CC3084, Data Science
Sección 20
Sara Echeverría, 21371
Melissa Pérez, 21385

Laboratorio No. 1
Análisis Exploratorio, PCA y Apriori

Repositorio: https://github.com/bl33h/cervicalCancerRiskFactors

1. *Realice una exploración rápida de sus datos. Puede usar alguna forma automatizada de hacer análisis exploratorio siempre y cuando explique los resultados que arrojan los módulos/paquetes.*
    - En el análisis exploratorio del conjunto de datos sobre factores de riesgo del cáncer cervical, se llevaron a cabo múltiples procesos para entender mejor las características de los datos y su potencial impacto en futuros estudios. Primero, se realizó la limpieza de datos, reemplazando valores faltantes y corrigiendo tipos de datos para asegurar la precisión en los análisis posteriores. A lo largo del documento, se estará detallando cada uno de los hallazgos entre las variables, tablas de frecuencia y proporción, entre otros.

2. *Indique el tipo de cada una de las variables del conjunto de datos (cualitativa o categórica, cuantitativa continua, cuantitativa discreta)*
    a. Categóricas
        i. Smokes
        ii. Hormonal Contraceptives
        iii. IUD
        iv. STDs
        v. STDs.condylomatosis
        vi. STDs.cervical.condylomatosis
        vii. STDs.vaginal.condylomatosis
        viii. STDs.vulvo.perineal.condylomatosis
        ix. STDs.syphilis
        x. STDs.pelvic.inflammatory.disease
        xi. STDs.genital.herpes
        xii. STDs.molluscum.contagiosum

xiii. STDs.AIDS
xiv. STDs.HIV
xv. STDs.Hepatitis.B
xvi. STDs.HPV
xvii. Dx.Cancer
xviii. Dx.CIN
xix. Dx.HPV
xx. Dx
xxi. Hinselmann
xxii. Schiller
xxiii. Citology
xxiv. Biopsy

b. Cuantitativa Continuas
    i. Inicialmente en el conjunto de datos no existen variables que establezcan rangos para su clasificación. Pero en un futuro, puede establecerse la edad como rangos de edades para clasificar.

c. Cuantitativa Discretas
    i. Age
    ii. Number of sexual partners
    iii. First.sexual.intercourse
    iv. Num.of.pregnancies
    v. Smokes.years
    vi. Smokes.packs.per.year
    vii. Hormonal.Contraceptives.years
    viii. IUD.years
    ix. STDs.number
    x. STDs.Number.of.diagnosis
    xi. STDs.Time.since.first.diagnosis
    xii. STDs.Time.since.last.diagnosis

3. *Incluya los gráficos exploratorios, siendo consecuentes con el tipo de variable que están representando.*

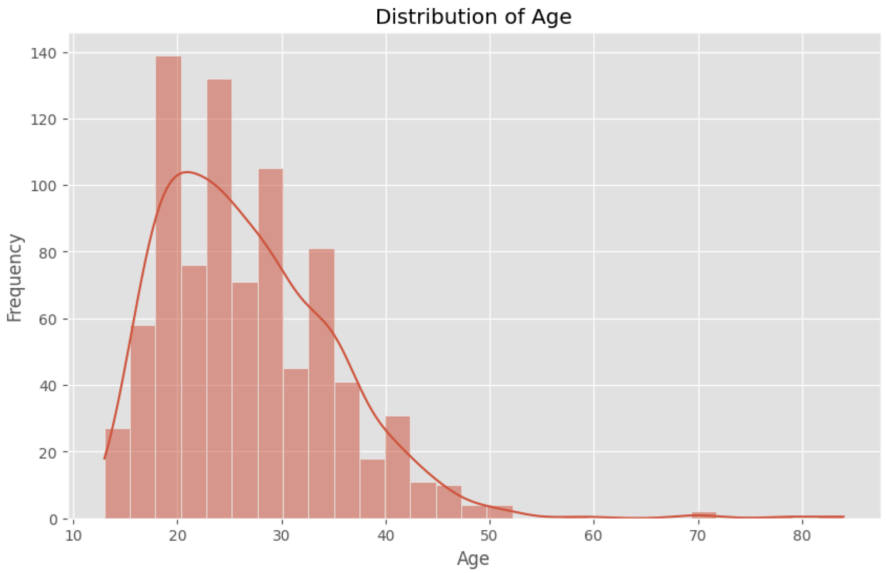

Gráfico 01. Distribución de edades dentro del conjunto de datos.

