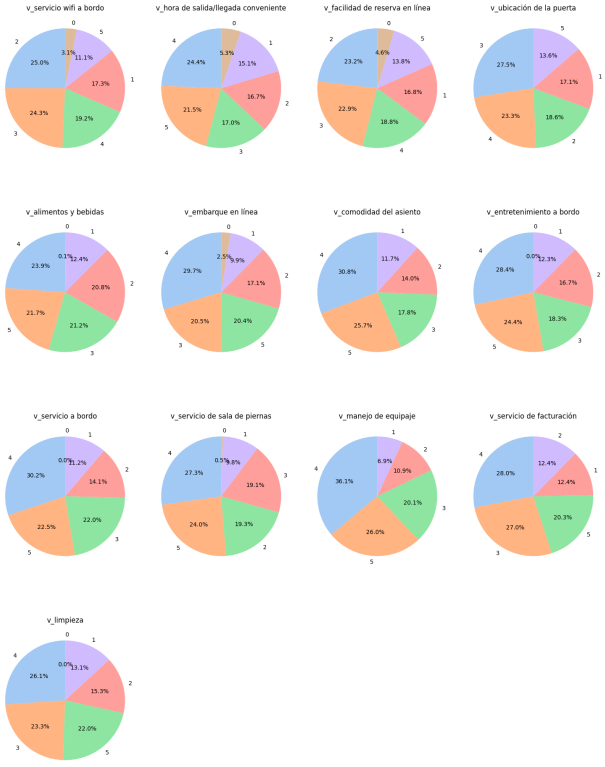
Práctica 1: Reducción de dimensionalidad

José María Reyes Figueroa
Ciencia de Datos - Módulo 3

Ejercicio 1

El primer dataset contiene información de puntuaciones que clientes de una aerolínea brindaron a ésta respecto a 14 servicios que ofrece. El rango de las puntuaciones es de 0-5, donde 0 indica que no se obtuvo respuesta del cliente en ese servicio y 1-5 indica la puntuación que este le dio siendo 5 la más alta.

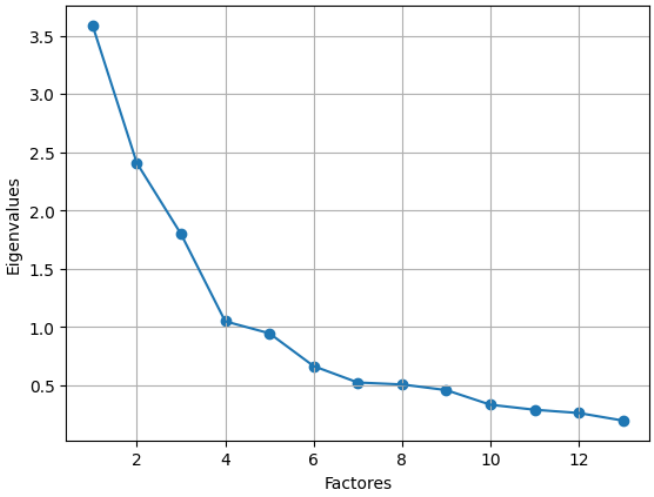
Las variables no contenían missings y tenían la distribución de respuestas que se muestra en el gráfico. Donde al menos de manera visual, cada resultado de la encuesta tiene un valor similar por respuesta salvo en las preguntas donde la opción cero ocupa un porcentaje pequeño, lo que podría indicar una agrupación de los clientes



Analizando los gráficos de caja, se identificaron outliers con el método de cercas percentílicas para los perceptibles 20 y 80 y con un factor de 1.5 dadas las distribuciones de las variables donde no se encontraron valores atípicos en las distribuciones.

Se escalaron los datos y se hizo la prueba de esfericidad, donde se obtuvo que un valor alto de χ^2 de 133457.6 y $p - value$ de 0, por lo que se puede concluir que existe correlación entre las variables.

Luego en la prueba de KMO, se obtiene una proporción de varianza total de 0.7605 y para la determinación del número de factores, se escogerán solo aquellas variables cuyo eigen value sea mayor a 1 por la elección del escalamiento y como se muestra en la gráfica, la elección serían 4 factores.



