Universidad del Valle de Guatemala

Data Science

Sección 20

Laboratorio 1 - Avances

Brian Carrillo - 21108

Carlos López - 21666

Guatemala, 18 de julio del 2024

Informe de análisis exploratorio

Variables del conjunto de datos y sus tipos

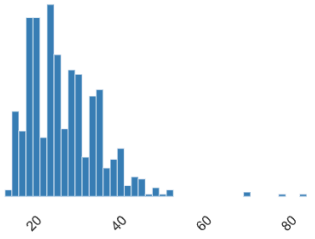
- Age: cuantitativa discreta
- Number.of.sexual.partners: cuantitativa discreta
- First.sexual.intercourse: cuantitativa discreta
- Num.of.pregnancies: cuantitativa discreta
- Smokes: categórica
- Smokes.years: cuantitativa discreta
- Smokes.packs.per.year: cuantitativa discreta
- Hormonal.Contraceptives: 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

Gráficos exploratorios relevantes

Age

Real number (ℝ)

Distinct	43	**Minimum**	13
Distinct (%)	5.7%	**Maximum**	84
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	26.863338	**Memory size**	28.1 KiB

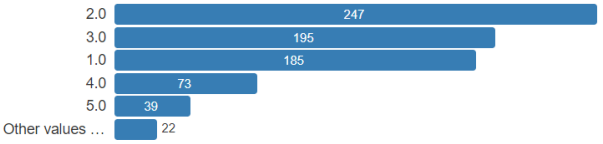


More details

Number of sexual partners

Categorical

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

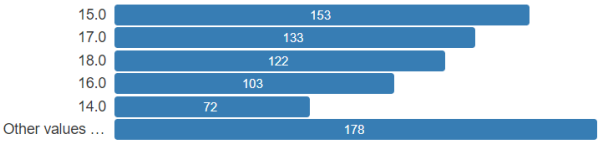


More details

First sexual intercourse

Categorical

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



More details

Hormonal Contraceptives

Categorical

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

More details

IUD

Categorical

HIGH_CORRELATION IMBALANCE

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

More details

STDs

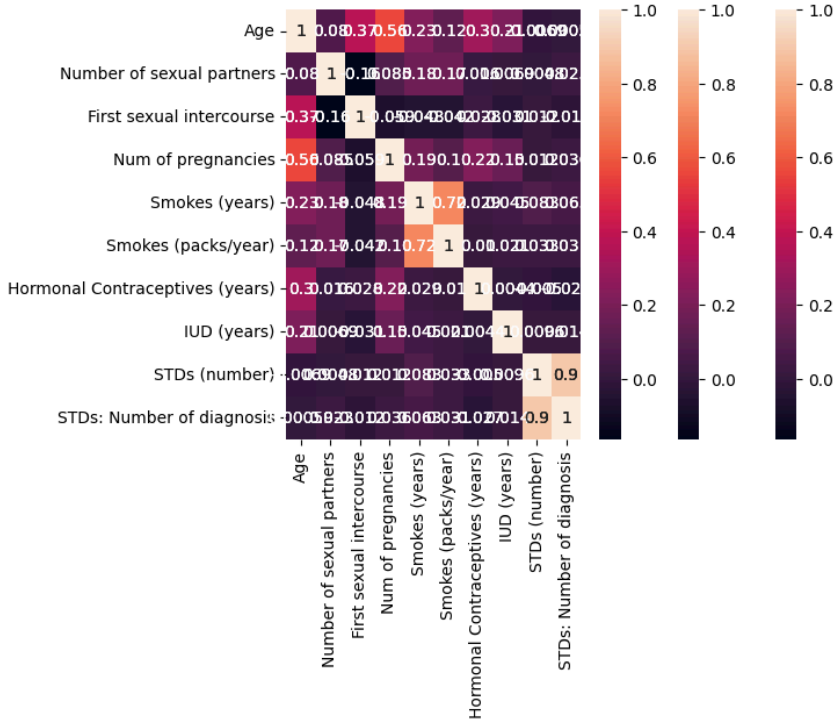
Categorical

HIGH CORRELATION IMBALANCE

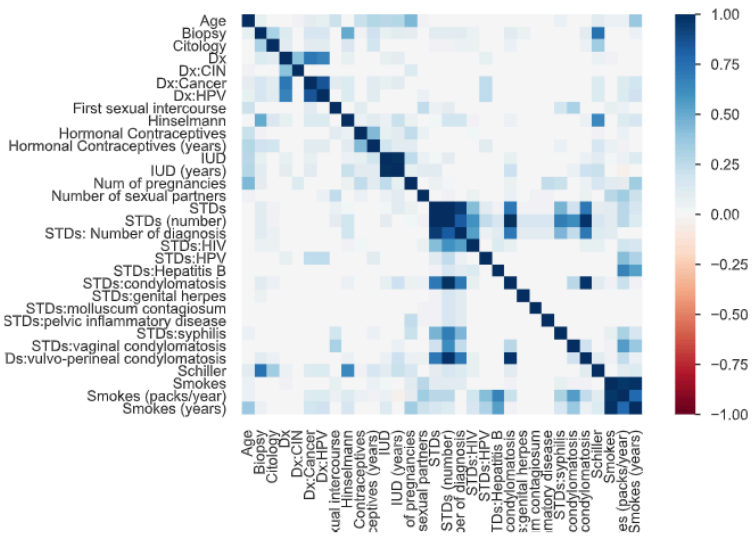
Distinct	2
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	60.8 KiB

More details

Correlación



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.



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).