- La mayoría de las personas en el conjunto de datos, tienen entre 15 y 25 años. Puede decirse que predominan en el rango de edad reproductiva.

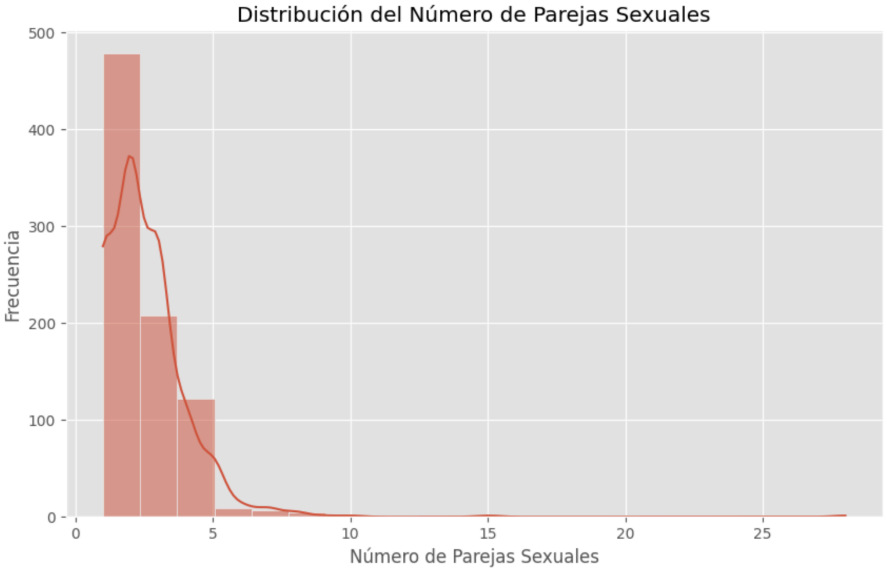

Gráfico 02. Distribución del número de parejas sexuales.

- La mayoría de la población tiene un número relativamente bajo de parejas sexuales.

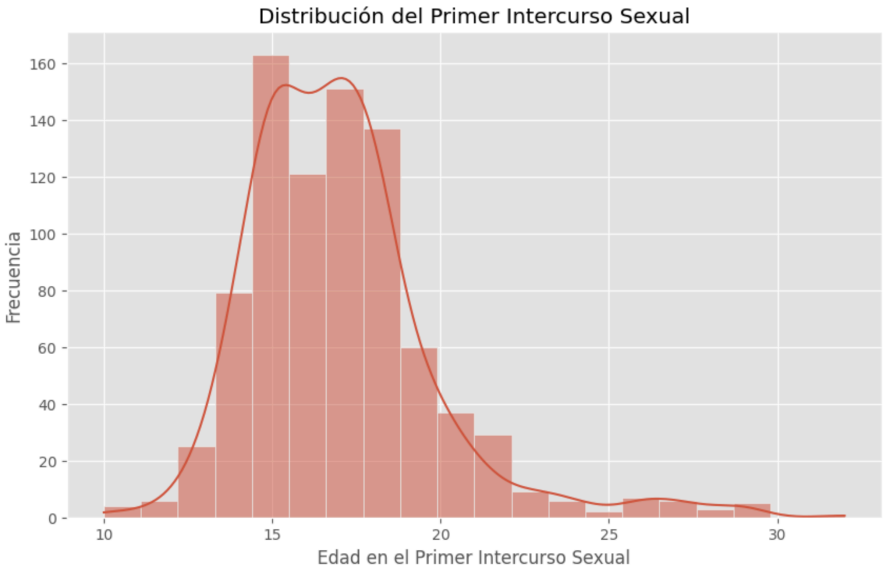

Gráfico 03. Distribución de las edades del primer intercambio sexual.

- La edad del primer intercambio sexual está concentrada en la adolescencia.

