Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Analisis Exploratorio lab 1DS

Introducción

Para este laboratorio se realizó como primera etapa una exploración una limpieza de la data , para la exploración se utilizó el pandas Profiler y en este informe se discutirán los resultados. Antes de las exploraciones se realizó una limpieza también detallada en este informe. Adjunto con el informe el informe de profiler completo y además el código usado.

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Exploración rápida de los datos (1)

Estadísticas Descriptivas:

	Age	Number of sexual partners	First sexual intercourse	\
count	858.000000	832.000000	851.000000	
mean	26.820513	2.527644	16.995300	
std	8.497948	1.667760	2.803355	
min	13.000000	1.000000	10.000000	
25%	20.000000	2.000000	15.000000	
50%	25.000000	2.000000	17.000000	
75%	32.000000	3.000000	18.000000	
max	84.000000	28.000000	32.000000	

	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	\
count	802.000000	845.000000	845.000000	845.000000	
mean	2.275561	0.145562	1.219721	0.453144	
std	1.447414	0.352876	4.089017	2.226610	
min	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	0.000000	0.000000	0.000000	
50%	2.000000	0.000000	0.000000	0.000000	
75%	3.000000	0.000000	0.000000	0.000000	
max	11.000000	1.000000	37.000000	37.000000	

	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	\
count	750.000000	750.000000	741.000000	
mean	0.641333	2.256419	0.112011	
std	0.479929	3.764254	0.315593	
min	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	
50%	1.000000	0.500000	0.000000	
75%	1.000000	3.000000	0.000000	
max	1.000000	30.000000	1.000000	

	...	STDs: Time since first diagnosis	STDs: Time since last diagnosis	\
count	...	71.000000	71.000000	
mean	...	6.140845	5.816901	
std	...	5.895024	5.755271	
min	...	1.000000	1.000000	
25%	...	2.000000	2.000000	
50%	...	4.000000	3.000000	
75%	...	8.000000	7.500000	
max	...	22.000000	22.000000	

	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	\
count	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	
mean	0.020979	0.010490	0.020979	0.027972	0.040793	0.086247	
std	0.143398	0.101939	0.143398	0.164989	0.197925	0.280892	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

	Citology	Biopsy
count	858.000000	858.000000
mean	0.051282	0.064103
std	0.220701	0.245078
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	1.000000	1.000000

Y también se utilizó ydata-profiling de manera que se encuentra en un html.

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Descripción de las variables (2)

Age' es: Cuantitativa Discreta
'Number of sexual partners' es: Cuantitativa Continua
'First sexual intercourse' es: Cuantitativa Continua
'Num of pregnancies' es: Cuantitativa Continua
'Smokes' es: Categórica
'Smokes (years)' es: Cuantitativa Continua
'Smokes (packs/year)' es: Categórica
'Hormonal Contraceptives' es: Categórica
'Hormonal Contraceptives (years)' es: Cuantitativa Continua
'IUD' es: Categórica
'IUD (years)' es: Categórica
'STDs' es: Categórica
'STDs (number)' es: Categórica
'STDs:condylomatosis' es: Categórica
'STDs:cervical condylomatosis' es: Categórica
'STDs:vaginal condylomatosis' es: Categórica
'STDs:vulvo-perineal condylomatosis' es: Categórica
'STDs:syphilis' es: Categórica
'STDs:pelvic inflammatory disease' es: Categórica
'STDs:genital herpes' es: Categórica
'STDs:molluscum contagiosum' es: Categórica
'STDs:AIDS' es: Categórica
'STDs:HIV' es: Categórica
'STDs:Hepatitis B' es: Categórica
'STDs:HPV' es: Categórica
'STDs: Number of diagnosis' es: Cuantitativa Discreta
'STDs: Time since first diagnosis' es: Categórica
'STDs: Time since last diagnosis' es: Categórica
'Dx:Cancer' es: Cuantitativa Discreta
'Dx:CIN' es: Cuantitativa Discreta
'Dx:HPV' es: Cuantitativa Discreta
'Dx' es: Cuantitativa Discreta
'Hinselmann' es: Cuantitativa Discreta
'Schiller' es: Cuantitativa Discreta
'Citology' es: Cuantitativa Discreta
'Biopsy' es: Cuantitativa Discreta

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Gráficos Exploratorios y tablas de frecuencia (3 y 5)

