Universidad del Valle de Guatemala

Data Science

Sección 20

Laboratorio 1

Brian Carrillo - 21108

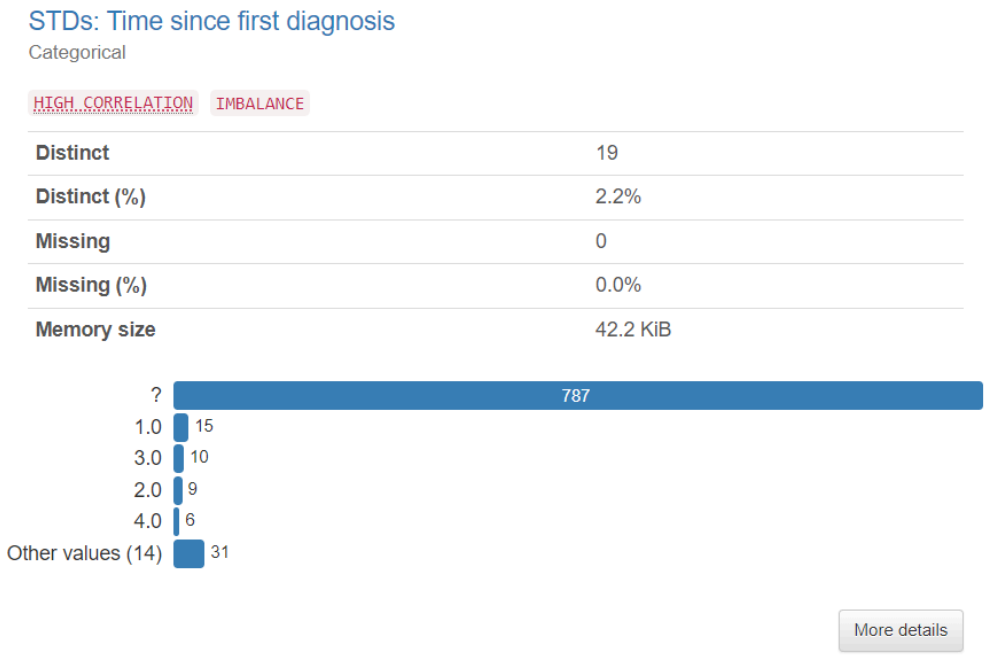
Carlos López - 21666

Guatemala, 21 de julio del 2024

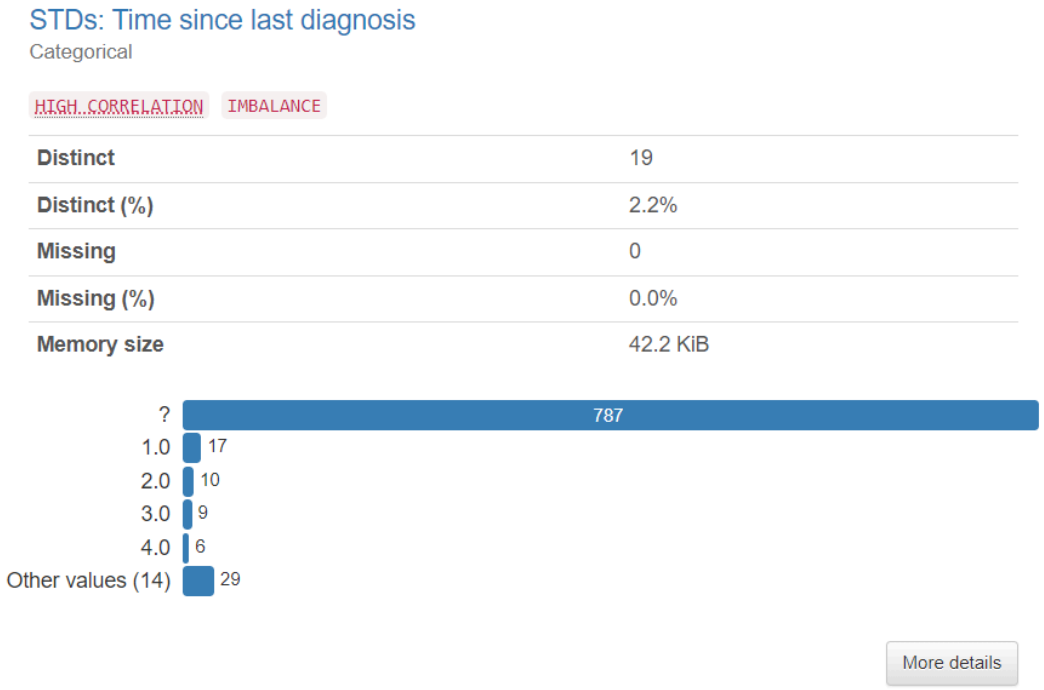
Informe de análisis exploratorio

Datos importantes sobre la exploración inicial

Herramienta automatizada utilizada: ProfileReport de la librería ydata_profiling

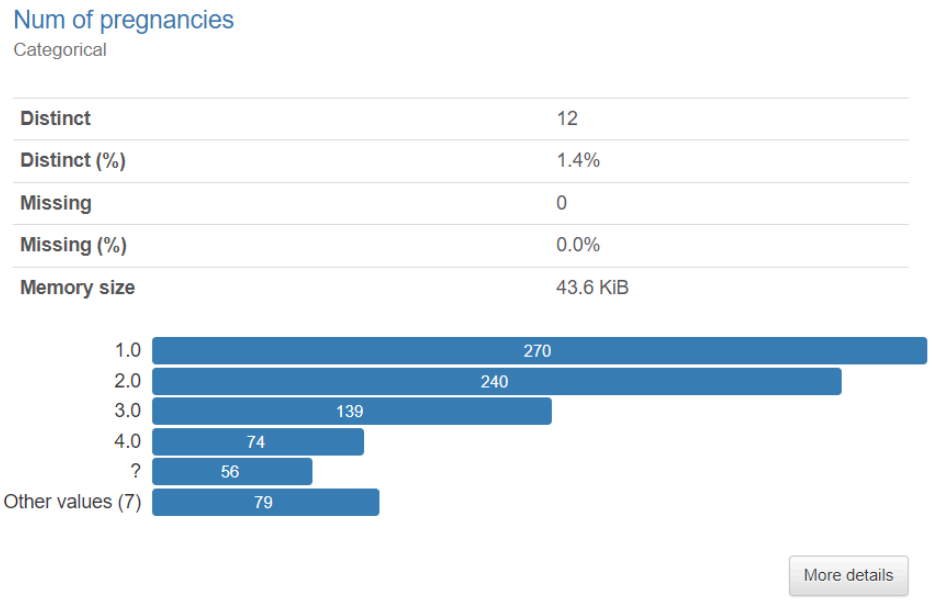


