Práctica 1: Reducción de Dimensiones

Alumno: Daniel Martínez Zapata
Profesora: Carla Reséndiz
Módulo III – Diplomado en Ciencia de Datos

1 Ejercicio I - Análisis Factorial

1.1 Introducción

El conjunto de datos contiene 15 categorías con 25976 registros que contienen una serie de preguntas realizadas a los usuarios de una aerolínea, esto con el fin de conocer su opinión de los distintos aspectos y comodidades del viaje. Mediante un modelo de Análisis de Factores se busca generar una nueva representación de las variables originales.

1.2 EDA

Mediante una exploración de los datos se observa que no hay registros faltantes, ni valores atípicos ya que las variables toman valores entre 0 y 5.

De acuerdo a la imagen 1, las calificaciones se agrupan de manera general de la siguiente forma:

- Calificaciones altas entre 4-5: Alimentos y bebidas, Servicio a bordo, Limpieza, Hora de salida/llegada conveniente, Embarque en línea., Servicio de sala de piernas, Comodidad del asiento, Manejo de equipaje, Entretenimiento a bordo.
- Calificaciones entre 0-3: Servicio wifi a bordo, Facilidad de reserva en línea, Ubicación de la puerta, Servicio de facturación.

A pesar de que las variables toman valores entre 0 y 5 se estadarizan para implementar el Análisi Factorial.

Prueba de esfericidad de Bartlett

La prueba de esfericidad de Bartlett proporciona un valor del estadístico de la prueba p de 0, por lo que se concluye que existe correlación entre las variables y es posible aplicar Análisis Factorial.

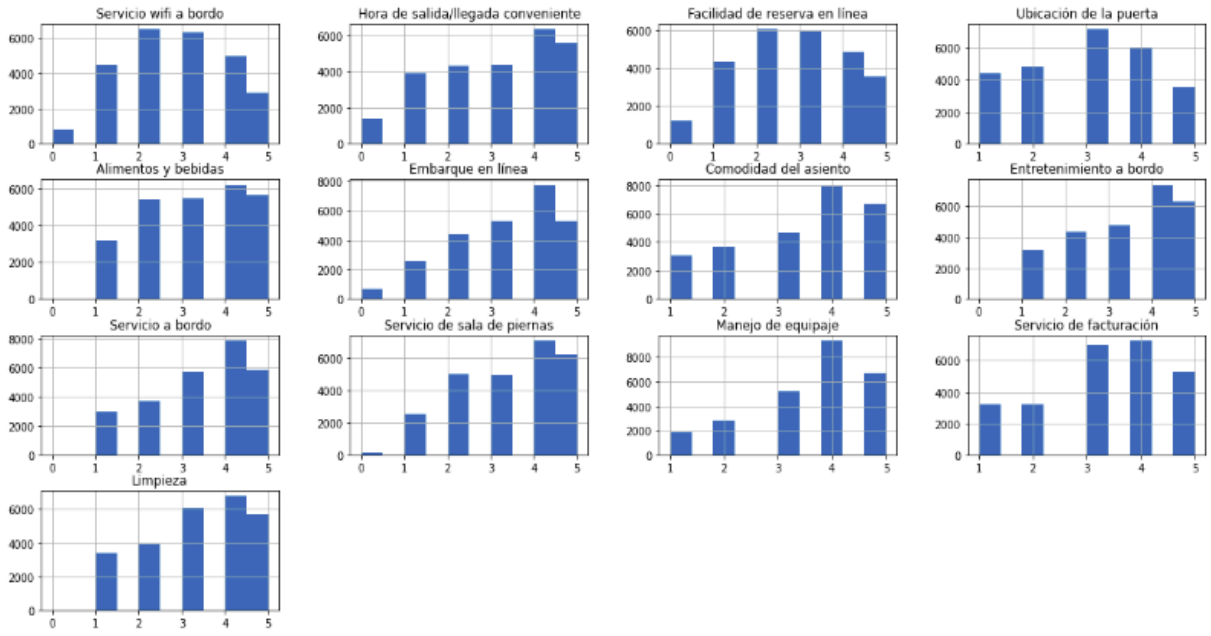

Figure 1: Distribución de las variables que contienen información acerca de las preguntas realizadas a los usuarios de una aerolínea.

Prueba KMO

El valor correspondiente a la prueba de KMO es de 0.76 lo cual indica que existe una alta correlación entre las variables lo cual permite implementar Análisis Factorial.