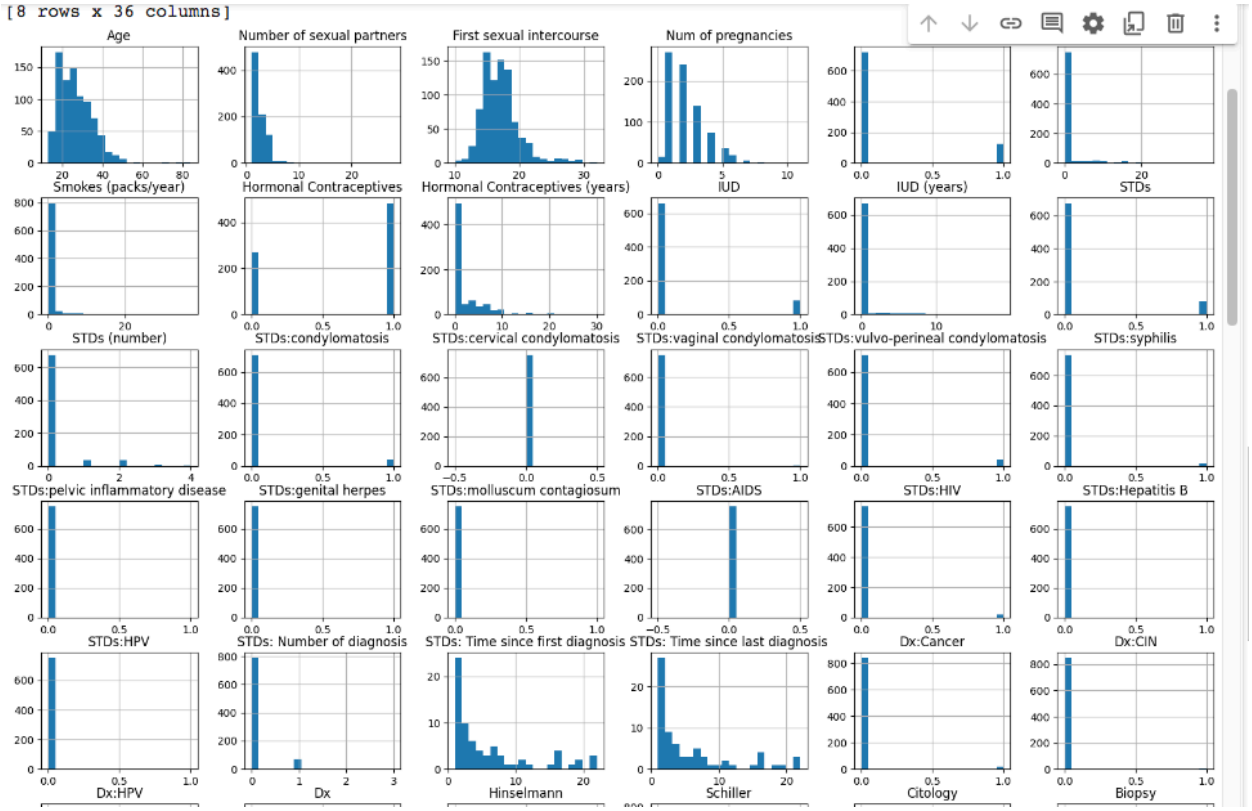
Procedimiento de Limpieza de Datos

El siguiente procedimiento detalla los pasos realizados para limpiar el conjunto de datos `risk_factors_cervical_cancer.csv`, asegurando la integridad y la calidad de los datos para un análisis preciso.

1. **Carga del Conjunto de Datos**: Se cargó el conjunto de datos utilizando la función `pd.read_csv` de la biblioteca `pandas`.
2. **Reemplazo de Valores Faltantes**: Los valores faltantes representados por '?' se reemplazaron con `NaN` para facilitar el manejo posterior de estos valores.
3. **Visualización de Valores Faltantes**: Se mostró la cantidad de valores faltantes por variable antes de realizar la limpieza, proporcionando una visión general del estado inicial del conjunto de datos.
4. **Eliminación de Columnas con Altos Valores Faltantes**: Se eliminaron las columnas que tenían más del 30% de sus valores como faltantes, asegurando que solo se mantuvieran las variables con datos más completos.
5. **Conteo de Filas Antes de la Limpieza**: Se contó el número de filas antes de comenzar la limpieza más intensiva para tener una referencia de la magnitud de los datos iniciales.
6. **Conversión de Columnas a Tipos Numéricos**: Las columnas relevantes se convirtieron a tipos de datos numéricos para permitir un análisis cuantitativo adecuado.
7. **Imputación de Valores Faltantes**:
    - Para las variables numéricas, los valores faltantes se imputaron utilizando la media de cada columna, asegurando que los datos fueran consistentes y representativos.
    - Para las variables categóricas, los valores faltantes se imputaron utilizando la moda (el valor más frecuente) de cada columna.
8. **Eliminación de Filas con Valores Faltantes Excesivos**: Se eliminaron las filas que tenían más del 35% de sus valores como faltantes para mejorar la calidad general del conjunto de datos. Se contabilizó el número de filas antes y después de este paso para determinar cuántas fueron eliminadas.
9. **Detección y Manejo de Outliers**: Se utilizó el método del Rango Intercuartílico (IQR) para identificar y manejar los valores atípicos. Se calcularon los límites inferior y superior, y se filtraron los datos que se encontraban fuera de estos límites, eliminando así los outliers.