Con esta cantidad de factores y sin rotación, se obtienen los siguientes valores de las cargas.

	0	1	2	3
v_servicio wifi a bordo	0.440007	0.641630	-0.040986	-0.104044
v_hora de salida/llegada conveniente	0.149339	0.516230	-0.013227	0.247136
v_facilidad de reserva en línea	0.328432	0.815753	-0.092654	-0.050616
v_ubicación de la puerta	0.121029	0.556741	-0.113300	0.401024
v_alimentos y bebidas	0.633539	-0.273371	-0.324174	0.142716
v_embarque en línea	0.563311	0.239859	-0.114392	-0.556693
v_comodidad del asiento	0.701890	-0.255424	-0.261339	-0.031412
v_entretenimiento a bordo	0.835296	-0.276392	0.090525	0.185102
v_servicio a bordo	0.393598	-0.029262	0.613621	0.009114
v_servicio de sala de piernas	0.307115	0.032515	0.416700	-0.019603
v_manejo de equipaje	0.362644	-0.011856	0.644376	0.077693
v_servicio de facturación	0.244928	-0.034270	0.202404	-0.116240
v_limpieza	0.746813	-0.311010	-0.282487	0.094131

Con estos valores el grupo 0 se puede clasificar como los clientes que tienen más interés en los servicios a bordo del avión, como el entretenimiento, la comodidad de sus asientos, la limpieza, los alimentos y bebidas que este pueda ofrecer embarque y servicio de wifi en el avión. Por otra parte el grupo 1 parece estar más interesado en las facilidades del servicio ya que las variables de facilidad de reserva en línea, wifi, ubicación de sus puertas y salidas y llegadas en horas convenientes tienen mayor peso para este grupo. El grupo 2 parece darle importancia a que exista variedad en los servicios así como eficacia en las responsabilidades

de la aerolínea como lo es el manejo del equipaje, que este sea eficaz y cuidado y finalmente, el grupo 4 parece tener preferencia por AAAAA.

Ejercicio 2

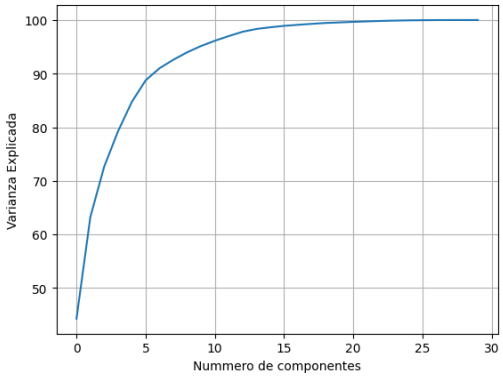
En el segundo dataset, se tiene información de 569 características calculadas a partir de una imagen digitalizada de una aspiración con aguja fina de una masa mamaria.

Los datos no presentaron missings y analizando sus distribuciones así como sus boxplots (imágenes en el código) y haciendo el conteo de outliers (también en el código), se determino que los outliers existentes pertenecen a la distribución natural de los datos, por lo que se conservan ya que como se muestra en el gráfico, hay mayor presencia de outliers en los tumores que resultan benignos (esta etiqueta solo se consideró para la visualización y no en el entrenamiento de los modelos) que los malignos, por lo que se puede concluir que conservarlos es de ayuda para la separación de individuos.

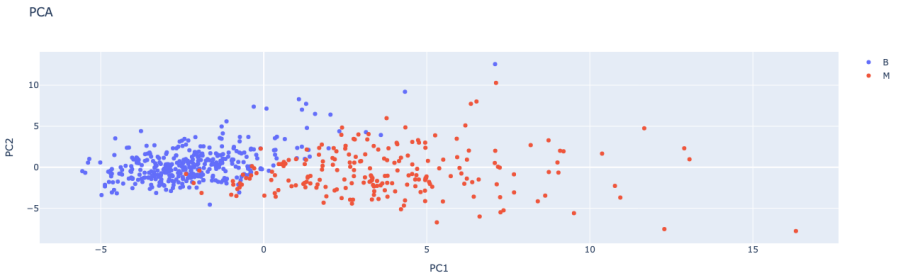
Posterior al escalamiento de datos, se intentarán los métodos de PCA, MDS, ISOMAP y t-SNE con 2 y 3 componentes para determinar cuál tiene mejor desempeño de ellos.