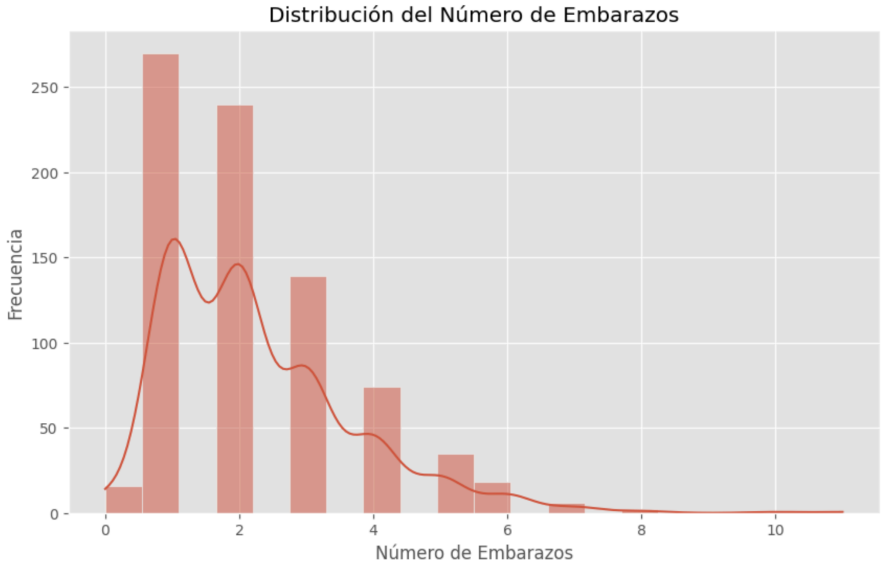

Gráfico 04. Distribución del número de embarazos según su frecuencia.

- La mayoría de la población tiene un número relativamente bajo de embarazos.

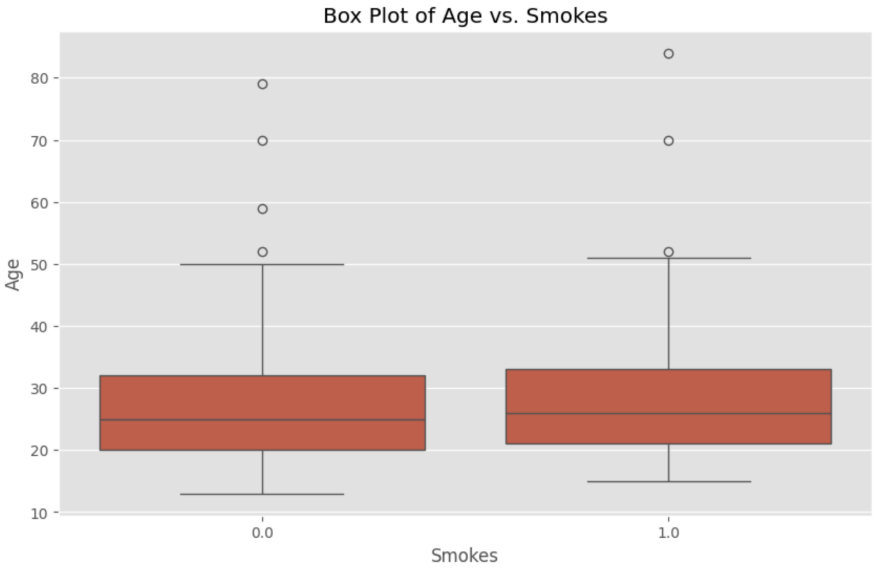

Gráfico 05. Diagrama de caja y bigotes que muestra la distribución de las edades en las que las personas fuman o no.

- No hay diferencia significativa en la distribución de las edades entre las personas que fuman y las que no. Los valores atípicos están presentes en ambos grupos, lo que indica que hay edades extremas en ambas categorías.