STDs: Time since first diagnosis frequencies

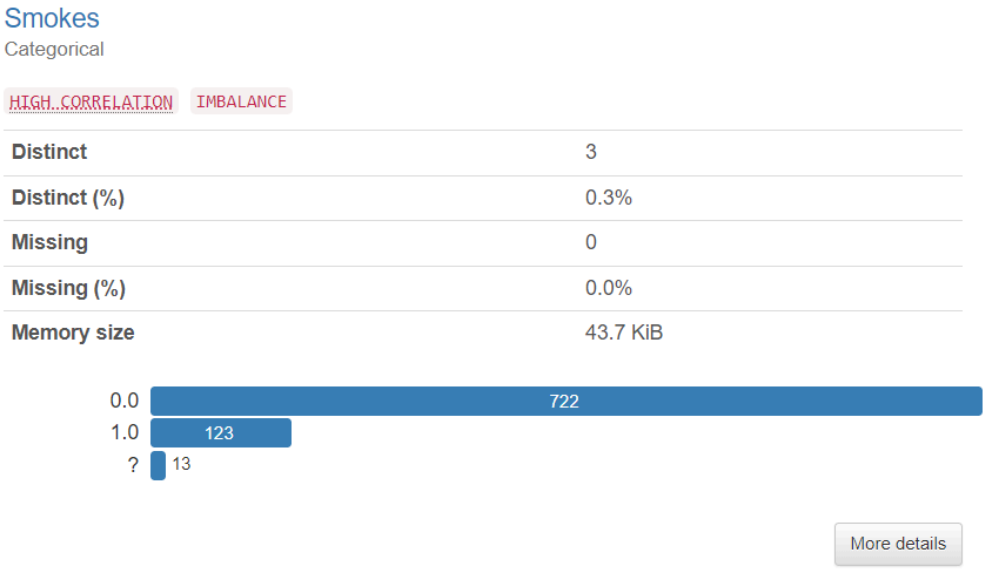


STDs: Time since last diagnosis frequencies

Las gráficas de barras permiten visualizar que estas variables categóricas presentan una amplia cantidad de valores “nulos”, que en este conjunto de datos están representados por el símbolo “?”. Por lo tanto, es posible inducir que es viable eliminar estas variables, en lugar de tratar sus valores faltantes, o bien eliminar las filas relacionadas.



Num of pregnancies frequencies



Smokes frequencies

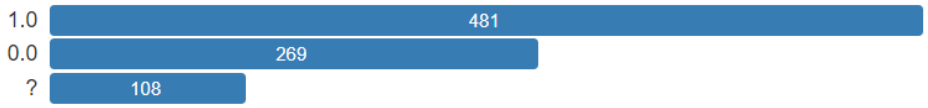
En estas variables cuantitativa y categórica respectivamente, se observa que el número de valores nulos es menor, por lo que es viable eliminar los registros con valores nulos sobre estos campos, puesto que no se tiene impacto significativo sobre la cantidad total de registros.