1.3 Análisis Factorial

Mediante el Análisis Factorial se obtienen 4 eigen valores mayor a 1 que serán los eigen valores utilizados para dicho análisis de factores.

En la imagen 2 se observa la implementación de las cargas de los 4 factores que agrupan la información en dos categorías principales que se denotan como factores internos y externos en los vuelos y una mezcla de ambas condiciones que se enlistan a continuación:

- Factores internos de los vuelos: Alimentos y bebidas, Comodidad del asiento, Entretenimiento a bordo, Limpieza.
- Factores externos de los vuelos: Servicio de wifi a bordo, Hora de salida/llegada conveniente, Facilidad de reserva en línea, Ubicación de la puerta.
- Factores Internos/Externos del Vuelo: Servicio a bordo, Servicio de sala de piernas, Manejo de equipaje.

	0	1	2	3
Servicio wifi a bordo	0.09	0.63	0.14	0.45
Hora de salida/llegada conveniente	-0.02	0.59	0.04	-0.00
Facilidad de reserva en línea	-0.03	0.78	0.04	0.42
Ubicación de la puerta	0.01	0.69	-0.06	-0.12
Alimentos y bebidas	0.77	0.02	0.00	0.04
Embarque en línea	0.27	0.13	0.13	0.77
Comodidad del asiento	0.76	-0.03	0.08	0.22
Entretenimiento a bordo	0.78	0.04	0.46	0.01
Servicio a bordo	0.07	0.01	0.73	0.03
Servicio de sala de piernas	0.06	0.05	0.51	0.07
Manejo de equipaje	0.05	0.05	0.74	-0.04
Servicio de facturación	0.10	-0.04	0.29	0.14
Limpieza	0.85	-0.01	0.09	0.11

Figure 2: Cargas de los factores obtenidos mediante el método de Análisis Factorial.

1.4 Conlusión

El método de de Análisis Factorial permite agrupar los datos que corresponden a ciertas características, en el caso de los vuelos se puede prestar atención a 3 grupos en lugares de 13 categorías individuales.

2 Ejercicio II - Diagnóstico Médico

2.1 Introducción

El conjunto de datos contiene 32 categorías con 569 registros que contienen información obtenida a partir de una imagen digitalizada de una aspiración con aguja fina (FNA) de una masa mamaria. Mediante reducción de dimensiones se busca generar nuevas variables que permitanrepresentar el comportamiento de los datos.

2.2 EDA

Mediante una exploración de los datos se observa que el diagnóstico está etiquetado como benigno (B) y maligno (M). Además se observa que no hay presencia de valores faltantes pero si se observan valores atípicos que son removidos a través de una cerca percentil e intercuartil.

2.3 PCA

En la figura 3 se aprecia que el método de PCA permite pasar de 32 variables a 7 variables que mantienen un valor de varianza explicada de 91% pero a cambio se pierde explicabilidad ya que las nuevas variables son combinación lineal de las variables originales.

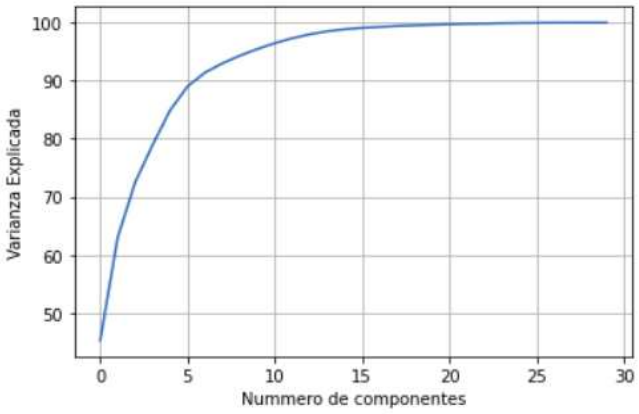

Figure 3: Variable Explicada registrada en función del número de componentes en los datos.

Por otro lado el método de PCA permite la visualización de la información, para ello se consideran solo 2 componentes y se asocia el diagnóstico resultante, lo cual es posible observar en la figura 4.