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

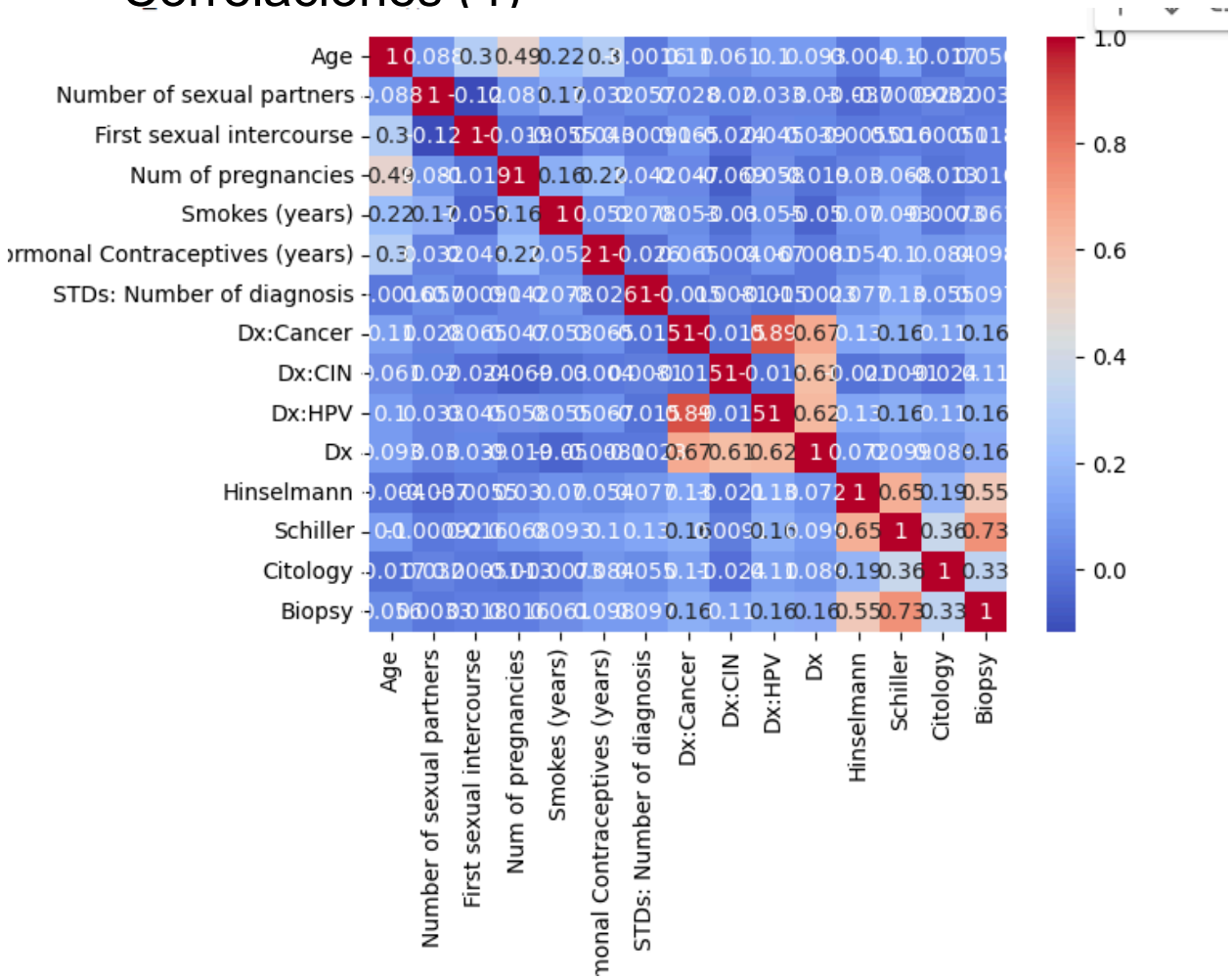
10. **Conversión de Variables Categóricas**: Finalmente, se convirtieron las variables categóricas al tipo de dato 'category' para un manejo más eficiente y preciso en análisis posteriores.