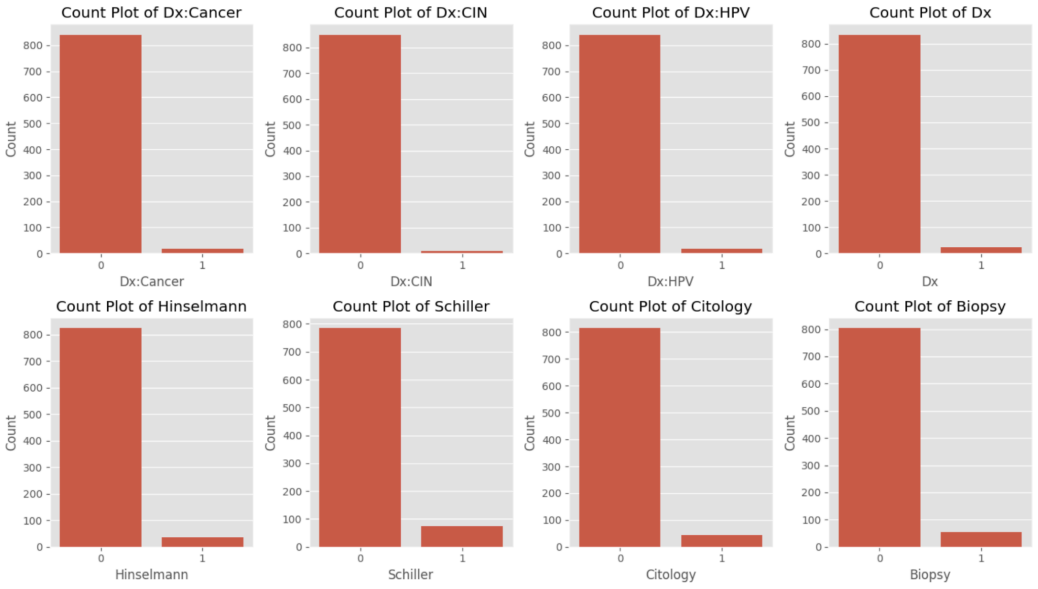

Gráfico 06. Cuenta de algunas variables categóricas.

- Para todas las pruebas y diagnósticos, la gran mayoría de los resultados son negativos.

4. *Aísle las variables numéricas de las categóricas, realice un análisis de correlación entre las mismas.*

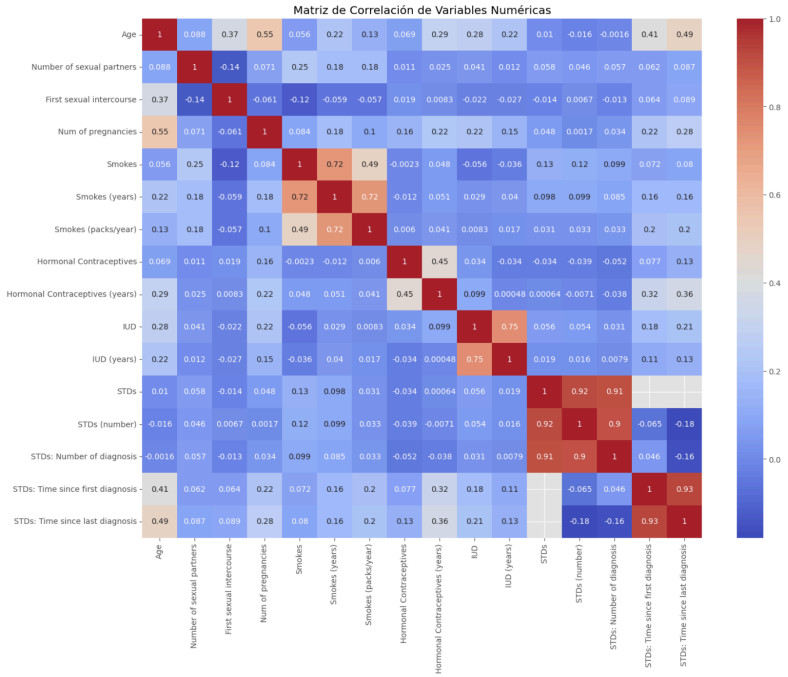

Gráfico 07. Matriz de correlación de variables numéricas.

- Se ha encontrado una alta correlación entre las siguientes variables:
    - Smokes y Smoke (years)
    - STDs, STDs (number), STDs: Number of diagnosis
    - Hormonal Contraceptives, Hormonal Contraceptives (years)
- Correlaciones moderadas entre las variables:
    - Age y el número de embarazos
    - Primer encuentro sexual y edad
- Correlaciones bajas con el resto de variables que no están fuertemente relacionadas.