Hormonal Contraceptives

Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	43.5 KiB



More details

Hormonal contraceptives frequencies

IUD

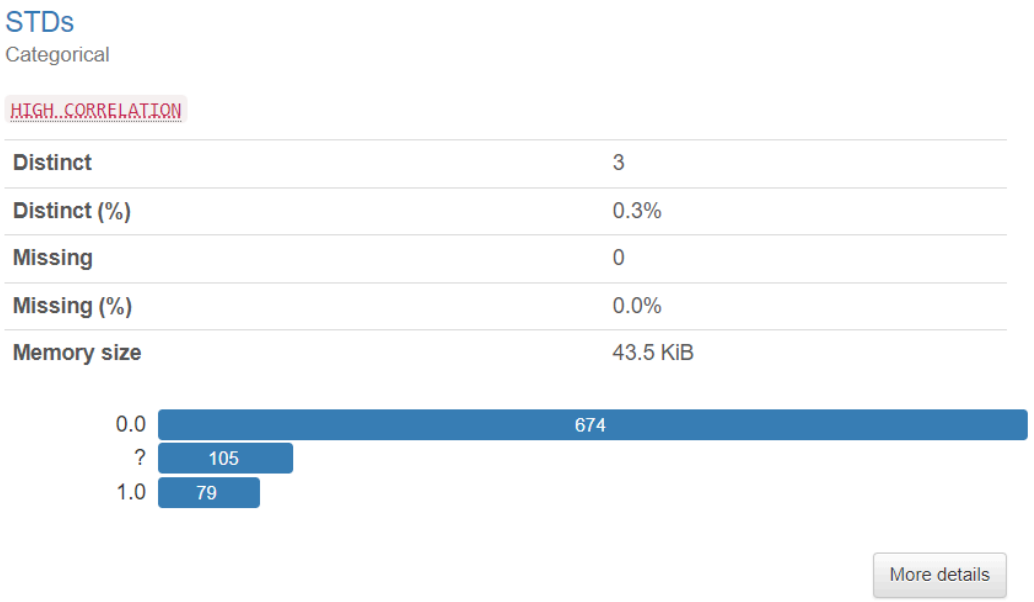
Categorical

HIGH CORRELATION

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	43.5 KiB

More details

IUD frequencies



STDs frequencies

Finalmente, con las variables Hormonal contraceptives, IUD, STDs y sus derivadas, se observa que la cantidad de valores nulos es moderada, por lo que no es viable eliminar dichos registros. En su lugar, es más óptimo el tratar los valores nulos como ceros, puesto que sus columnas derivadas también pueden ser cero como consecuencia de la variable principal.

Variables del conjunto de datos y sus tipos

- Age: cuantitativa discreta
- Number.of.sexual.partners: cuantitativa discreta
- First.sexual.intercourse: cuantitativa discreta
- Num.of.pregnancies: cuantitativa discreta
- Smokes: categórica
- Smokes.years: cuantitativa discreta
- Smokes.packs.per.year: cuantitativa discreta
- Hormonal.Contraceptives: categórica
- Hormonal.Contraceptives.years: cuantitativa discreta
- IUD: categórica
- IUD.years: cuantitativa discreta
- STDs: categórica
- STDs.number: cuantitativa discreta
- STDs.condylomatosis: categórica
- STDs.cervical.condylomatosis: categórica
- STDs.vaginal.condylomatosis: categórica
- STDs.vulvo.perineal.condylomatosis: categórica
- STDs.syphilis: categórica
- STDs.pelvic.inflammatory.disease: categórica
- STDs.genital.herpes: categórica
- STDs.molluscum.contagiosum: categórica
- STDs.AIDS. categórica

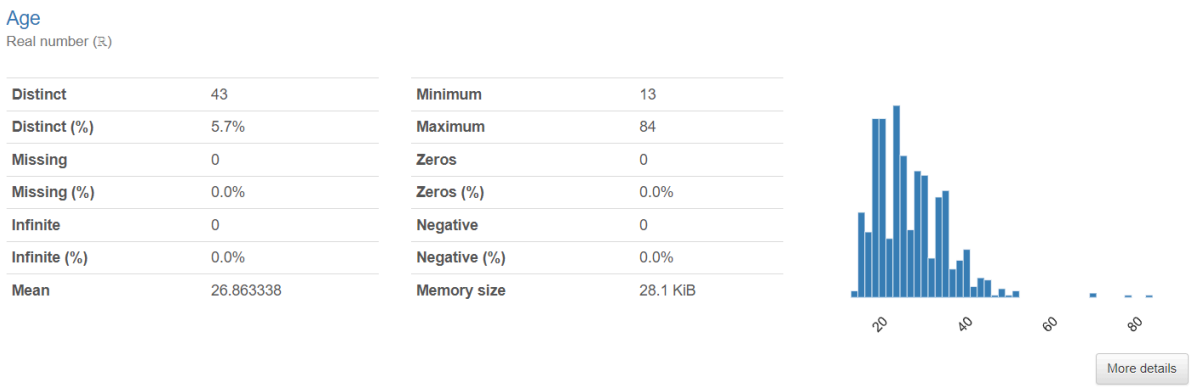
- STDs.HIV: categórica
- STDs.Hepatitis.B: categórica
- STDs.HPV: categórica
- STDs.Number.of.diagnosis: cuantitativa discreta
- STDs.Times.since.first.diagnosis: cuantitativa discreta
- STDs.Times.since.last.diagnosis: cuantitativa discreta
- Dx.Cancer: categórica
- Dx.CIN: categórica
- Dx.HPV: categórica
- Dx: categórica
- Hinselmann: categórica
- Schiller: categórica
- Citology: categórica
- Biopsy: categórica