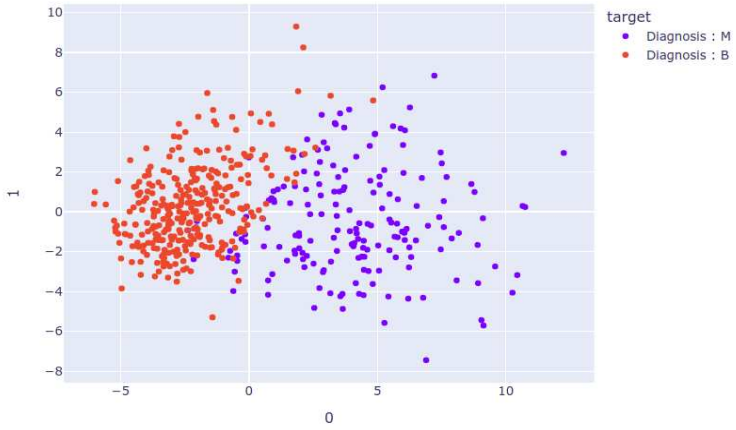

Figure 4: Visualización del Diagnóstico a partir de dos factores.

2.4 MDS

En la figura 5 se aprecia la visualización hecha a partir del método MDS utilizando una métrica de distancia Euclidiana y una métrica de distancia Manhattan, cuyos valores de stress son

303718.3 y 4613850.1, respectivamente. Por lo que en este caso se considera al modelo MDS-distancia Euclidiana el que reporta un valor más bajo de stress y una agrupación más densa de los diagnós ticos

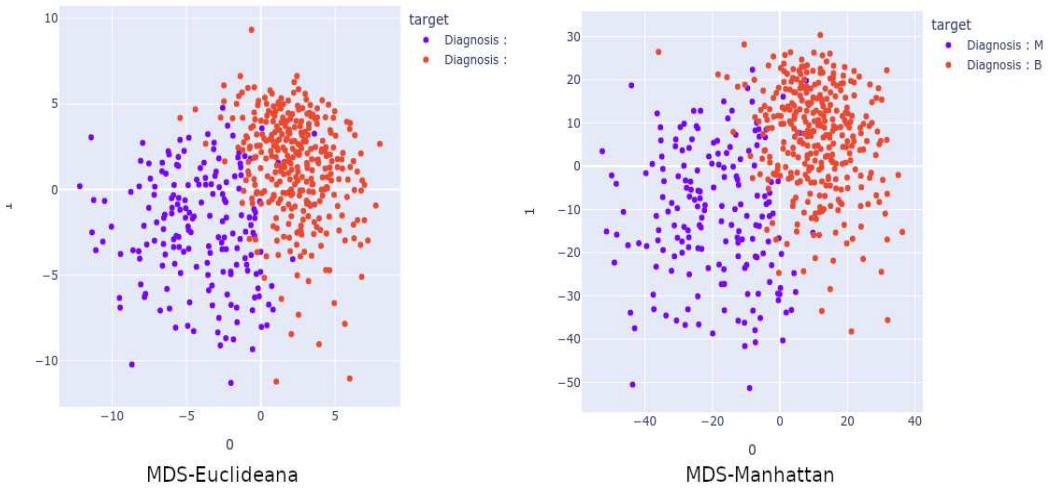

Figure 5: Visualización del Diagnóstico a partir de MDS-Euclidiano y MDS-Manhattan.

2.5 ISOMAP

En la figura 6 se aprecia la visualización hecha a partir del método ISOMAP para dos componentes.

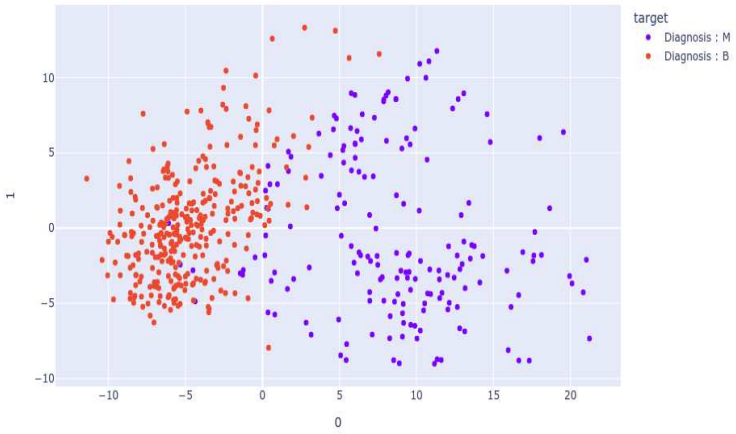

Figure 6: Visualización del Diagnóstico a partir de ISOMAP.

Visualmente se aprecia que el diagnóstico M aparece menos denso y los puntos más dispersos, lo cual tendría que analizarse con mayor detenimiento.

2.6 t-SNE

En la figura 7 se aprecia la visualización hecha a partir del método t-SNE basado en una distribución de probabilidad de tipo t-student para clasificar las distancias entre los puntos.