Al estudiar las distribuciones de las diferentes variables se puede resaltar que existe gran desbalance en muchas de las variables categóricas. Un tipo es considerablemente más común que los demás se podría usar alguna técnica de balanceo o de reducción como smote. En cuanto a las numéricas muchas están sesgadas a la izquierda.

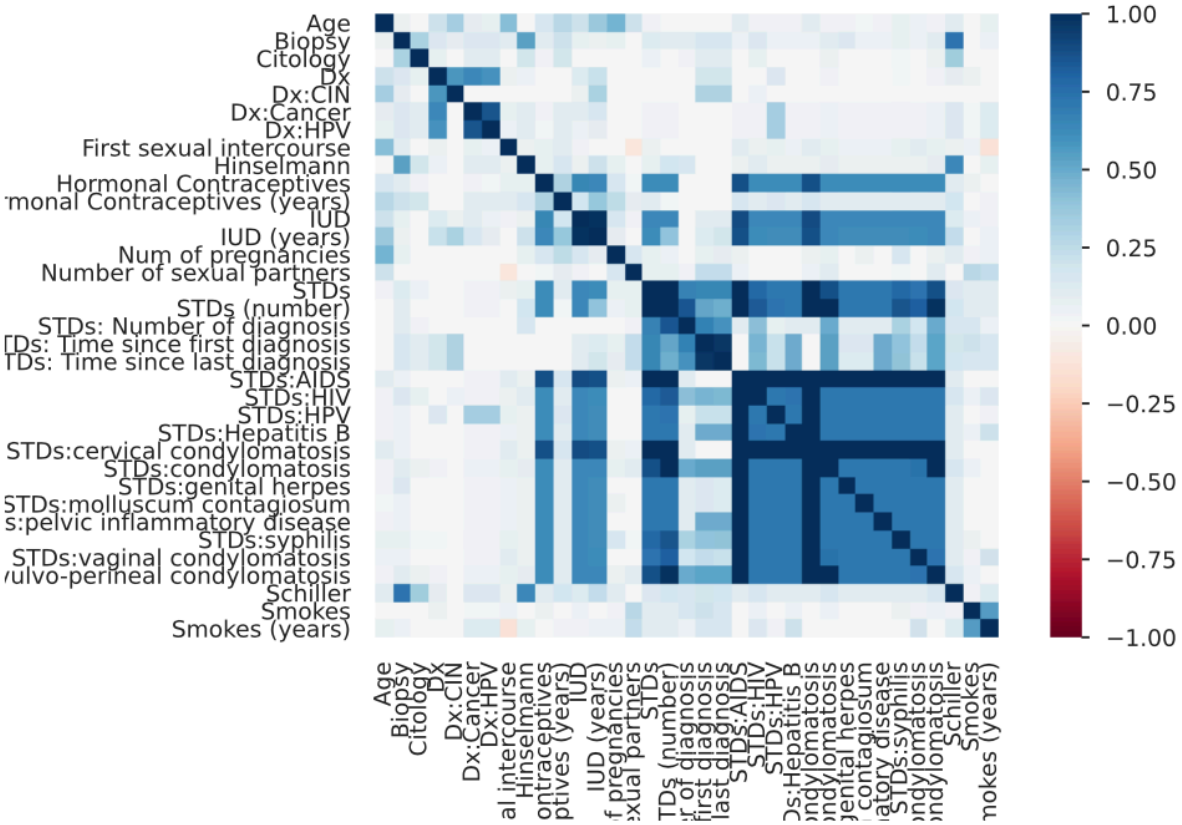
Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Correlaciones (4)



Al realizar una correlación entre variables, pero únicamente para las cualitativas, se donde se observó más de cerca la correlación entre el cáncer cervical y las distintas posibles causas, la más correlacionada fue en este caso el diagnóstico de VPH.

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172



Al estudiar la correlación entre todas la variables se puede ver que las todas las variables de STD están fuertemente correlacionadas especialmente la de AIDS por lo que sería útil usar esa variable como representación de resto. Un Caso similar ocurre con las variables de diagnóstico , pero hay que tener en cuenta el desbalance como una posible causa de la aparente correlación.

Tratamiento de valores faltantes (6)

Tal cómo se explicó durante el procedimiento de limpieza de datos, los valores faltantes en las variables numéricas se utilizó la media de cada columna de manera que los datos fueran consistentes y representativos. Así mismo, se utilizó la moda para poder tratar los datos faltantes en el caso de las variables categóricas, a su vez, en caso de que una fila no contara con el mínimo umbral de valores llenos se borraría dicha fila dado que realmente no sería representativa.