PCA

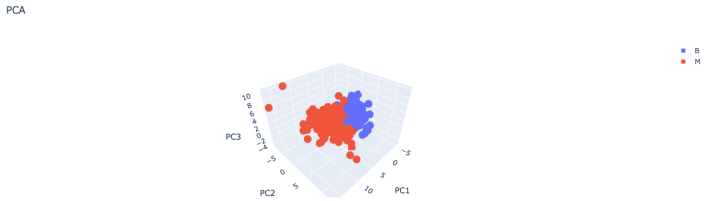
Para el método de PCA, se graficó primero la varianza explicada dado el numéro de componentes, para dos y tres componentes se tiene. Aproximadamente un valor entre 60% y 75% de varianza explicada.



Con PCA para 2 y 3 componentes se obtuvieron los siguientes gráficos.



PCA 2 componentes

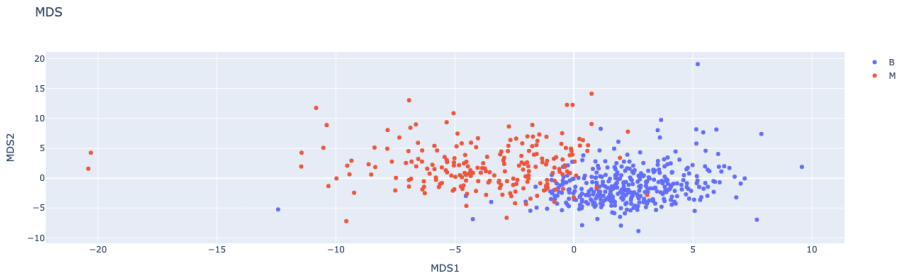


PCA 3 componentes

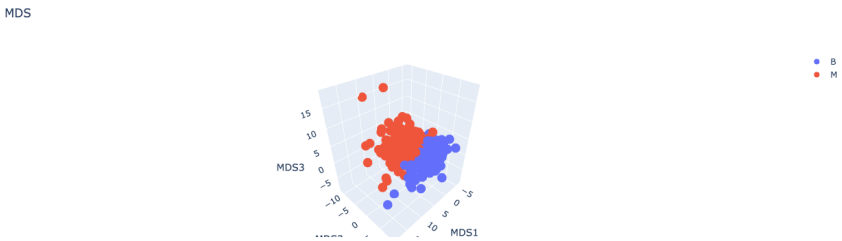
Las varianzas explicadas de ambos métodos son respectivamente 0.63 y 0.73 como se había mencionado por el primer gráfico.

MDS e ISOMAP (método ganador a 2 componentes).

Los métodos de MDS con 2 y 3 componentes tuvieron un stress de 325908.14 y 113073.11 respectivamente y mejoraron visualmente respecto a PCA como se muestra a continuación.

