Laboratorio 1.
Análisis Exploratorio, PCA y Apriori

INSTRUCCIONES:

Utilice el dataset Cervical Cancer (Risks Factors) que comparte UC Irvine Machine Learning Repository. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es clasificar si una paciente será diagnosticada con cáncer cervical o no. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual. Lleve a cabo un análisis de componentes principales. Este laboratorio debe realizarse en **PAREJAS.** Para que se pueda calificar su laboratorio debe estar inscrito en algún grupo de canvas.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

El conjunto de datos fue recolectado en el "Hospital Universitario de Caracas" en Venezuela. Comprende información demográfica, hábitos y registros históricos de 858 pacientes.

Variables

Consta de 36 variables y 858 filas:

- **Age:** Edad de la paciente
- **Number.of.sexual.partners:** Cantidad de parejas sexuales que ha tenido la paciente desde que inició su vida sexual
- **First.sexual.intercourse:** Edad a la que tuvo el primer encuentro sexual.
- **Num.of.pregnancies:** Cantidad de embarazos
- **Smokes:** Si fuma o no
- **Smokes.years:** Años que lleva fumando
- **Smokes.packs.per.year:** Cajetillas de cigarrillos por año que fuma la paciente
- **Hormonal.Contraceptives:** Si usa anticonceptivos hormonales o no
- **Hormonal.Contraceptives.years:** Años que lleva usando anticonceptivos hormonales
- **IUD:** Si tiene colocado algún dispositivo intrauterino (DIU)
- **IUD.years:** Años que lleva usando un DIU
- **STDs:** Si ha tiene enfermedades de transmisión sexual (ETS)
- **STDs.number:** Cuantas ETS ha tenido
- **STDs.condylomatosis:** si ha tenido condilomatosis
- **STDs.cervical.condylomatosis:** si ha tenido condilomatosis cervical
- **STDs.vaginal.condylomatosis:** Si ha tenido condilomatosis vaginal
- **STDs.vulvo.perineal.condylomatosis:** Si ha tenido condilomatosis vulvo perineal
- **STDs.syphilis:** Si ha tenido Sífilis
- **STDs.pelvic.inflammatory.disease**: Si ha tenido inflamaciones pélvicas
- **STDs.genital.herpes:** si ha tenido herpes genital

- **STDs.molluscum.contagiosum:** Si ha tenido molusco contagioso
- **STDs.AIDS:** Si tiene SIDA
- **STDs.HIV:** Si tiene VIH
- **STDs.Hepatitis.B:** si ha tenido o tiene hepatitis B
- **STDs.HPV:** Si ha tenido o tiene Virus del Papiloma Humano (VPH)
- **STDs.Number.of.diagnosis:** Cantidad de diagnósticos de ETS
- **STDs.Time.since.first.diagnosis:** Tiempo desde el primer diagnóstico
- **STDs.Time.since.last.diagnosis**: Tiempo desde el último diagnóstico
- **Diagnósticos**:
  - **Dx.Cancer:** Si tiene diagnóstico de cáncer o no
  - **Dx.CIN**: Si tiene diagnóstico de NIC (Neoplasia Intraepitelial Cervical)
  - **Dx.HPV**: Si tiene diagnóstico de Virus del Papiloma Humano
  - **Dx:** Si tiene diagnóstico
- **Pruebas para diagnosticar**
  - **Hinselmann:** Si hicieron Colposcopía
  - **Schiller:** Si hicieron la prueba de Schiller
  - **Citology:** Si hicieron citología o no.
  - **Biopsy:** Si hicieron Biopsia o no

