UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Departamento de Informática
Universidad Técnica Federico Santa María

Análisis Inteligente de Datos-Tarea 1

Prof. Ricardo Ñanculef

Integrantes: Diego Salazar Barrera
Felipe Flores Valdivia

1.-Manipulación Básica de *Dataframes* y Visualización I

El trabajo realizado consiste en aprender a utilizar la librería pandas y matplotlib a través de una serie de ejercicios sobre un *dataset* con información de los pasajeros del Titanic.
Se procede a cargar el *dataset* de 891 entradas y 12 atributos, con algunos elementos nulos para algunas entradas. A partir del análisis de los datos es posible determinar que sobrevivieron aproximádamente un 74% de las pasajeros mujeres y un 19% de los hombres, de lo que se puede suponer una correlación entre el género y la probabilidad de supervivencia (favorable a las mujeres). Por otro lado, es posible notar que la cantidad de niños y adultos mayores fallecidos es considerablemente inferior a la cantidad de adultos (~20-60 años) igualmente fallecidos. Aún así, dado que las edades de todos los pasajeros se tienden a concentrar alrededor de los 30 años, la mayoría de muertos y sobrevivientes se encuentran dentro de dicho rango.

Para realizar los análisis, se tuvo que asignar una edad estimada a los pasajeros a los que se les desconocía la edad. Se decidió aplicar la función piso a la media aritmética de las edades de hombres y mujeres. De acuerdo a la definición de los datos, para estimar edades se debe cumplir con el formato ‘xx.5’, donde xx es la edad estimada. De este modo se asignaron las nuevas edades 30.5 y 27.5 para hombres y mujeres respectivamente.

Los pasajeros estaban divididos en 3 clases. De los fallecidos, un 68%, 18% y 15% de los pasajeros eran de 3º, 2º y 1º clase respectivamente. Observando los datos, es posible notar que el 88% de las mujeres fallecidas corresponden a la 3º clase, por lo que se puede deducir que para ellas no se cumple con tanta frecuencia la norma de que sobreviven por sobre los hombres.

A partir de todo lo anterior, se realiza la predicción de que todas las personas de 1º clase y las mujeres de 2º clase sobreviven. Al contrastar la predicción con los datos reales, se llega a un nivel de predicción de un 64% y un recall de 81%.

Finalmente, se decide crear una nueva partición socioeconómica según costo de pasaje, ya que efectivamente algunos de primera clases pagaron menos que otros. Se eliminan los datos *outlier* y se realiza una nueva partición en 5 clases económicas. Observando un histograma con la nueva partición, se realiza una nueva regla predictiva: los únicos muertos son de la clase 5 (la que se presume más pobre). Esto da una precisión de un 34%, pero un recall del 100%.

Al aplicar ambas reglas sobre los datos de prueba, se descubrió que la regla 1 era más precisa y exhaustiva que la regla 2.

2.-PCA y Visualización II

Luego de obtener las dos componentes principales a través del PCA, y llevar esto a un gráfico de puntos, es posible observar que aquellos países que tienen mayor cantidad de infectados, se encuentran a la derecha del gráfico. Por esto, es posible concluir que la primera componente principal (aquella que se encuentra en el eje x), representa la cantidad promedio de infectados por país. Por lo cual, al usar un código de colores secuencial, aquellos países con mayor incidencia de la enfermedad se encuentran en rojo más intenso.

Al obtener un scatter plot de las componentes principales, utilizando un código de colores divergentes, se diferencian aquellos países donde existe una mayor variación en el número de infectados a lo largo de los años de estudio. Debido a la dispersión existente entre estos datos y al no existir zonas claramente identificadas como en el scatter plot anterior, es difícil determinar a qué corresponde cada dato. Al realizar un gráfico de burbujas, se puede apreciar de mejor manera la diferencia de color existente entre unos puntos y otros, por lo que, a pesar de no cumplir con el principio de la tinta de Tufte, se presentan los datos de manera más íntegra y permiten una mejor visualización de estos.

Al agregar los nombres de los países a los puntos del gráfico, nuevamente se rompe la regla de la tinta de Tufte, e incluso, si esto no se acota, se generaría un desorden que no permitiría observar la información de manera correcta. Ahora bien, la visualización se realiza de manera correcta cuando se logra transmitir lo que el analista quiere transmitir, por lo que si la intención es mostrar que aquellos países que poseen una mayor infección a lo largo de los años, sería una buena herramienta, que permitirá al observador apreciar de manera rápida, cuáles son los países más afectados por esta enfermedad.

Al realizar las mismas operaciones sobre el dataframe de tuberculosis, se observan resultados similares, donde son más los países con bajo índice de enfermedad. Además, existe una mayor variación en los índices de ambas enfermedades en aquellos países donde estas se encuentran más arraigadas. Lo que es esperable, pues en aquellos países donde se encuentran mayores casos de personas enfermas, es donde más se debe atacar la enfermedad. Al comparar los gráficos de burbuja obtenidos y considerando que los datos fueron normalizados, es posible concluir que la tuberculosis es una enfermedad con mayor número de infectados que el VIH, debido a que las burbujas son mucho mayores en el caso de esta enfermedad.

Con respecto a los gráficos de la varianzas explicadas, es posible mencionar que en ambos casos las dos primeras componentes principales representan más del 95%, siendo la primera la que acapara cerca del 90% de información en ambos casos, por lo cual la representación bidimensional conserva gran cantidad de la información del dataset original, permitiendo hacer un buen análisis de los datos.