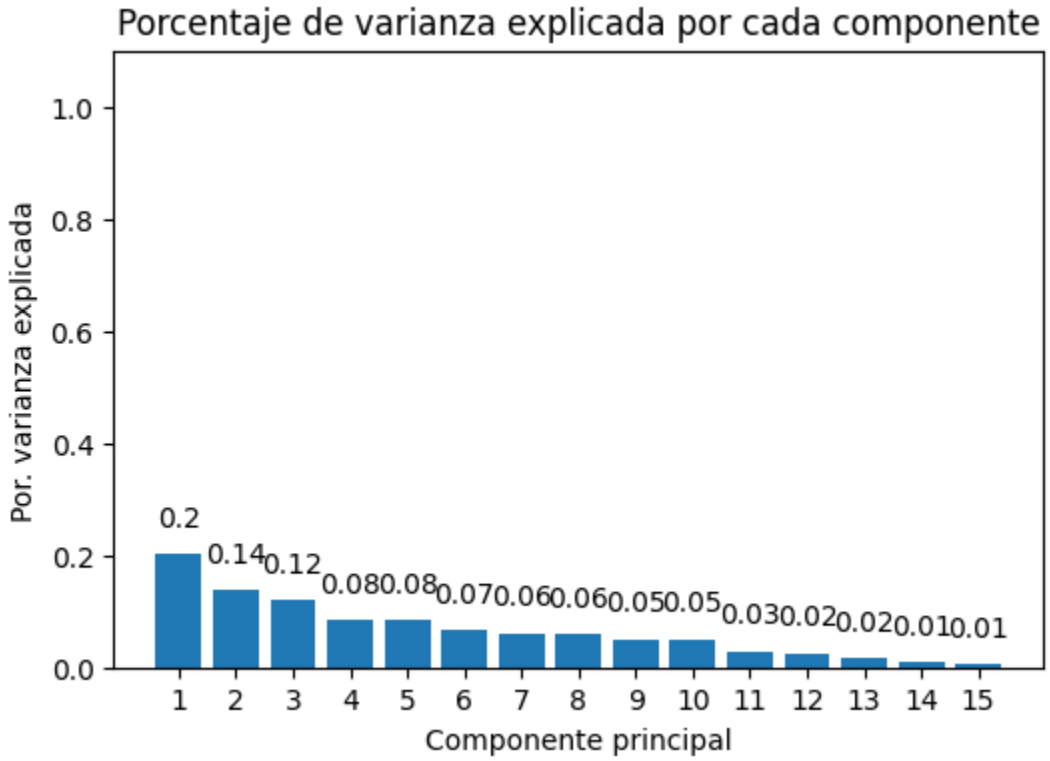
Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Transformaciones en las variables categóricas para el PCA (7)

Sï se podrían realizar algunas transformaciones en las variables categóricas para que fueran válidas como numéricas, sin embargo, realmente no serían de trascendencia para el PCA. ¿Por qué? Porque las variables categóricas presentes responden respuestas de sí o no, realmente mantendrían su esencia categórica y no numérica por lo que no serían significativas para el PCA.