Procesamiento previo

Para gestionar los valores nulos "?" se realizaron las siguientes acciones

- Eliminación de filas con valor ? en Number en Sexual Partners, puesto que el desconocimiento de su valores supone desconocer los valores First sexual intercourse, Num of pregnancies.
- Eliminación de filas con valor ? en Smokes y sus columnas derivadas, debido a su corta cantidad en proporción del total.
- Sustitución de valores ? por 0 en Hormonal Contraceptives y sus columnas derivadas, debido a su moderada cantidad.
- Sustitución de valores ? por 0 en IUD, y sus columnas derivadas, debido a su moderada cantidad.
- Sustitución de valores ? por 0 en STDs, y sus columnas derivadas, debido a su moderada cantidad.
- Eliminación de columnas 'STDs:Time since first diagnosis' y 'STDs:Time since last diagnosis', debido a la alta cantidad de valores ? y al poco impacto sobre el análisis.

Gráficos exploratorios relevantes tras procesamiento previo

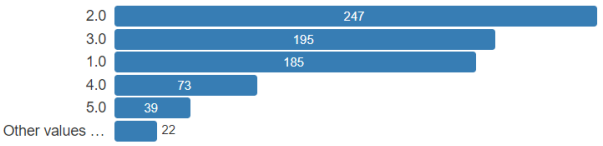


Age histogram

Number of sexual partners

Categorical

Distinct	10
Distinct (%)	1.3%
Missing	0
Missing (%)	0.0%
Memory size	60.8 KiB

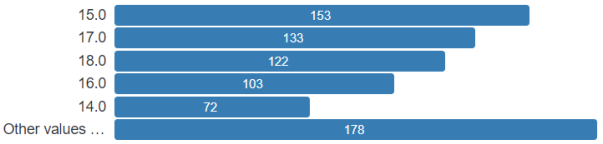


More details

First sexual intercourse

Categorical

Distinct	21
Distinct (%)	2.8%