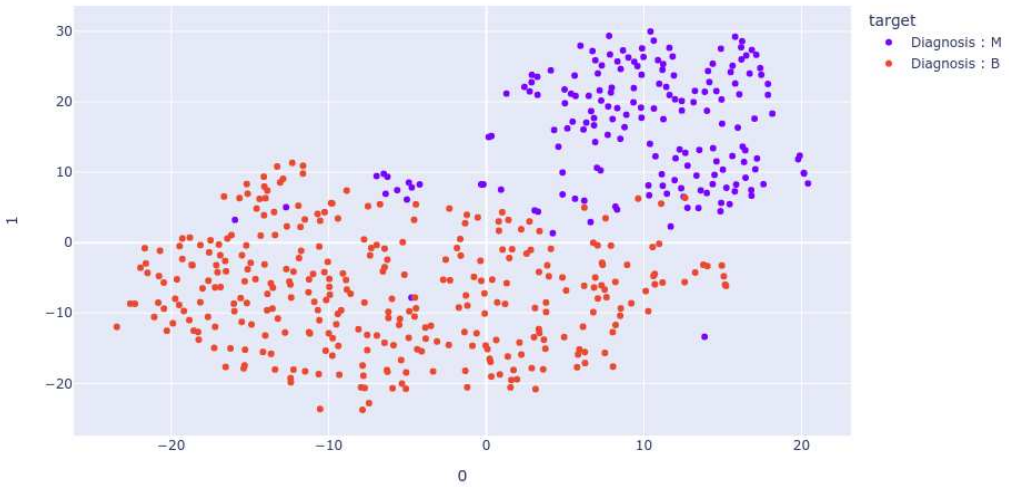

Figure 7: Visualización del Diagnóstico a partir de t-SNE.

Visualmente se observa que t-SNE presenta a los datos con conglomerados definidos y con proporciones adecuadas de acuerdo al número de diagnósticos M y B correspondientes.

2.7 Conlusión

Los métodos de reducción de dimensiones permiten reducir los tiempos de cómputo utilizando menos variables perdiendo explicatividad pero ayudan en la visualización de los datos. En este caso en particu lar se observa que la proporción de diagnósticos B=357 y M= 212 se puede visualizar con los distintos métodos mostrados anteriormente.

El método de PCA nos proporciona información de la cantidad de varianza explicada que puede representarse con un número bastante reducido de variables (7 componentes) y cuya visualización se aprecia con la densidad correspondiente a los diagnósticos reportados sin que contengan valores mezclados en los conglomerados a diferencia de MDS e ISOMAP, mientras que t-SNE también muestra una clasificación relativamente buena de los datos. De esta manera PCA proporciona mayor información acerca del conjunto de datos a través de la varianza explicada y una visualización del comportamiento de los datos siempre y cuando el conjunto de datos se pueda tratar de manera lineal.

3 Ejercicio III - Lenguaje de Manos

3.1 Introducción

El conjunto de datos contiene 32 categorías con 569 registros que contienen 26 imágenes que corresponden al lenguaje de señas con la mano de 28x28. Mediante reducción de dimensiones se busca generar nuevas variables que permitan clasificar las imágenes manteniendo alejadas aquellas imágenes donde los gestos de la mano son muy diferentes y mantienen cerca aquellas imágenes donde los gestos son similares.

3.2 EDA

Mediante una exploración de los datos se observa que no hay registros faltantes, ni valores atípicos ya que las variables corresponen a pixeles tomados de una imágen, tampoco hace falta escalar sin embargo se realiza el escalamiento como parte los pasos a realizar.

3.3 PCA

En la figura 8 se aprecia la clasificación obtenida mediante PCA la cual agrupa a todas las posiciones de la mano en un solo conglomerado sin ser clara su clasificación y también las clasifica de acuerdo a su tonalidad. Con 45 componentes es posible describir el 90% de la varianza explicada.