Resumen del conjunto de datos

```
summary(riesgoCancerCervical)
```

```
      Age        Number.of.sexual.partners First.sexual.intercourse
 Min.   :13.00   Min.   : 1.000            Min.   :10
 1st Qu.:20.00   1st Qu.: 2.000            1st Qu.:15
 Median :25.00   Median : 2.000            Median :17
 Mean   :26.82   Mean   : 2.528            Mean   :17
 3rd Qu.:32.00   3rd Qu.: 3.000            3rd Qu.:18
 Max.   :84.00   Max.   :28.000            Max.   :32
                 NA's   :26                NA's   :7
 Num.of.pregnancies Smokes     Smokes.years   Smokes.packs.per.year
 Min.   : 0.000     ?  : 13   Min.   : 0.00   Min.   : 0.0000
 1st Qu.: 1.000     0.0:722   1st Qu.: 0.00   1st Qu.: 0.0000
 Median : 2.000     1.0:123   Median : 0.00   Median : 0.0000
 Mean   : 2.276               Mean   : 1.22   Mean   : 0.4531
 3rd Qu.: 3.000               3rd Qu.: 0.00   3rd Qu.: 0.0000
 Max.   :11.000               Max.   :37.00   Max.   :37.0000
 NA's   :56                   NA's   :13      NA's   :13
 Hormonal.Contraceptives Hormonal.Contraceptives.years  IUD
 ?  :108                 Min.   : 0.000                ?  :117
 0.0:269                 1st Qu.: 0.000                0.0:658
 1.0:481                 Median : 0.500                1.0: 83
                         Mean   : 2.256
                         3rd Qu.: 3.000
                         Max.   :30.000
                         NA's   :108
```

```
  IUD.years        STDs      STDs.number     STDs.condylomatosis
Min.   : 0.0000   ?  :105   Min.   :0.0000   ?  :105
1st Qu.: 0.0000   0.0:674   1st Qu.:0.0000   0.0:709
Median : 0.0000   1.0: 79   Median :0.0000   1.0: 44
Mean   : 0.5148             Mean   :0.1766
3rd Qu.: 0.0000             3rd Qu.:0.0000
Max.   :19.0000             Max.   :4.0000
NA's   :117                 NA's   :105
STDs.cervical.condylomatosis STDs.vaginal.condylomatosis
?  :105                      ?  :105
0.0:753                      0.0:749
                             1.0:  4

STDs.vulvo.perineal.condylomatosis STDs.syphilis
?  :105                            ?  :105
0.0:710                            0.0:735
1.0: 43                            1.0: 18

STDs.pelvic.inflammatory.disease STDs.genital.herpes
?  :105                          ?  :105
0.0:752                          0.0:752
1.0:  1                          1.0:  1

STDs.molluscum.contagiosum STDs.AIDS STDs.HIV  STDs.Hepatitis.B STDs.HPV
?  :105                    ?  :105   ?  :105   ?  :105          ?  :105
0.0:752                    0.0:753   0.0:735   0.0:752          0.0:751
1.0:  1                              1.0: 18   1.0:  1          1.0:  2

STDs.Number.of.diagnosis STDs.Time.since.first.diagnosis
Min.   :0.00000          Min.   : 1.000
1st Qu.:0.00000          1st Qu.: 2.000
Median :0.00000          Median : 4.000
Mean   :0.08741          Mean   : 6.141
3rd Qu.:0.00000          3rd Qu.: 8.000
Max.   :3.00000          Max.   :22.000
                         NA's   :787


STDs.Time.since.last.diagnosis Dx.Cancer Dx.CIN  Dx.HPV  Dx      Hinselmann
Min.   : 1.000                 0:840     0:849   0:840   0:834   0:823
1st Qu.: 2.000                 1: 18     1:  9   1: 18   1: 24   1: 35
Median : 3.000
Mean   : 5.817
3rd Qu.: 7.500
Max.   :22.000
NA's   :787
```

```
Schiller Citology Biopsy
0:784    0:814    0:803
1: 74    1: 44    1: 55
```

EJERCICIOS

1. Haga una exploración rápida de sus datos.
2. Diga el tipo de cada una de las variables del dataset (cualitativa o categórica, cuantitativa continua, cuantitativa discreta)
3. Incluya los gráficos exploratorios siendo consecuentes con el tipo de variable que están representando.
4. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.
5. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos.
6. Determine el comportamiento a seguir con los valores faltantes. Explique si necesita remover alguna variable por la cantidad de valores faltantes que tiene. ¿Es factible eliminar todos los valores faltantes de todas las variables?
7. Estudie si es posible hacer transformaciones en las variables categóricas para incluirlas en el PCA, ¿valdrá la pena?
8. Estudie si es conveniente hacer un Análisis de Componentes Principales. Recuerde que puede usar el índice KMO y el test de esfericidad de Bartlett. Haga un análisis de componentes principales con las variables numéricas, discuta los resultados e interprete los componentes.
9. Obtenga reglas de asociación interesantes del dataset. Recuerde discretizar las variables numéricas. Genere reglas con diferentes niveles de confianza y soporte. Discuta los resultados. Si considera que debe eliminar variables porque son muy frecuentes y con eso puede recibir más insigths de la generación de reglas. Hágalo y discútalo.

EVALUACIÓN

NOTA: La evaluación de cada integrante del grupo será de acuerdo con sus contribuciones al trabajo grupal

(42 puntos) Análisis Exploratorio:

- Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
- Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos
- Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.
- Estudia las variables categóricas
- Elabora gráficos de barra, tablas de frecuencia y de proporciones
- Elabora gráficos adecuados según el tipo de dato que representan
- Explica muy bien todos los procedimientos y los hallazgos que va haciendo.
- Decide que hacer con los valores faltantes y con las variables a utilizar en los algoritmos.

(18 puntos) Análisis de componentes Principales

- Estudia la matriz de correlación, la agrega y explica lo que observa en ella
- Determina si es posible usar la técnica de análisis factorial para hallar las componentes principales
- Determina si vale la pena aplicar las componentes principales interpretando la prueba de esfericidad de Bartlett
- Obtiene los componentes principales y explica cuántos seleccionará para explicar la mayor variabilidad posible.
- Interpreta los coeficientes principales.

(18 puntos) Reglas de asociación

- Construye reglas de asociación usando el algoritmo a priori.
- Prueba con varios valores de confianza y soporte, y decide si quitar o no características para obtener mejores hallazgos.
- Discute sobre las reglas de asociación más interesantes teniendo en cuenta sus niveles de confianza y soporte.

(22 puntos) Hallazgos y conclusiones.

- Hace un resumen de los hallazgos en el análisis exploratorio
- Llega a conclusiones sobre el análisis de componentes principales
- Determina las reglas de asociación más interesantes.

MATERIAL A ENTREGAR

- Archivo .pdf con el informe de análisis exploratorio.
- Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado
- Link de github o el versionador que se utilizó.