Missing	0
Missing (%)	0.0%
Memory size	61.5 KiB



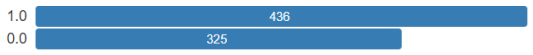
More details

Number of sexual partners and First sexual intercourse frequencies

Hormonal Contraceptives

Categorical

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	60.8 KiB



More details

Hormonal contraceptives frequencies

IUD

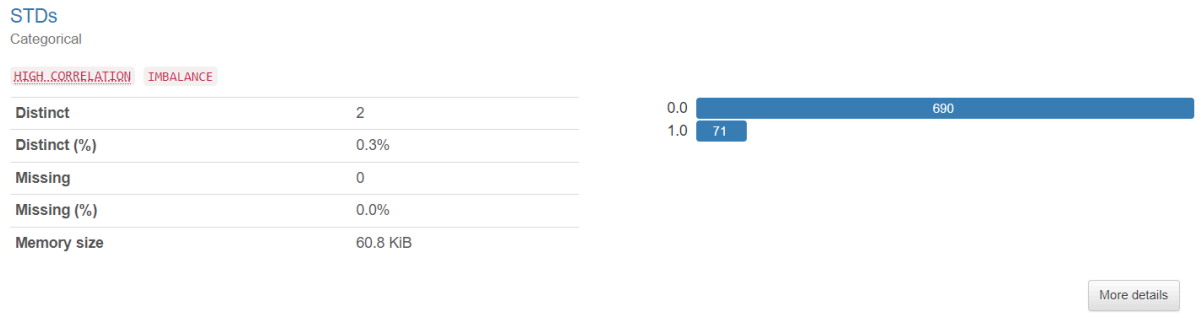
Categorical

HIGH_CORRELATION IMBALANCE

Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	60.8 KiB

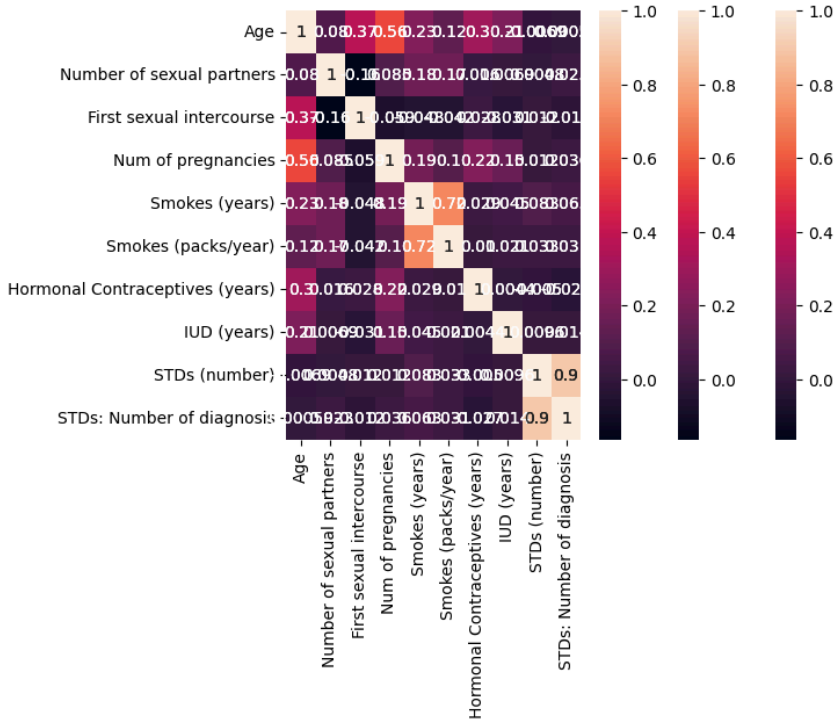
More details

IUD frequencies



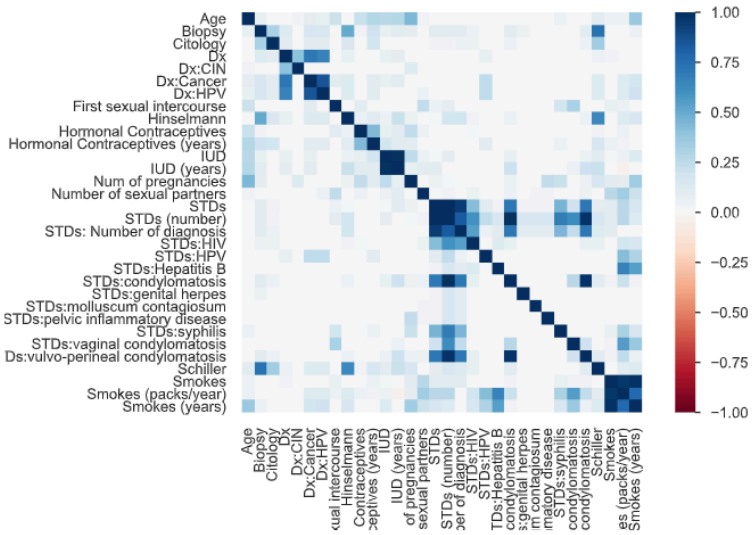
STDs frequencies

Correlación



Matriz de correlación de variables cuantitativas

Se observa una alta correlación entre las variables Smokes (years), Smokes (packs/year) y STDs (number), STDs: Number of diagnosis. Es posible inducir que fumar durante más años conlleva fumar más paquetes por año, así como un mayor número de ETS se refleja en un mayor número de diagnósticos. También se observan moderadas correlaciones negativas entre Age, First sexual intercourse y Age, Num of pregnancies. Esto sugiere que las personas de mayor edad tuvieron su primer encuentro sexual con menor edad. Así mismo, se induce que las mujeres mayores tienden a tener menos embarazos.