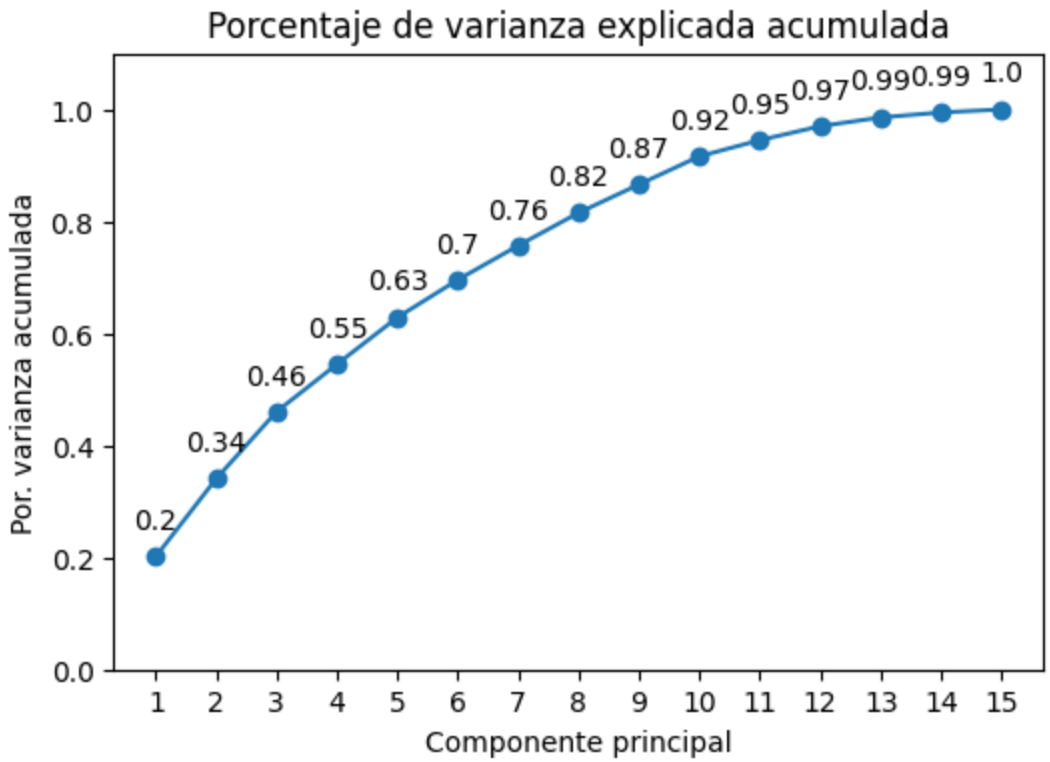
Análisis de la necesidad de los componentes principales (8)

Así mismo, al realizar el test de esfericidad de Bartlett se obtuvo 0.0 de índice lo que rechaza la hipótesis nula, por lo tanto las variables están suficientemente correlacionadas para poder realizar el PCA pertinente. También se realizó la prueba de adecuación de Kaiser-Meyer Olkin (KMO) para determinar el grado de relación conjunta de las variables, se obtuvo un 0.60 de correlación, a pesar de que bajo la literatura para realizar la factorización requiere que al menos se obtenga un KMO de 0.8. Por lo tanto, ambos tests concluyen que existen variables correlacionadas siendo recomendable realizar el PCA.



Tal cómo se observa, el componente PC1 es con una diferencia bastante sustancial el que más explica la varianza, seguido por el PC2 y PC3.

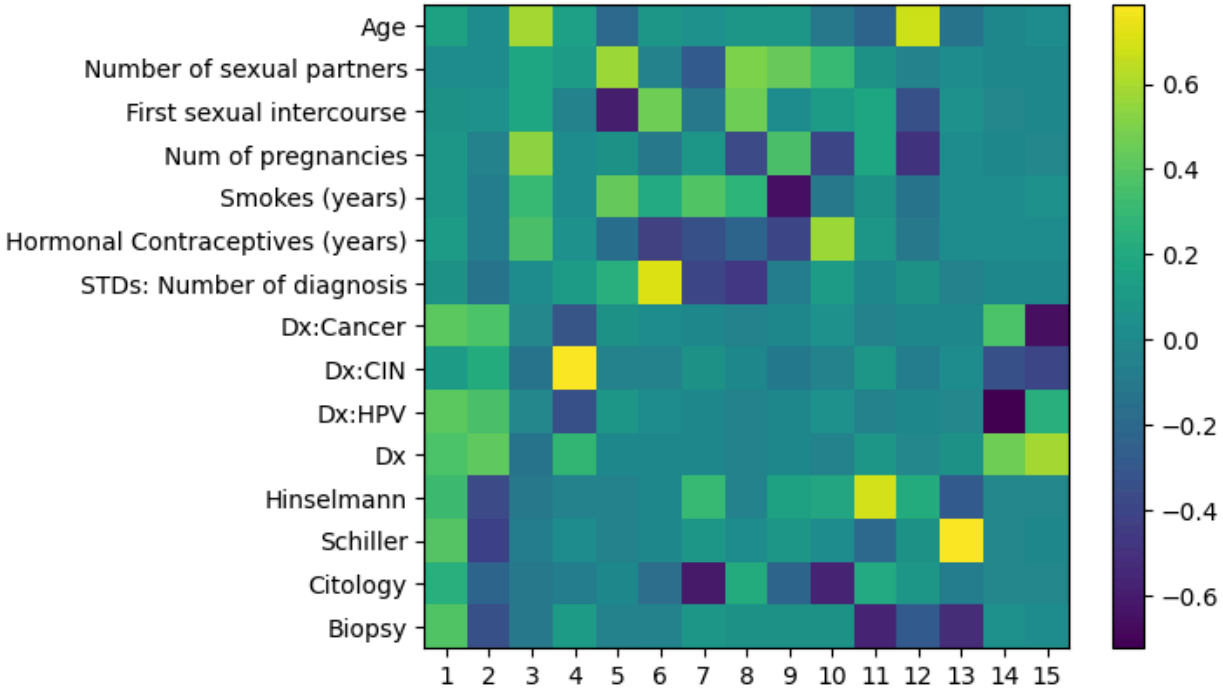
Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172