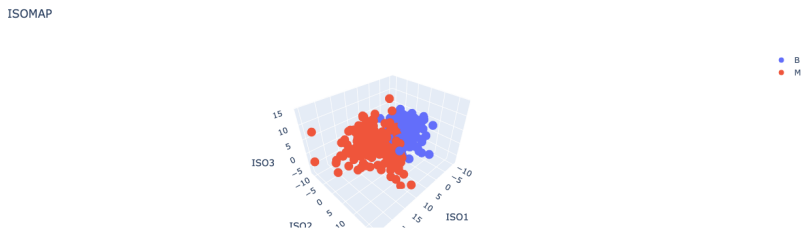


MDS 2 componentes

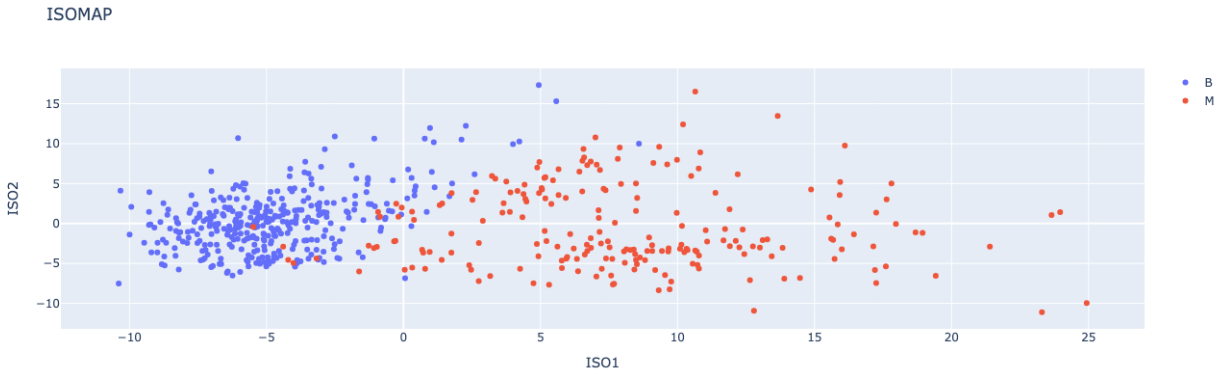


MDS 3 componentes

Los algoritmo de ISOMAP lograron replicar la varianza el 100% en el caso de dos componentes y 98% en el caso de 3 por lo que el primero es el método ganador.

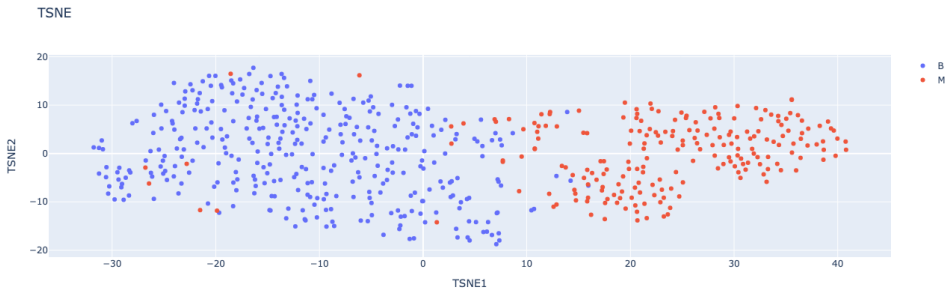


ISOMAP 3 componentes

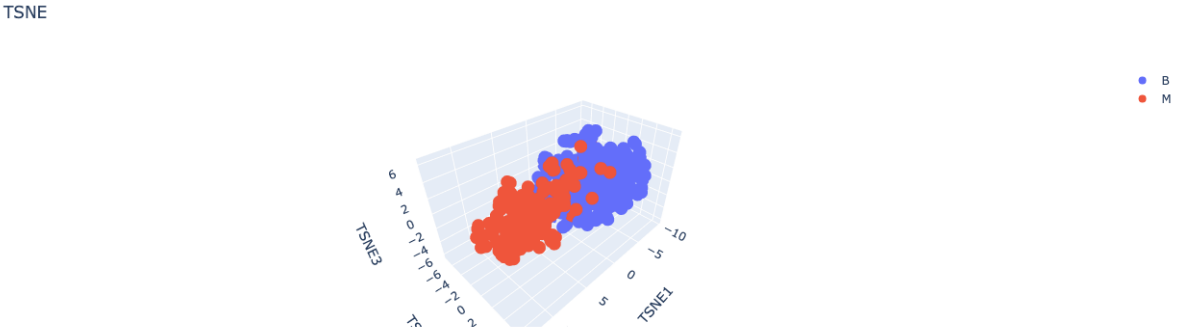


ISOMAP 2 componentes (Método **GANADOR**)

Finalmente se realizó t-SNE como visualmente no tiene la misma distribución del método ganador ni de ISOMAP con 3 componentes se concluye que tampoco tuve un desempeño alto, similar a PCA.



t-SNE 2 componentes



t-SNE 3 componentes