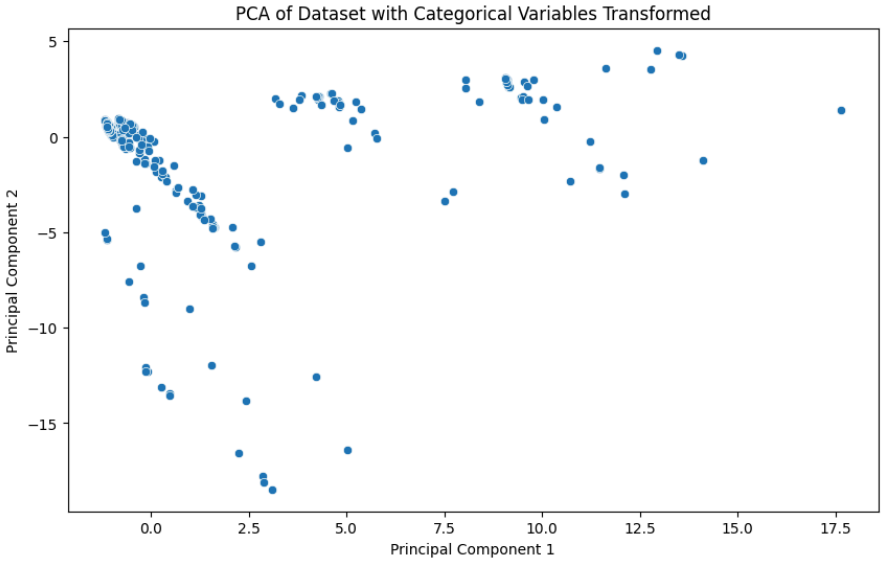


Matriz de correlación de todas las variables

Este mapa de calor, en el que se considera el valor numérico de las categorías sí/no de ciertas variables, refuerza las observaciones anteriores, y añade que existe correlación entre Dx:Cancer y Dx:CIN, lo cual indica que ciertas tipos de cáncer pueden estar relacionados con infecciones por HPV. Hormonal Contraceptive (years) y IUD (years) indican que las personas que utilizan anticonceptivos hormonales durante más tiempo, tienden a utilizar DIU durante largos periodos de tiempo.

Transformación de variables categóricas para PCA

Dada la naturaleza binaria de las variables categóricas del conjunto de datos, es posible hacer transformaciones para incluirlas en un análisis de componentes principales (PCA).



PCA del conjunto de datos utilizando las variables categóricas transformadas

Los dos primeros componentes principales explican solo el 25% de la varianza total (14% el primero y 11% el segundo). Este valor es muy bajo, indicando que los componentes principales no están capturando mucha información relevante de los datos originales. La gráfica muestra una dispersión considerable sin una estructura clara, sugiriendo que las variables categóricas transformadas no aportan una agrupación significativa o información adicional útil para el análisis.

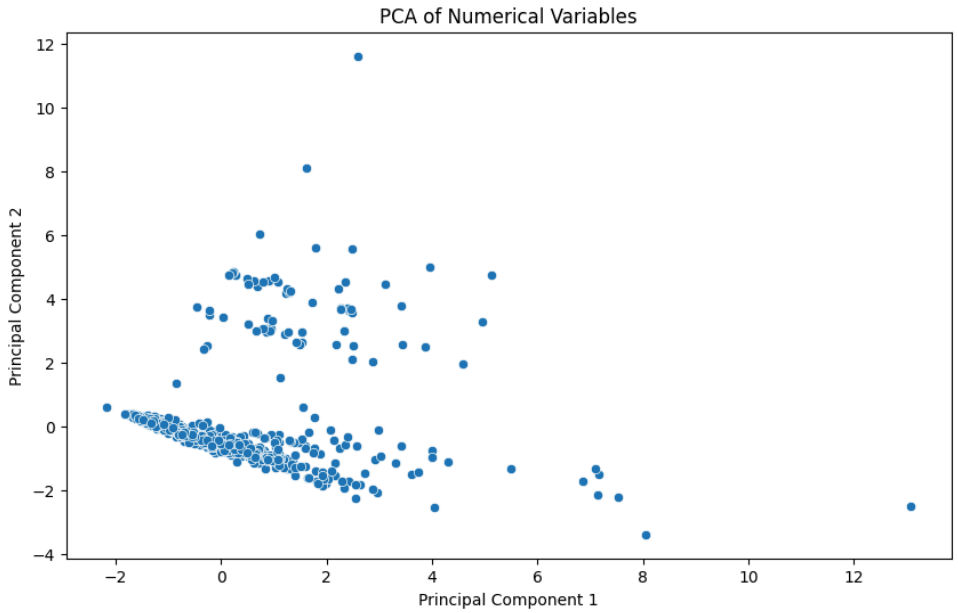
Dado que la varianza explicada es muy baja y la gráfica no muestra una estructura clara, no vale la pena incluir las variables categóricas transformadas en el PCA en este caso.

Índice de KMO y test de esfericidad de Bartlett

Variable	Valor
Índice KMO	0.49
Test de enfericidad de Bartlett	Chi cuadrado: 2568.86, p-valor: 0.00

Los valores sugieren que existe correlación entre las variables, por lo que es conveniente realizar PCA.

PCA



PCA de variables numéricas

Principal Component 1:
Carga Alta: Age (0.45), Smokes (years) (0.48), Smokes (packs/year) (0.43), Num of pregnancies (0.41). El primer componente parece estar influenciado principalmente por la edad, el historial de fumar (años y paquetes por año) y el número de embarazos. Este componente puede interpretarse

como una medida de experiencia de vida y comportamientos de salud relacionados con la edad y el tabaquismo.

Principal Component 2:
Carga Alta: STDs (number) (0.66), STDs: Number of diagnosis (0.66). El segundo componente está altamente influenciado por el número de enfermedades de transmisión sexual y el número de diagnósticos de ETS. Este componente puede interpretarse como una medida de la historia de ETS.

Varianza explicada por el primer componente: 0.22 (22%), Varianza explicada por el segundo componente: 0.19 (19%). Los dos primeros componentes principales explican el 41% de la varianza total. Esto sugiere que estos componentes capturan una cantidad moderada de la variabilidad en los datos, pero no una mayoría significativa.

Reglas de asociación

Reglas Generadas con Diferentes Niveles de Confianza y Soporte
Reglas con Soporte Mínimo de 0.2 y Confianza Mínima de 0.6:
Se generaron un total de 1154 reglas. Algunas reglas comunes incluyen asociaciones entre diferentes tipos de diagnósticos y enfermedades de transmisión sexual (ETS).

Reglas con Soporte Mínimo de 0.1 y Confianza Mínima de 0.6:
Se generaron más reglas, dado el menor umbral de soporte. Las asociaciones son similares a las anteriores, pero incluyen más combinaciones debido al menor umbral.

Reglas con Soporte Mínimo de 0.1 y Confianza Mínima de 0.7:
La mayor confianza mínima resultó en un menor número de reglas, pero con asociaciones más fuertes. Las reglas se centran en variables relacionadas con ETS y diagnósticos específicos.

Identificación de Variables Muy Frecuentes:
Las variables más frecuentes en las reglas de asociación incluyen:
STDs:genital herpes_0.0
STDs:AIDS_0.0
STDs:Hepatitis B_0.0
Dx:CIN_0
Estas variables dominan muchas de las reglas generadas, lo que puede ocultar otras relaciones interesantes.

Conclusiones

Respecto al análisis exploratorio, el PCA y las reglas de asociación, es posible concluir:

- El conjunto de datos inicial contiene una amplia cantidad de valores nulos “?”. También, se determinó correlación entre variables del conjunto.
- El test de Bartlett determinó la viabilidad de realizar PCA. Los componentes principales obtenidos logran explicar parte de la varianza. Sin embargo, estos valores no son totalmente satisfactorios.
- Las principales reglas de asociación obtenidas se centraron en las variables de ETS y los diagnósticos específicos.