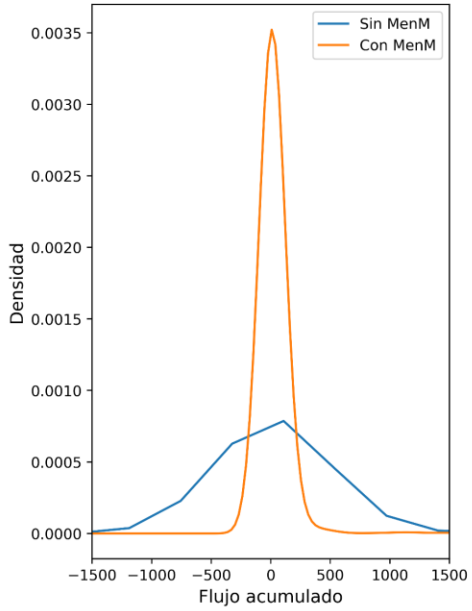












Los histogramas de densidad acá presentes funcionan para ver en que rangos se acumulan con mayor facilidad los movimientos en masa.

Los puntos donde se interceptan son clave para generar luego posibles clasificaciones.

Correlación con los Movimientos en Masa

Se realiza una correlación a partir de la media de cada variable independiente de acuerdo al inventario de movimientos en masa.

inventario	geologia	geomorfologia	pendiente	curvatura	aspecto	flujo acumulado	altitud
0.0	2.554721	6.002436	22.956055	-0.001281	170.708145	389.447327	1203.725342
1.0	2.825000	5.107803	25.942549	0.042358	179.660507	63.878902	1340.093872

Para todas las variables, es importante entender que, si la diferencia entre la media de si hay movimientos en masa o no, es inferior al 5%, entonces, la variable será estadísticamente significativa. Se puede comprobar fácilmente con el valor p que se muestra a continuación.