Al utilizar todos los componentes principales (el 100% de varianza acumulada), se conserva la totalidad de la varianza original, lo que significa que no se pierde información en el proceso de reducción de dimensionalidad. Sin embargo, es posible que no todos los componentes sean igualmente informativos o relevantes para el análisis, esto se observa debido a que solo en los primeros 5 componentes principales se conserva más del 50% de varianza acumulada lo cual es cantidad significativa de la varianza original.

	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes (years)	Hormonal Contraceptives (years)	STDs: Number of diagnosis	Dx:Cancer	Dx:CIN	Dx:HPV	Dx	Hinselmann	Schiller	Citology	Biopsy
PC1	0.148392	0.034423	0.053063	0.089942	0.081011	0.116114	0.060519	0.411067	0.122452	0.404499	0.370489	0.325286	0.390698	0.228126	0.383186
PC2	0.031514	0.015101	0.046986	-0.038096	-0.078637	-0.060569	-0.132983	0.372395	0.204040	0.360788	0.426290	-0.367481	-0.410954	-0.212752	-0.355236
PC3	0.590063	0.167767	0.174272	0.539164	0.306037	0.359031	0.013041	-0.021822	-0.126900	-0.018243	-0.136378	-0.119735	-0.071066	-0.105025	-0.118112
PC4	0.143565	0.125823	-0.046474	0.022961	0.013864	0.041857	0.118901	-0.326216	0.787756	-0.338024	0.290546	-0.034993	0.025153	-0.062936	0.115402
PC5	-0.198693	0.568614	-0.595862	0.055750	0.429046	-0.156178	0.228830	0.063937	-0.054978	0.080107	-0.006731	-0.029215	-0.034859	-0.004223	-0.050447
PC6	0.072487	-0.049811	0.462471	-0.102308	0.198014	-0.434411	0.710461	0.032850	-0.051145	0.019696	0.004096	0.003649	0.002588	-0.173453	-0.032181
PC7	0.040190	-0.288372	-0.101796	0.083055	0.377697	-0.329115	-0.397539	0.002377	0.054338	0.004465	0.000776	0.306560	0.091068	-0.614859	0.075191
PC8	0.082684	0.501404	0.465145	-0.360206	0.263606	-0.213377	-0.474202	-0.039741	0.000244	-0.052468	-0.032611	-0.050235	0.019017	0.211127	0.053927
PC9	0.102556	0.443496	0.024674	0.358355	-0.649367	-0.382661	-0.057261	-0.007784	-0.088658	-0.004642	0.007070	0.146897	0.083858	-0.229366	0.056293
PC10	-0.119915	0.303757	0.120225	-0.393168	-0.109165	0.563807	0.129421	0.044223	-0.044527	0.042769	-0.054858	0.195381	0.020673	-0.572165	0.054956
PC11	-0.221969	0.054735	0.169694	0.174517	0.059599	0.074793	0.000690	-0.036069	0.093594	-0.048778	0.099762	0.681895	-0.189491	0.202615	-0.561251
PC12	0.680960	-0.046305	-0.346972	-0.483922	-0.132889	-0.109532	0.053253	0.007564	-0.059211	-0.007249	-0.017389	0.219168	0.051839	0.076562	-0.294254
PC13	-0.130542	0.009456	0.045718	0.031276	0.010147	0.031504	-0.037662	-0.004663	0.030419	-0.021159	0.053579	-0.270508	0.785297	-0.082847	-0.525634
PC14	-0.001666	-0.003323	-0.020940	-0.002687	0.036090	0.032740	-0.004575	0.373681	-0.338214	-0.723881	0.464830	-0.012671	-0.012314	-0.023378	0.046801
PC15	0.009929	-0.006749	-0.003975	-0.019750	0.041944	0.035755	-0.004293	-0.660050	-0.400869	0.230883	0.587045	-0.010671	-0.003858	-0.025825	0.036151

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172



Esos son los resultados del PCA y el posible significado de los componentes más significativos (aunque el PC4 y PC5 podrían descartarse por el porcentaje de varianza explicada):