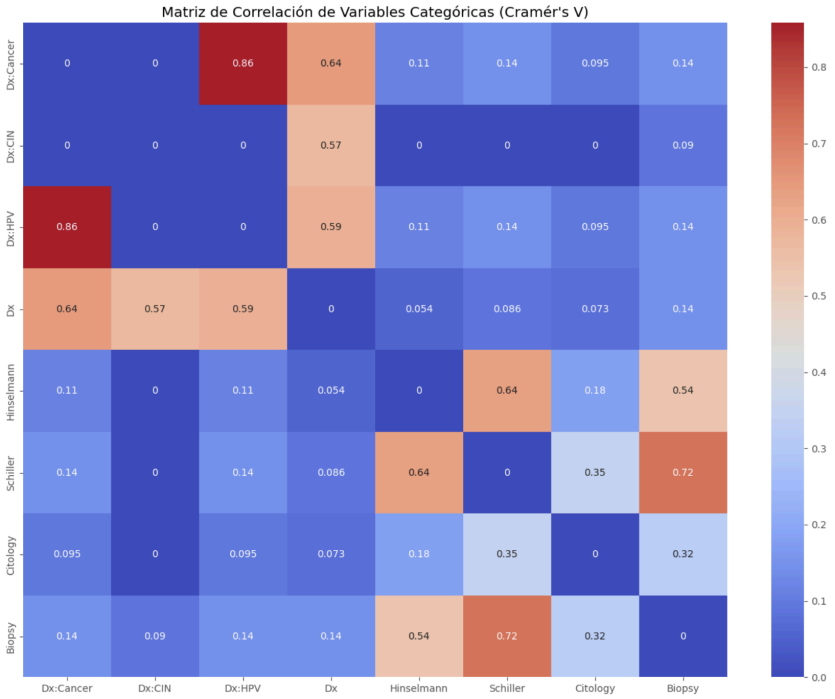

Gráfico 08. Matriz de correlación de las variables categóricas.

- Correlación alta entre:
  - Dx: Cancer y Dx: HPV
  - Dx: Cancer y Dx
  - Schiller y Biopsy
- Correlación moderada:
  - Dx:CIN y Dx:HPV
  - Hinselmann y Biopsy
- Correlaciones bajas con el resto de variables que no están fuertemente relacionadas.

Como análisis general, puede indicarse que las variables relacionadas con fumar y las relacionadas con las ETS muestran altas correlaciones dentro de sus grupos, lo cual es lógico. Por otro lado, las correlaciones entre variables categóricas muestran asociaciones entre ciertos diagnósticos y resultados de pruebas. La mayoría de las otras correlaciones entre variables numéricas y categóricas son bajas, lo que indica que no hay muchas relaciones fuertes.

5. *Utilice las variables categóricas, cree tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos.*

Se obtuvieron los siguientes resultados:

	Frecuencia		Proporción (%)	
Dx: Cancer	0	840	0	0.979021
	1	18	1	0.020979
Dx: CIN	0	849	0	0.98951
	1	9	1	0.98951
Dx: HPV	0	840	0	0.979021
	1	18	1	0.020979
Dx	0	834	0	0.972028
	1	24	1	0.027972
Hinselmann	0	823	0	0.959207
	1	35	1	0.040793
Schiller	0	784	0	0.913753
	1	74	1	0.086247
Citology	0	814	0	0.948718
	1	44	1	0.051282
Biopsy	0	803	0	0.935897
	1	55	1	0.064103

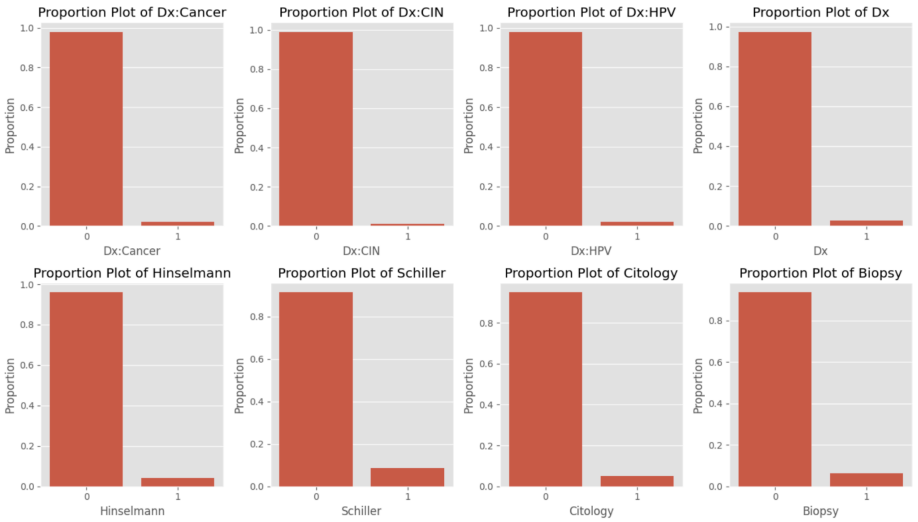

Gráfico 09. Gráfico de barras normalizado para visualizar las proporciones de las variables categóricas.

Las gráficas de proporción confirman visualmente las observaciones de las tablas de frecuencia y proporción. En todas las variables categóricas, la proporción de resultados negativos es mayor que la de resultados positivos. De manera general puede decirse que la mayoría de personas no tienen diagnósticos positivos de alguna de las enfermedades.