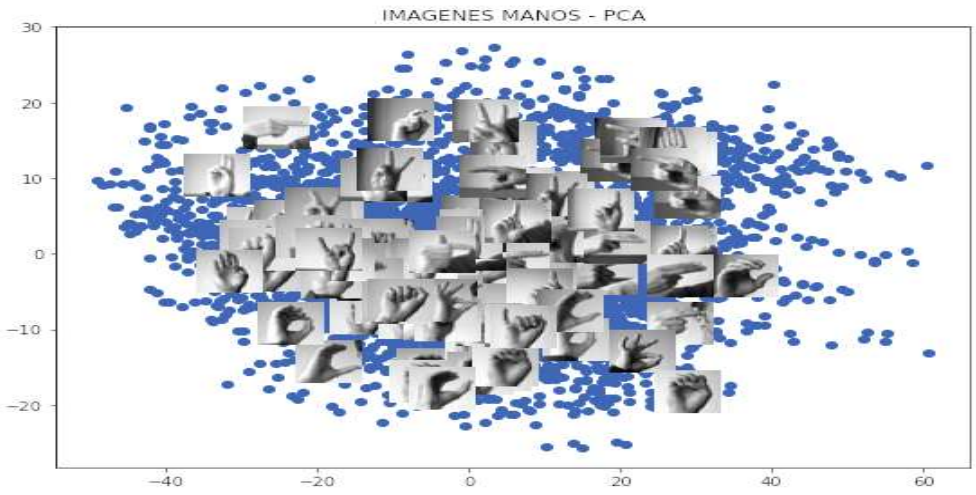

Figure 8: Visualización de los gestos de las manos con 2 factores utilizando PCA.

3.4 MDS

En la figura 9 se aprecia la visualización hecha a partir del método MDS, en el cual como en el caso anterior también se observa que los gestos de las manos se encuentran aglomerados al

centro, si se identifican grupos donde las manos comparten el mismo comportamiento pero no clara la separación.

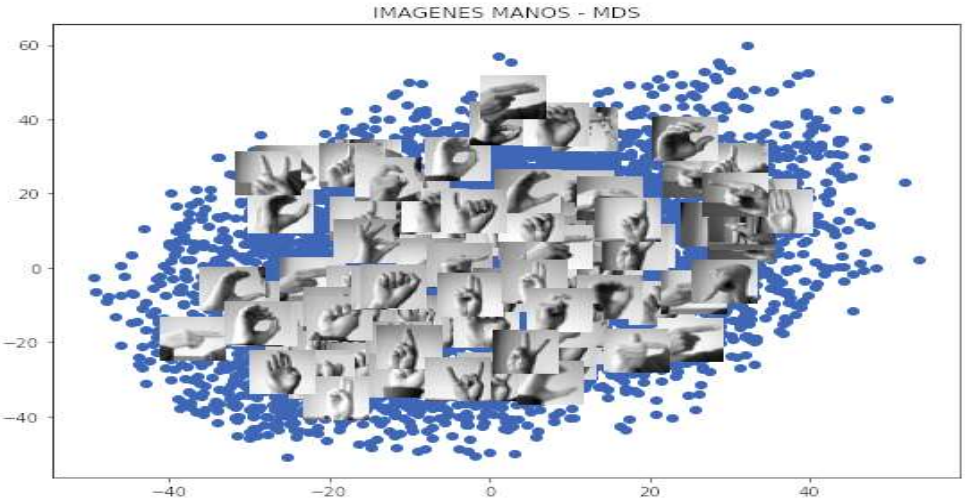

Figure 9: Visualización de los gestos de las manos a partir de MDS.

3.5 ISOMAP

En la figura 10 se aprecia la visualización hecha a partir del método ISOMAP para dos componentes. En dicho modelo se aprecia una clasificación de los gestos de las manos mucho más precisa identificando cada gesto con una dirección particular.

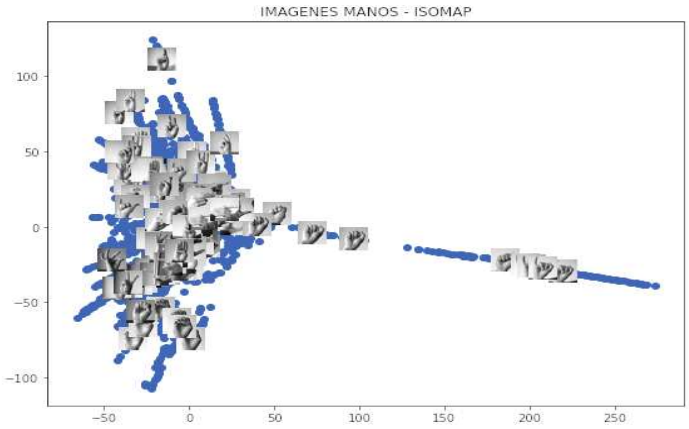

Figure 10: Visualización de los gestos de las manos a partir de ISOMAP.

3.6 Conclusión

El método de ISOMAP permite la reducción de variables pero además identifica el comportamiento específico de cada gesto de las manos por lo que es el adecuado para clasificar este tipo de imágenes.