1. PC1: está fuertemente influenciada por Dx: Cáncer, Dx: HPV, Dx, Schiller, Biopsia. Esto significa que realmente este componente estaría evaluando sí los pacientes fueron sometidos a alguna clase de diagnóstico o no, dado que está fuertemente por diagnósticos (cáncer, virus del papiloma humano, diagnósticos en general) y pruebas (biopsia, colonoscopía, Schiller).
2. PC2: está influenciada únicamente por los diagnósticos y por las pruebas inversamente, por lo que determina si los pacientes tuvieron algún tipo de diagnóstico en general lo que genera un excluyente que los mismos pacientes hayan tenido pruebas, o viceversa. Por ende, este componente sería un complemento del componente 1.
3. PC3: está influenciada por: número de embarazos, la edad, años que lleva fumando, cantidad de años con uso de anticonceptivos y un poco por variables cómo la cantidad de parejas sexuales y la edad en la que fue desvirgado. Por lo que este componente podría determinar la actividad sexual que poseen los pacientes.
4. PC4: está influenciada también por los diagnósticos por lo tanto sería otro complemento para los primeros dos componentes principales.

5. PC5: está influenciada variables como el número de parejas, primer encuentro sexual, número de diagnósticos de STD's, de manera que este componente sería otro complemento al componente PC2.

Reglas de asociación (9)

1. Smokes
2. Smokes (packs/year)
3. Hormonal Contraceptives
4. IUD
5. IUD (years)
6. STDs
7. STDs (number)
8. STDs:condylomatosis
9. STDs:cervical condylomatosis
10. STDs:vaginal condylomatosis

```
Probando con soporte=0.1 y confianza=0.6
  Base       Add   Support  Confidence  Lift
0            0.0  1.000000    1.000000   1.0
1            1.0  0.806452    0.806452   1.0
2       1.0, 0.0  0.806452    0.806452   1.0
3  0.0       1.0  0.806452    0.806452   1.0
4  1.0       0.0  0.806452    1.000000   1.0
```

```
Probando con soporte=0.1 y confianza=0.7
  Base       Add   Support  Confidence  Lift
0            0.0  1.000000    1.000000   1.0
1            1.0  0.806452    0.806452   1.0
2       1.0, 0.0  0.806452    0.806452   1.0
3  0.0       1.0  0.806452    0.806452   1.0
4  1.0       0.0  0.806452    1.000000   1.0
```

Guillermo Alfonso Furlán Estrada 20713
Diego Andrés Alonzo Medinilla 20172

Probando con soporte=0.05 y confianza=0.6

	Base	Add	Support	Confidence	Lift
0		0.0	1.000000	1.000000	1.00
1		1.0	0.806452	0.806452	1.00
2		1.0, 0.0	0.806452	0.806452	1.00
3	0.0		1.0	0.806452	0.806452	1.00
4	1.0		0.0	0.806452	1.000000	1.00
5	2.0		0.0	0.064516	1.000000	1.00
6	2.0		1.0	0.064516	1.000000	1.24
7	2.0	1.0, 0.0	0.064516	1.000000	1.24
8	2.0, 0.0		1.0	0.064516	1.000000	1.24
9	1.0, 2.0		0.0	0.064516	1.000000	1.00

Probando con soporte=0.05 y confianza=0.7

	Base	Add	Support	Confidence	Lift
0		0.0	1.000000	1.000000	1.00
1		1.0	0.806452	0.806452	1.00
2		1.0, 0.0	0.806452	0.806452	1.00
3	0.0		1.0	0.806452	0.806452	1.00
4	1.0		0.0	0.806452	1.000000	1.00
5	2.0		0.0	0.064516	1.000000	1.00
6	2.0		1.0	0.064516	1.000000	1.24
7	2.0	1.0, 0.0	0.064516	1.000000	1.24
8	2.0, 0.0		1.0	0.064516	1.000000	1.24
9	1.0, 2.0		0.0	0.064516	1.000000	1.00

De manera que se encontraron diferentes reglas utilizando diferentes parámetros, soporte de 5% con una confiabilidad del 70% o del 60%.
Hay reglas obvias que se descartarán como el hecho de que si está presente que fuma entonces habrá paquetes consumidos y viceversa.
Sin embargo, el hecho de que sí existen conceptivos hormonales está presente que fuman con un soporte de 0.06 y con una confianza del 1.0 de manera que casi que siempre que se utilizan conceptivos las personas fuman. El lift de 1.24 sugiere que la presencia de que se utilicen anticonceptivos aumenta la probabilidad de que fumen en un 24% comparado con lo esperado si fueran independientes realmente.