6. *Determine el tratamiento 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?*

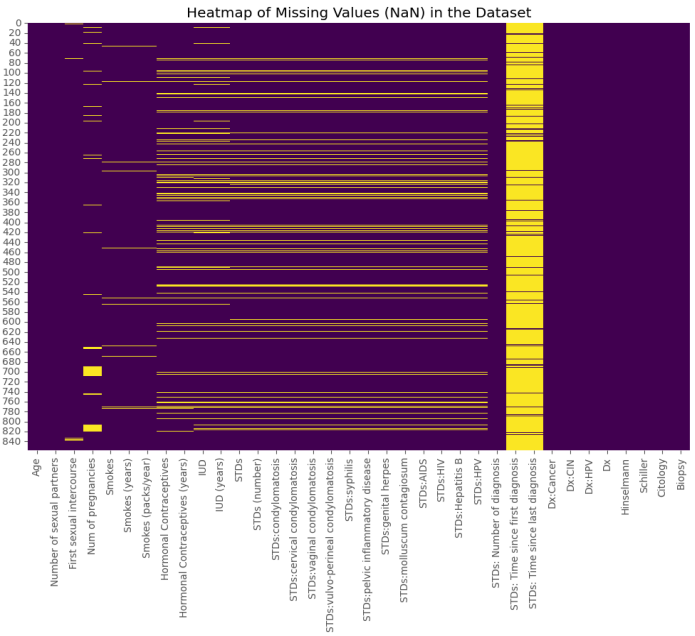

Gráfico 10. Mapa de calor de los valores faltantes en cada columna del dataset

Los valores faltantes en el dataset reflejados en el *Gráfico 10*, se observa una distribución significativa a lo largo de las variables.Debido a la cantidad de valores NaN, que asciende a un total de 3596, se determinó que no era factible eliminar todos los valores faltantes sin comprometer la integridad del conjunto de datos.

Por lo tanto, se utilizó una estrategia para manejarlos acorde al contexto, es decir, para las variables categóricas se utilizó la moda y para las cuantitativas discretas la mediana.

Además, se procedió a eliminar aquellas variables que presentaban una baja correlación y un alto volumen de valores faltantes.

7. *Estudie si es posible hacer transformaciones en las variables categóricas para incluirlas en el PCA, ¿valdrá la pena?*
    - Sí vale la pena incluir las variables categóricas transformadas porque contienen información relevante que no se capturan en las variables numéricas. En este conjunto de datos, será relevante porque de tal forma se puede obtener información de los diagnósticos. Aún así, debe analizarse cuáles variables son las que aportan información valiosa para ese tipo de análisis. Vale la pena luego de realizar las pruebas preliminares para interpretar los componentes principales y que mejoren la comprensión de los datos.

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. Realice un análisis de componentes principales con las variables numéricas, discuta los resultados e interprete los componentes.*
    - Se obtuvo una matriz de correlación muy similar a la del Gráfico 07, con la diferencia que ya no se incluyen las variables que han sido eliminadas o limpiadas. Luego de realizar el test de esfericidad de Bartlett y KMO, se obtuvieron los resultados:

      ```
      KMO index: 0.6511079595277759
      Bartlett's test chi-square value: 7209.422966129833
      Bartlett's test p-value: 0.0
      ```

        - El índice de KMO aunque no es tan alto, es aceptable y sugiere una porción moderada de la varianza.

    - El test de esfericidad tiene un valor de chi-cuadrado alto y el valor de p es cero, lo que indica que las variables están correlacionadas.
- Con el análisis de componentes principales, se realizó un gráfico de varianza acumulada por PCA que es el siguiente:

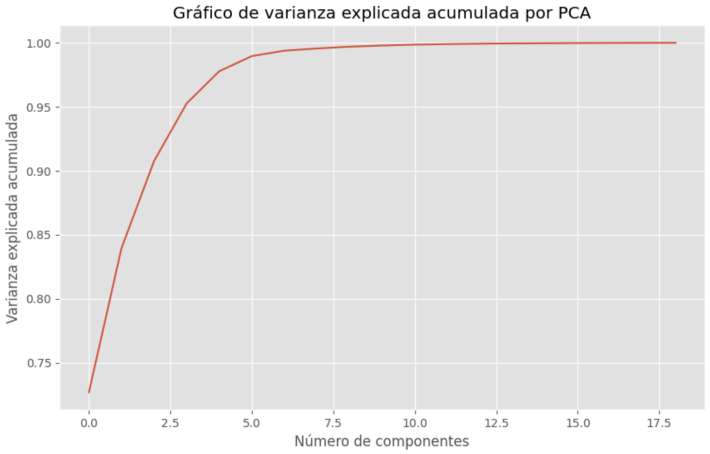

Gráfico 11. Varianza acumulada por PCA.