geologia:
ValorP: 5.037887590530396e-170

geomorfologia:
ValorP: 4.724478297572678e-124

pendiente:
ValorP: 2.4494556921841466e-109

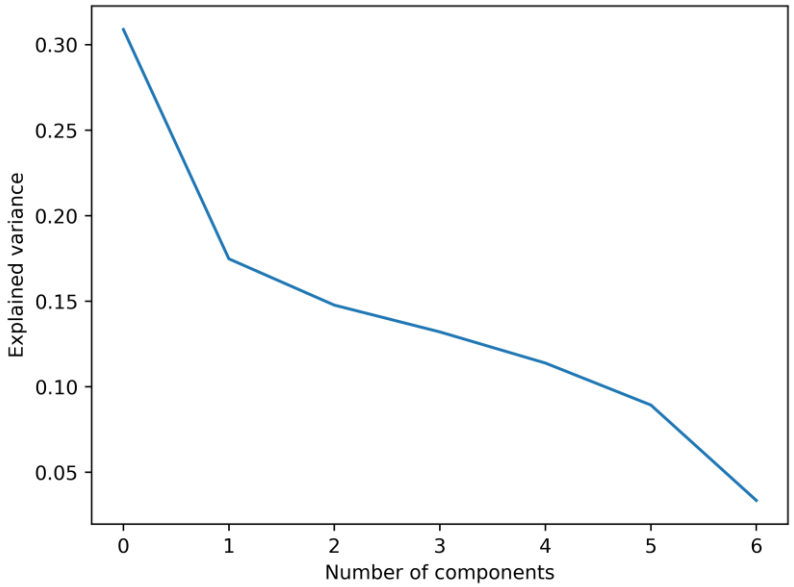
curvatura:
ValorP: 0.023082594980421 79

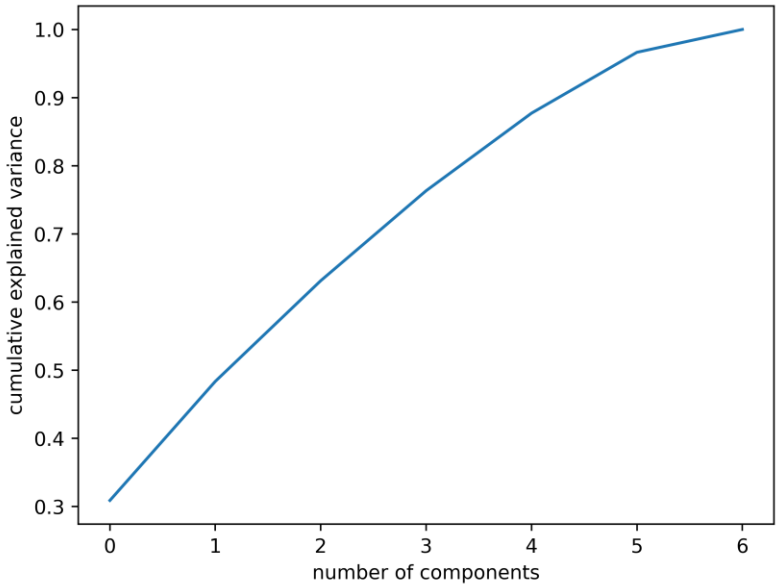
aspecto:
ValorP: 1.4970915178654643e-10

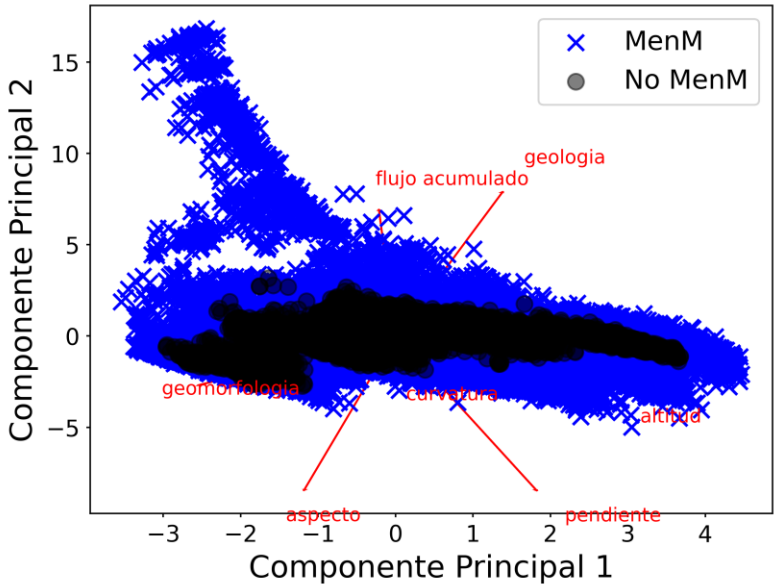
flujo acumulado:
ValorP: 1.3726996368939298e-05

altitud:
ValorP: 2.375611216655719e-136

PCA: Análisis de Componentes Principales







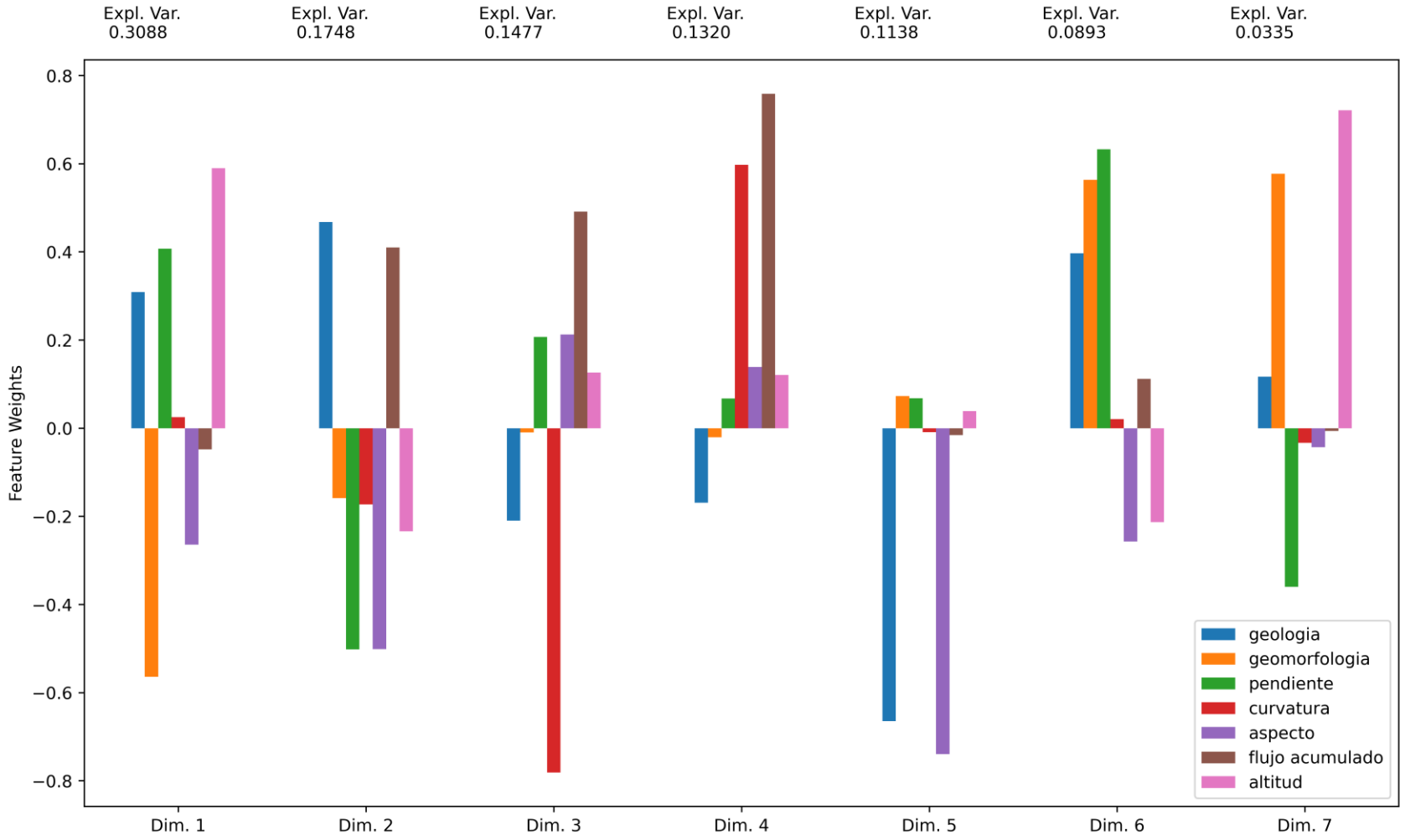
El PCA, es un método no supervisado muy utilizado de machine learning, que permite disminuir la cantidad de variables en un modelo en el que se planean implementar una buena cantidad. Pues, permite conocer que cantidad de variables le aportan más varianza al modelo a realizar.

En la gráfica de la izquierda, es notorio que la varianza baja a medida que se incrementan las variables, pero lo importante es entender cuántas variables son necesarias para tener la mínima varianza posible.

La grafica de la mitad, permite comprender el acumulado de la varianza en función de las variables utilizadas.

Por últmo, la gráfica de la izquierda permite entender cómo se relacionan 2 componentes principales, en este caso, el 1 para el eje X y el 2para el Y. Allí se muestran, cuales variables se pueden correlacionar mejor con los componentes 1 y 2.

PCA: Análisis de Componentes Principales



Este gráfico, permite resumir, cuales variables inciden más, de acuerdo con el número de variables seleccionadas.