Con este, podemos darnos una idea de lo que explica cada componente. Pero con los resultados del array que explica la varianza, indica que el primer componente explica aproximadamente el 72.6% de la variabilidad total, con dos componentes el 83.9% y alrededor de cinco componentes más del el 97.7% de la variabilidad.

- Con lo anterior, se decidió tomar los primeros tres componentes para analizar cómo cada variable contribuye a ellos.
    - PC1: representa una variación fuerte con la edad de los individuos y el tiempo que han usado anticonceptivos hormonales.
    - PC2: captura la variabilidad con la duración del uso de anticonceptivos hormonales.
    - PC3: relaciona aspectos de la edad de inicio sexual con el consumo de tabaco.

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 obtener más insights de la generación de reglas, hágalo y discútalo.*

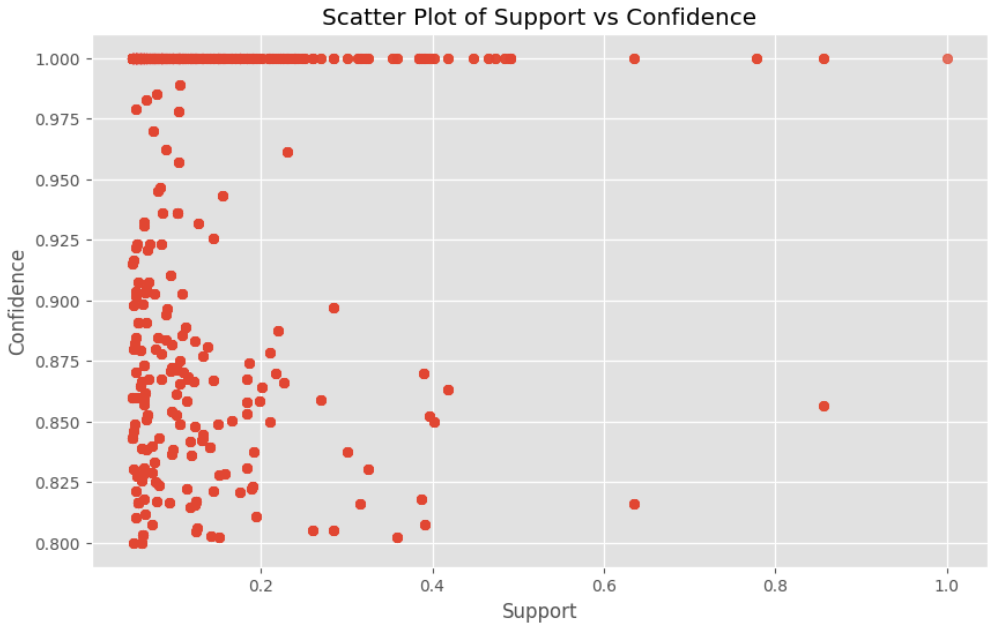

Gráfico 12. Dispersión del soporte vs confianza

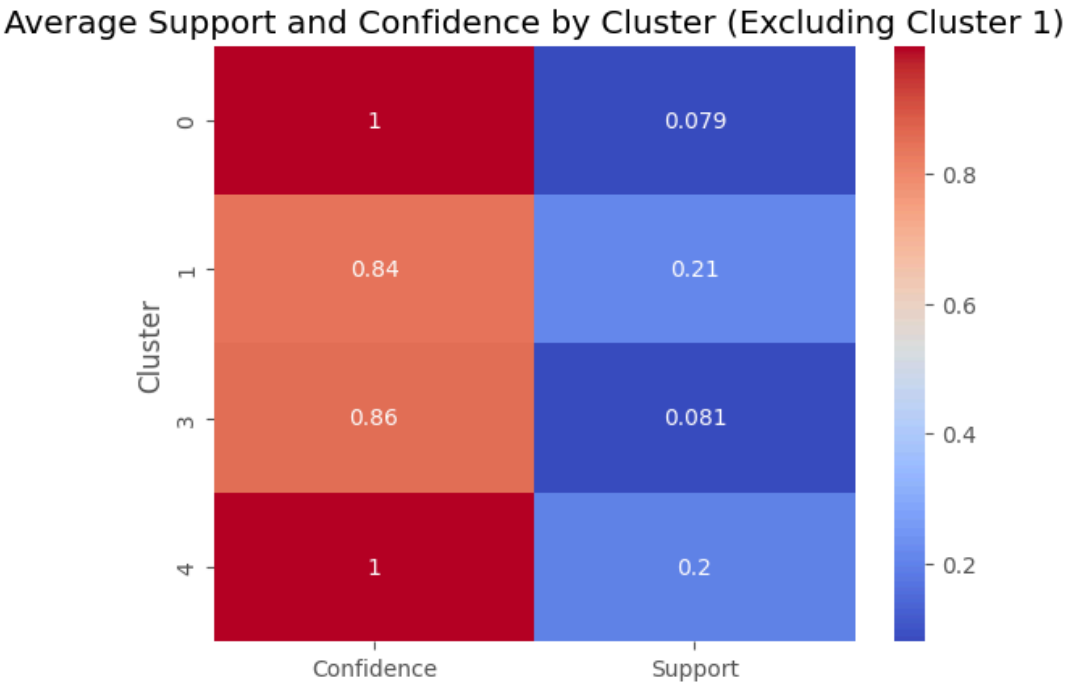

Gráfico 13. Soporte y confianza por clúster

Inicialmente, al aplicar Apriori se identificaron más de 22,000 reglas de asociación, lo que llevó a la necesidad de simplificar el análisis mediante la eliminación de variables muy comunes.

Esto último resultó en la creación de clusters, cada uno con características únicas de soporte y confianza, que muestran variaciones significativas en la fuerza y frecuencia de las asociaciones entre diferentes factores de riesgo, como se observa en el Gráfico 12. En este gráfico, es evidente que la mayoría de las reglas tienen un alto grado de confianza, superando el 90%, aunque el soporte varía ampliamente, reflejando la diversidad en la frecuencia con la que ocurren ciertas combinaciones de factores en el dataset manifestado en el Gráfico 13.

El Cluster 0, caracterizado por una alta confianza pero soporte bajo, incluye reglas que relacionan rangos de edad específicos con comportamientos como el no fumar. Aunque no son comunes, estas relaciones son consistentes cuando ocurren, como indica la regla más destacada de este cluster, que asocia a adolescentes de 15 a 18 años que no fuman con un bajo riesgo de cáncer cervical, reflejado por el resultado '0'.

El Cluster 1, con soporte moderado y alta confianza, vincula a adultos jóvenes de 20 a 30 años con un posible factor positivo o de control, representado por '1.0', sugiriendo una mayor probabilidad de tener cáncer cervical. Este cluster muestra una variabilidad en el soporte que refleja una frecuencia relativamente más alta de estas combinaciones en comparación con el Cluster 0.

Similar al Cluster 0, el Cluster 3 muestra patrones menos frecuentes pero confiables entre personas mayores de 21 años y el no fumar, asociados con un mayor riesgo de cáncer cervical, marcado por '1.0'. Esta asociación, también con confianza alta pero soporte bajo, resalta cómo ciertos comportamientos y edades están ligados a un incremento en el riesgo.

Finalmente, el Cluster 4, con una confianza perfecta y soporte moderado, resalta fuertemente la relación entre no fumar y jóvenes de 15 a 18 años con un bajo riesgo de cáncer cervical, evidenciado por el resultado '0'. Este cluster subraya la importancia de ciertos comportamientos saludables en grupos demográficos específicos y cómo estos se reflejan en los datos analizados.

Hallazgos

- Las variables relacionadas con el fumar y las ETS mostraron fuertes correlaciones, lo que destaca la importancia de estos comportamientos en la incidencia del cáncer cervical.
- Las reglas de asociación indicaron que jóvenes que no fuman presentan un menor factor de riesgo.
- El análisis de componentes principales señaló que la edad y el uso de anticonceptivos hormonales son factores significativos en la variabilidad de los datos.

Conclusiones

- Los patrones encontrados sugieren que intervenciones educativas y preventivas enfocadas en hábitos de fumar y uso de anticonceptivos desde una edad temprana podrían ser clave para reducir el riesgo de cáncer cervical.
- La efectividad de las medidas preventivas podría incrementarse al centrarse en los grupos de riesgo identificados, especialmente en jóvenes y adolescentes.

Referencias

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors