

Consumo de Alcohol en Estudiantes

CIENCIA DE DATOS - TERCERA GENERACIÓN



Contenido

- Antecedentes
- Objetivo
- Desarrollo
 - Presentación del dataset
 - Diccionario de Datos
 - Variable Objetivo
 - Selección y tratamiento de los datos
 - Histogramas
- Presentación del modelo
 - Comparación de los modelos
 - Resultado
- Conclusión
- Bibliografía

Antecedentes

De acuerdo a la OMS, el consumo de alcohol ocupa el tercer lugar mundial entre los factores de riesgo de enfermedades y de discapacidad

Aproximadamente 1 de cada 10 jóvenes mueren a causa del alcohol

Los problemas de salud son las principales consecuencias del consumo de alcohol, relacionado como la causa de 60 tipos de enfermedades tanto agudas como crónicas

Canva

Antecedentes

El consumo de alcohol se asocia con consecuencias psicosociales generalizadas como la violencia, el abandono, el maltrato y el ausentismo en el lugar de trabajo, entre otros

Los resultados de investigaciones en adolescentes muestran el daño neuronal secundario por consumo de alcohol en edades tempranas, presentando alteraciones de la conducta, de la memoria y de los procesos relacionados con el aprendizaje



Antecedentes

CONSECUENCIAS

- Los accidentes de tráfico, suicidios y homicidios.
- Actitud negativa, bajo rendimiento académico, problemas de disciplina, que conllevan al abandono total de la escuela.
- Movimientos motores menos coordinados, reflejos lentos, afeción del control de los músculos del habla y la actividad de los ojos.
- Conflictos familiares, distanciamiento y hostilidad.



Objetivo

Diseñar un modelo que indique los factores que influyen en el consumo de alcohol en jóvenes de 15 a 22 años, durante los fines de semana.



Alcoholismo en Portugal

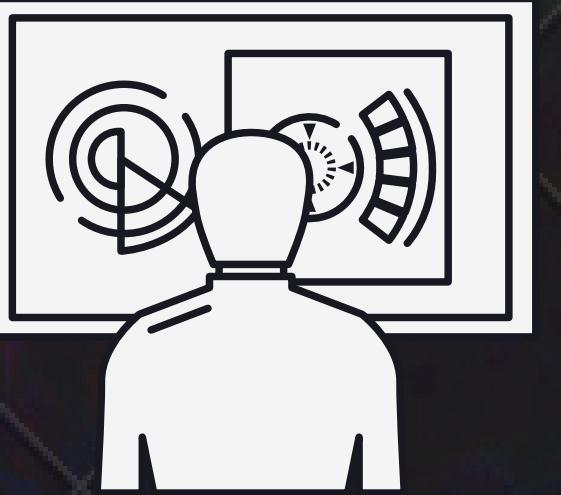
De acuerdo a la OMS, el consumo de alcohol per capita fue el más alto en toda europa con 10.6 litros

La mayor parte de las infraccionesen Portugal son causadas por el alcohol o el consumo de drogas

70% de los jovenes entre los 16 y 25 años toman alcohol regularmente

Desarrollo

Presentación del Dataset



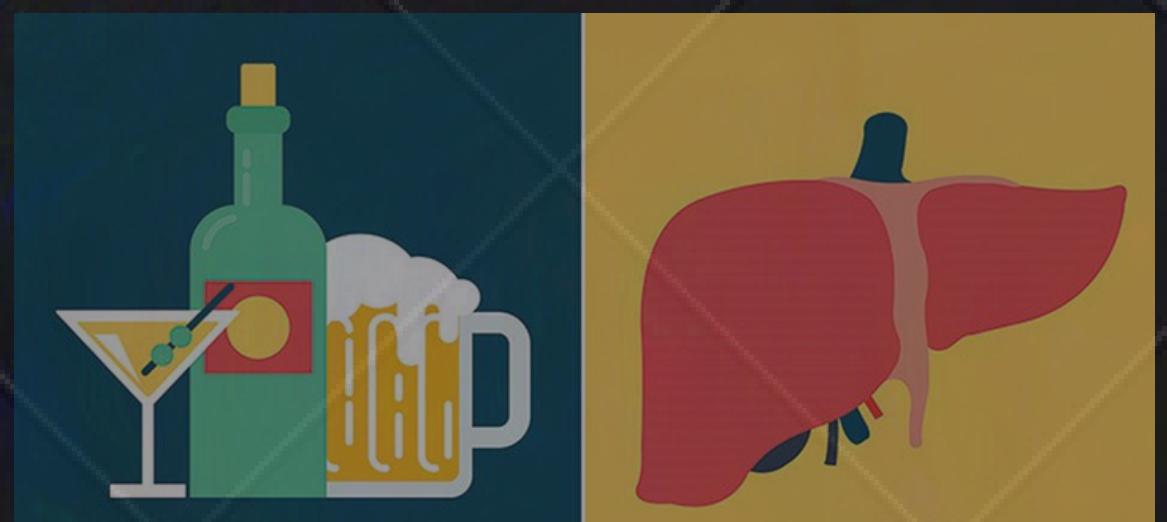
La información se obtuvo del dataset público *Student Alcohol Consumption* disponible en [Kaggle](#), cuyos datos se recabaron en una encuesta estudiantil realizada en Portugal, a alumnos pertenecientes a los cursos de matemáticas y portugués a nivel bachillerato. Este dataset brinda información tal como género, edad, horas de estudio, relación familiar y amorosa, entre otros.



Student Alcohol Consumption

Social, gender and study data from secondary school students

[kaggle.com](https://www.kaggle.com)



Diccionario de Datos

- school - escuela del estudiante (binario: 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira)
- sex - sexo del estudiante (binario: 'F' - femenino o 'M' - masculino)
- age - edad del estudiante (numérico: de 15 a 22)
- address - tipo de dirección del hogar del estudiante (binario: 'U' - urbano o 'R' - rural)
- famsize - tamaño de la familia (binario: 'LE3' - menor o igual a 3 o 'GT3' - mayor que 3)
- Pstatus - estado de convivencia de los padres (binario: 'T' - viviendo juntos o 'A' - separados)
- Medu - educación de la madre (numérico: 0 - ninguno, 1 - educación primaria (4 ° grado), 2 - 5 ° a 9 ° grado, 3 - educación secundaria o 4 - educación superior)
- Fedu - educación del padre (numérico: 0 - ninguno, 1 - educación primaria (4° grado), 2 - 5° a 9° grado, 3 - educación secundaria o 4 - educación superior)



Diccionario de Datos

- Mjob - trabajo de la madre (nominal: 'maestra', relacionado con el cuidado de la 'salud', 'servicios' civiles (por ejemplo, administrativo o policial), 'en_casa' u 'otro')
- Fjob - trabajo del padre (nominal: 'maestro', relacionado con el cuidado de la salud, 'servicios' civiles (por ejemplo, administrativo o policial), 'en_home' u 'otro')
- reason - motivo para elegir esta escuela (nominal: cerca de 'casa', 'reputación' de la escuela, preferencia de 'curso' u 'otro')
- guardian - tutor del estudiante (nominal: 'madre', 'padre' u 'otro')
- traveltime - tiempo de viaje de la casa a la escuela (numérico: 1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hora, o 4 -> 1 hora)
- studytime - tiempo de estudio semanal (numérico: 1 - <2 horas, 2 - 2 a 5 horas, 3 - 5 a 10 horas, o 4 -> 10 horas)
- failures - número de fallos de clase anteriores (numérico: n si $1 \leq n < 3$, en caso contrario 4)



Diccionario de Datos

- schoolup - apoyo educativo adicional (binario: sí o no)
- famsup - apoyo educativo familiar (binario: sí o no)
- paid - clases pagas adicionales dentro de la asignatura del curso (matemáticas o portugués) (binario: sí o no)
- activities - actividades extracurriculares (binario: sí o no)
- nursery - asistió a la guardería (binario: sí o no)
- higher - quiere cursar estudios superiores (binario: sí o no)
- Internet - acceso a Internet en casa (binario: sí o no)
- romantic - con una relación romántica (binario: sí o no)
- famrel - calidad de las relaciones familiares (numérico: de 1 - muy mala a 5 - excelente)
- freetime- tiempo libre después de la escuela (numérico: de 1 - muy bajo a 5 - muy alto)



Diccionario de Datos

- goout - salir con amigos (numérico: de 1 - muy bajo a 5 - muy alto)
- Dalc - consumo de alcohol en la jornada laboral (numérico: de 1 - muy bajo a 5 - muy alto)
- Walc - consumo de alcohol durante el fin de semana (numérico: de 1 - muy bajo a 5 - muy alto)
- health - estado de salud actual (numérico: de 1 - muy malo a 5 - muy bueno)
- absences - número de ausencias escolares (numérico: de 0 a 93)
- Estas calificaciones están relacionadas con la asignatura del curso, Matemáticas o Portugués:
 - G1 - grado del primer período (numérico: de 0 a 20)
 - G2 - grado del segundo período (numérico: de 0 a 20)
 - G3 - calificación final (numérica: de 0 a 20, objetivo de rendimiento)



Variable Objetivo

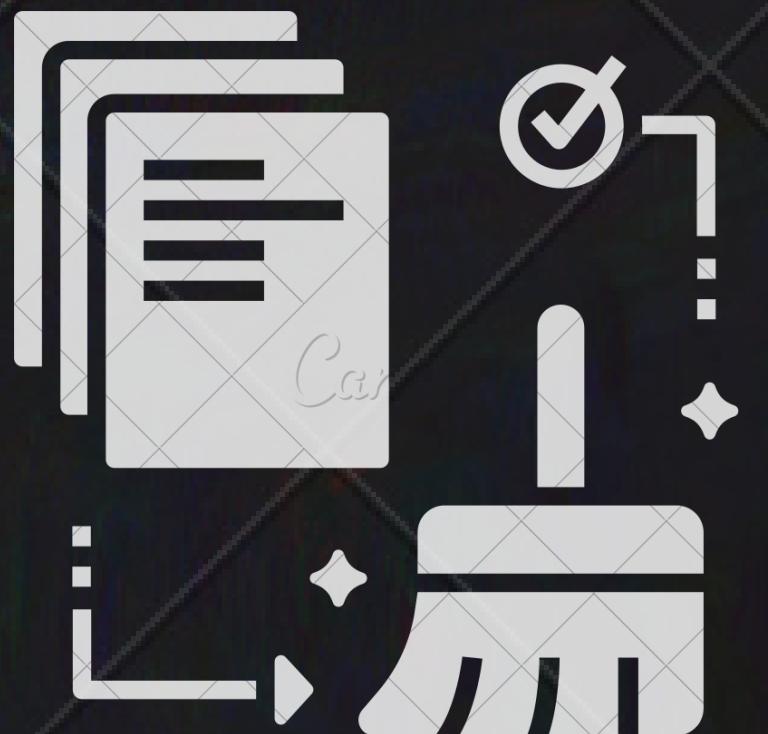
Dado que se busca predecir el consumo de alcohol en jóvenes durante los fines de semana, la variable objetivo de tipo continua es "Walc", que representa el consumo de alcohol durante el fin de semana (numérico: de 1 - muy bajo a 5 - muy alto).



Selección y limpieza de datos (EDA)

Para nuestro análisis requerimos filtrar los datos, empezando por eliminar los registros duplicados, posteriormente se procede a hacer la selección de variables que nos serán útiles a la hora de realizar nuestro modelo, estas variables deben proporcionarnos un mejor ajuste para que el modelo sea óptimo y acertado.

De esta manera, del diccionario de datos (33 atributos) tomamos solo 24 de esta manera procedemos a realizar la clasificación de variables.



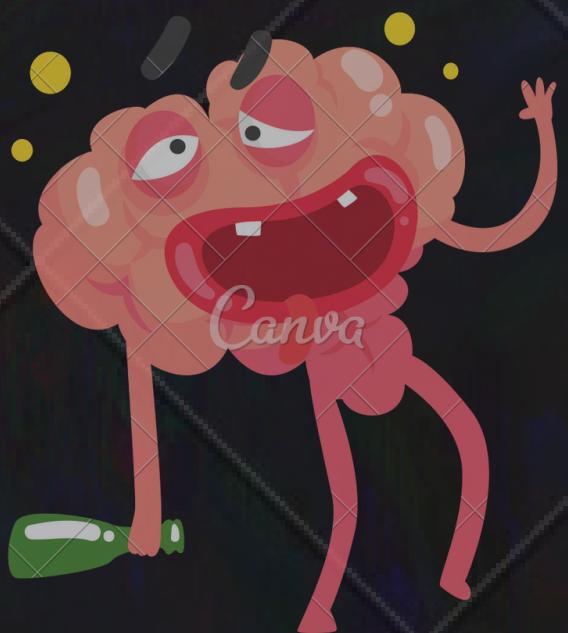
Variables Discretas

Tomamos como discretas 13 variables, para saber su comportamiento sacamos la frecuencia absoluta, la frecuencia relativa y la frecuencia acumulada

Feature: school				
	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
GP	729	73.34%	729	73.34%
MS	265	26.66%	994	100.00%

Escuela

La encuesta se realizó en dos escuelas, Gabriel Pereira y Mousinho da Silveira, la mayoría de los alumnos asisten a la secundaria Gabriel Pereira.



Variables Discretas

Feature: address				
	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
U	719	72.33%	719	72.33%
R	275	27.67%	994	100.00%



Vivienda

Esta variable nos dice el tipo de asentamiento en el que viven los estudiantes, si bien una gran parte vive en espacios urbanos, muchos entrevistados habitan en lugares rurales.

Feature: famsize				
	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
GT3	704	70.82%	704	70.82%
LE3	290	29.18%	994	100.00%

Integrantes por familia

Esto da a entender que la mayoría de los estudiantes tiene una familia numerosa y de aquí se puede sacar otro tipo de información como qué tanto influye el ámbito familiar en el consumo de alcohol.

Variables Discretas

Feature: Pstatus

	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
1	878	88.33%	878	88.33%
0	116	11.67%	994	100.00%

Relación de los padres

Solo el 11.67% de los entrevistados tiene padres separados.

Feature: famsup

	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
1	616	61.97%	616	61.97%
0	378	38.03%	994	100.00%

Apoyo escolar por parte de la familia

La mayoría cuenta con el.



Variables Discretas

Feature: romantic				
	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
0	642	64.59%	642	64.59%
1	352	35.41%	994	100.00%

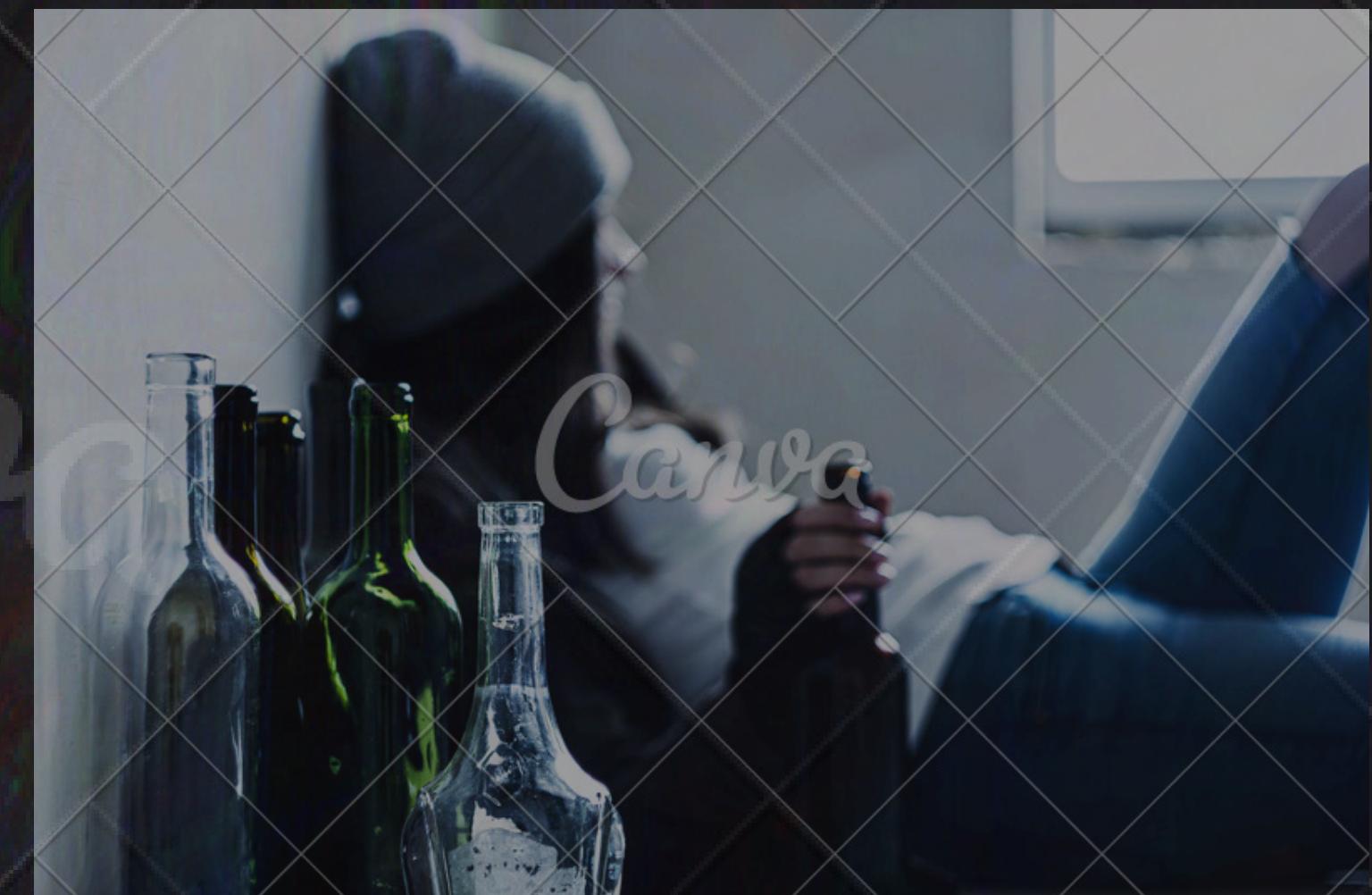
Se encuentra en una relación amorosa

En general, los alumnos de estas escuelas no cuentan con una.

Feature: internet				
	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
yes	787	79.18%	787	79.18%
no	207	20.82%	994	100.00%

Acceso a internet desde su hogar

Casi todos tienen acceso a internet.



Variables Discretas

Feature: activities				
	Absolute frequency	Relative frequency	Accumulated frequency	Accumulated %
no	501	50.40%	501	50.40%
yes	493	49.60%	994	100.00%

Actividades extra curriculares

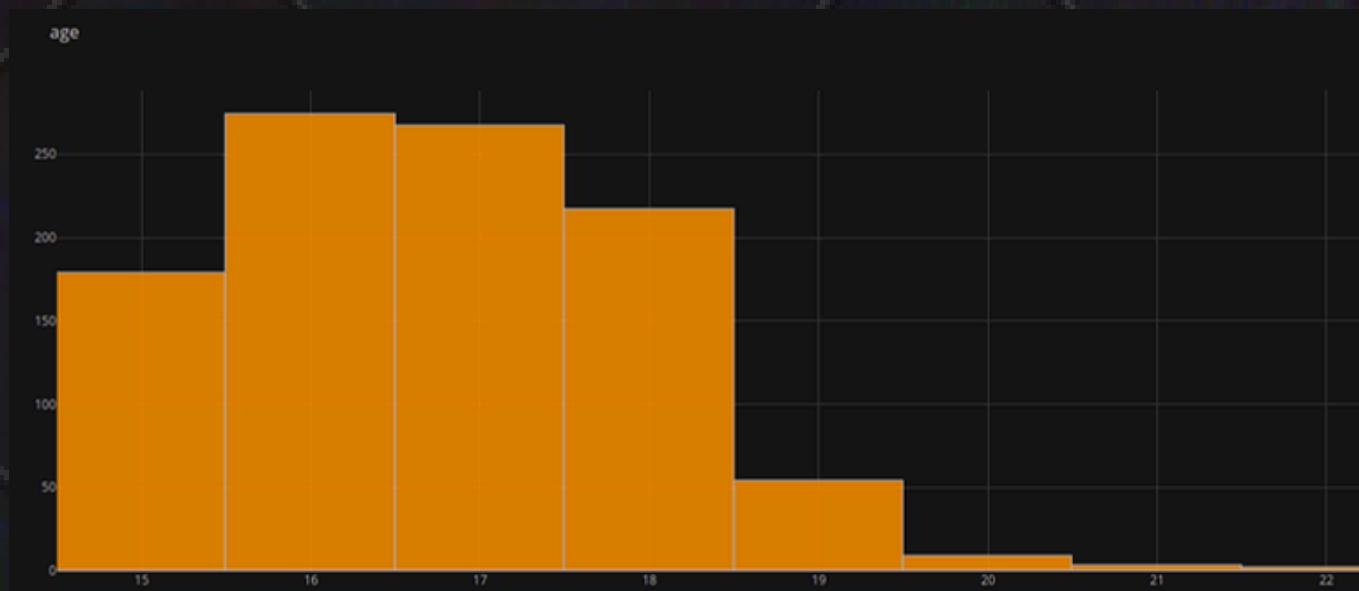
Poco más de la mitad del alumnado tiene alguna actividad aparte de la escuela.

Se verifican los missings o daltos faltantes de estas variables para descartar aquellas que tengan demasiados, en este caso no fue necesario.
Ahora se hace el análisis de las variables continuas.

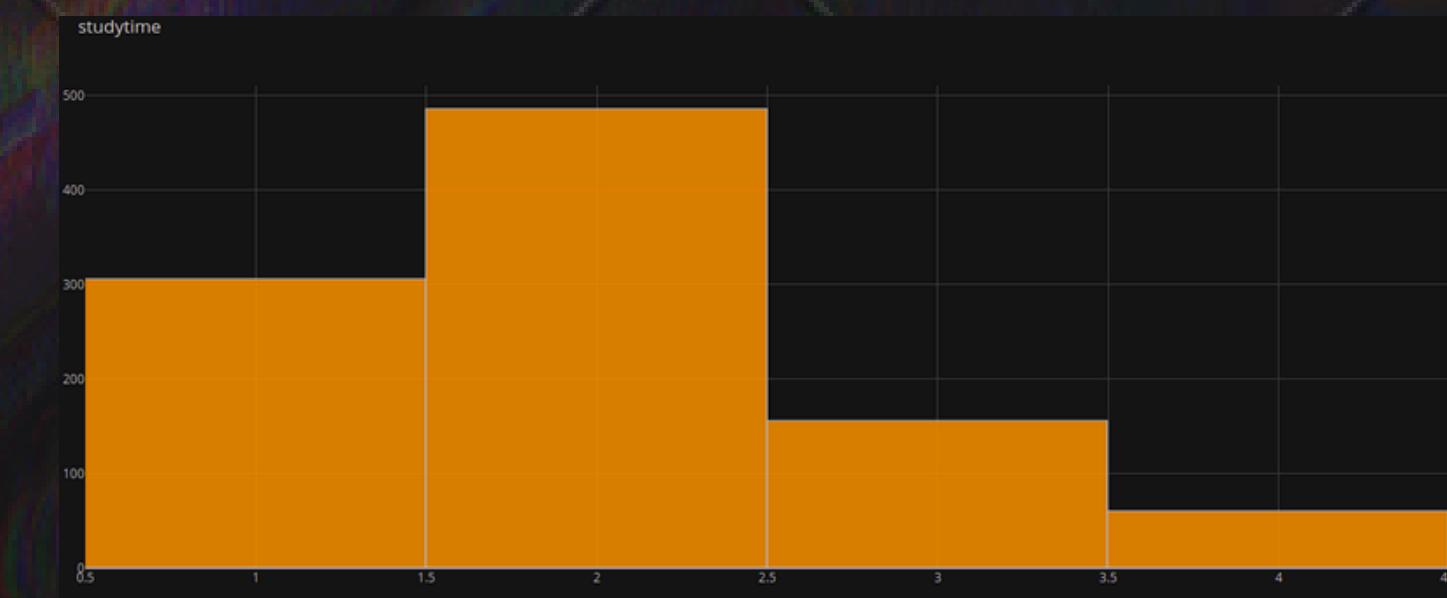


Variables Continuas

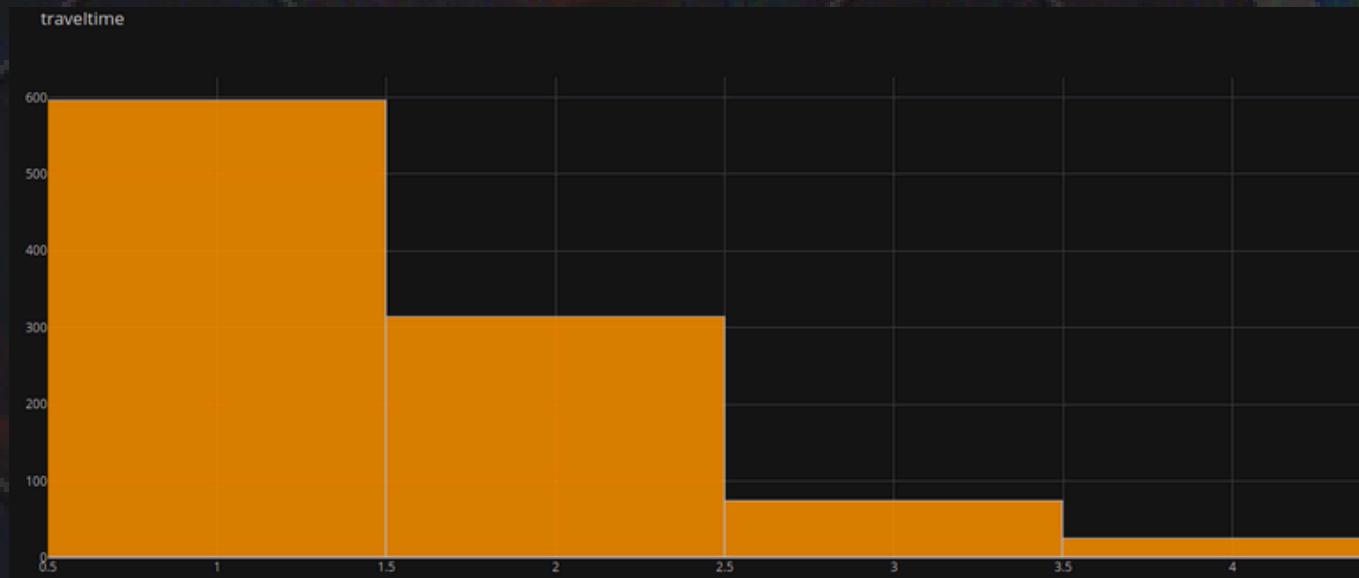
Edad



Studytime



Traveltime

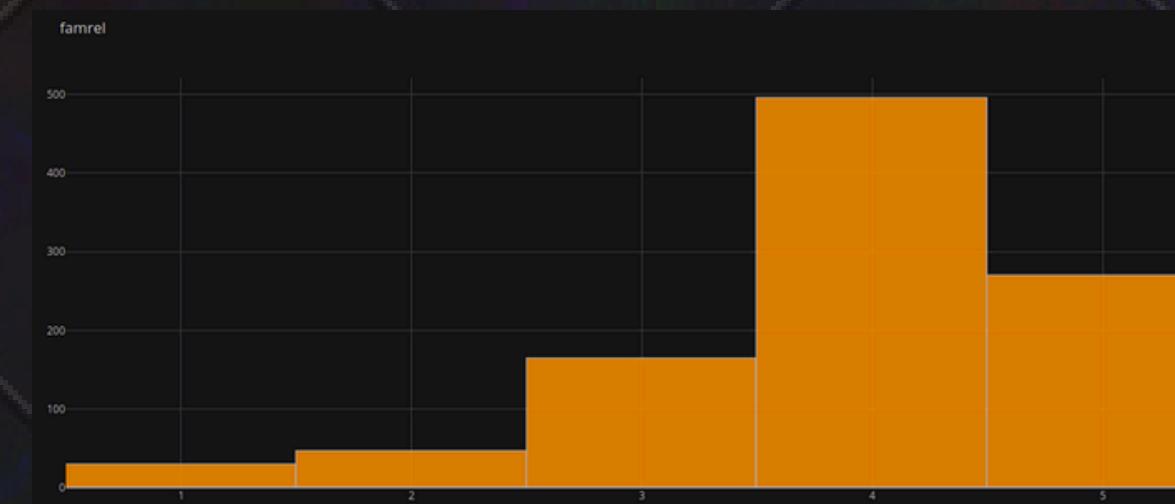


failures

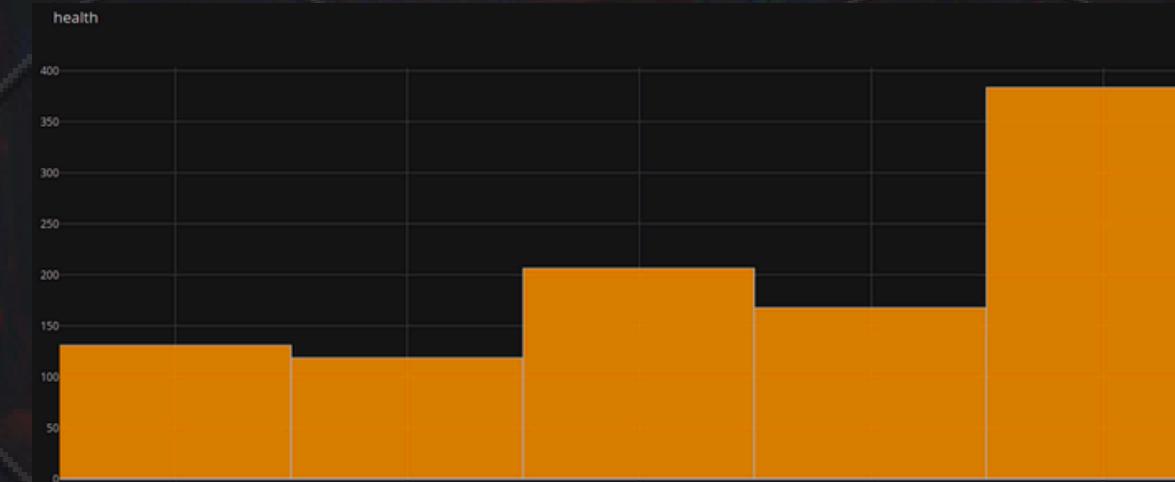


Variables Continuas

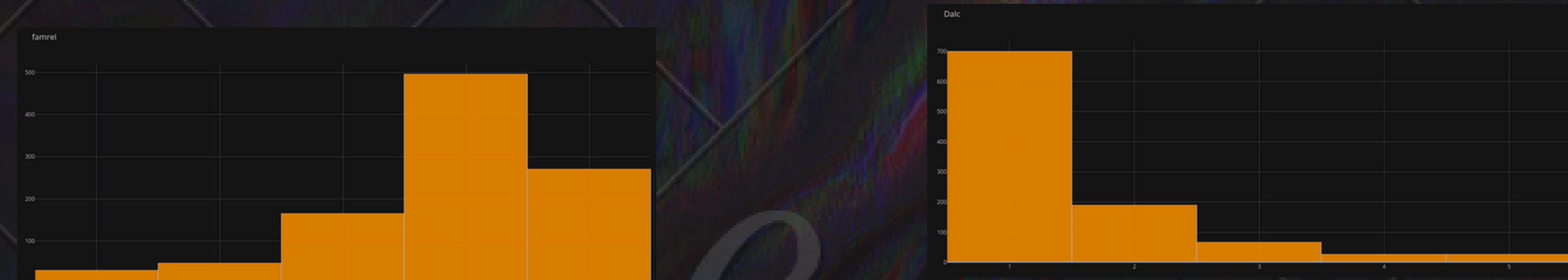
famrel



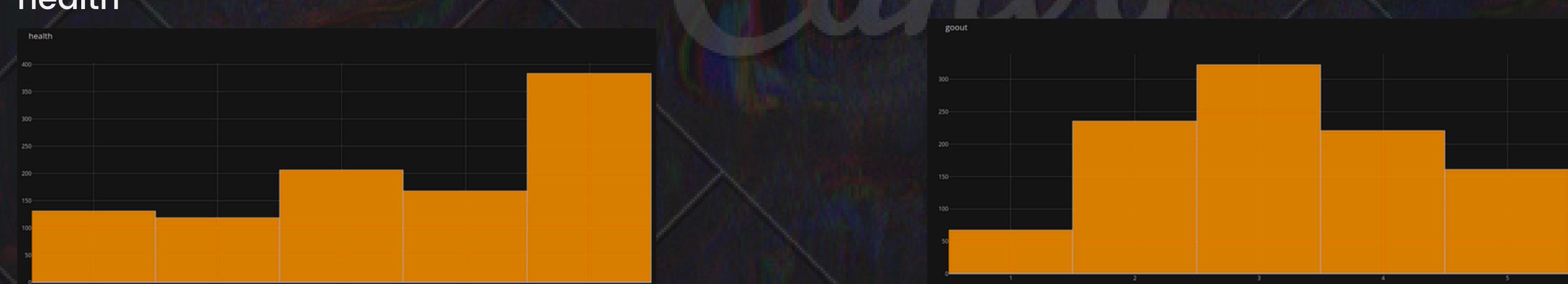
health



Dalc



goout



Variables Continuas

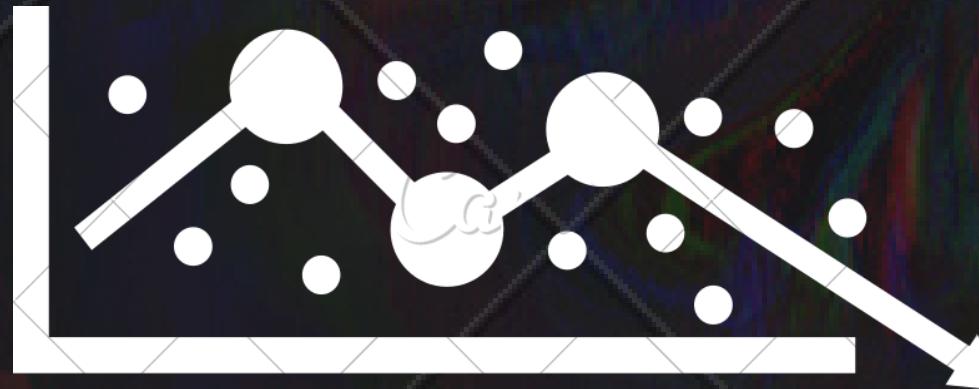
Nos deshacemos de los outlayers continuas usando el cuantil 99 y con un diccionario transformamos las variables binarias discretas que nos interesan a continuas



	romantic	Pstatus	sex	famsup
0	0	0	1	0
1	0	1	1	1
2	0	1	1	0
3	1	1	1	1
4	0	1	1	1

Presentación del modelo

Escogemos un modelo de regresión pues nuestra variable objetivo es de índole continuo, queremos describir cómo el género, las relaciones interpersonales, la escuela, el tiempo de estudio, entre otros influyen en la cantidad de alcohol que ingiere un joven en un fin de semana.



Presentación del modelo

Hacemos el ajuste del modelo tomando a nuestras variables continuas como X y a nuestra variable objetivo como y.

Creamos las variables de entrenamiento y de prueba y las normalizamos con Min-Max Scaler



Regresión Logística

Primero probamos con regresión logística dándole 10000 como máximo de iteraciones .

Hacemos un cross validation de 4 pliegues y obtenemos un promedio de aprox. 0.203836 y una desviación estándar de 0.032658, en sí no es un mal modelo, al hacer el score obtenernos una precisión del 0.502008 aprox. de aquí sacamos la siguiente tabla que indica cuáles factores influyen más, para la regresión logística es el tiempo de estudio.

	0	1
7	Dalc	-2.15314
6	goout	-0.66698
11	Pstatus	-0.39960
8	health	-0.20962
0	age	-0.05401
9	absences	-0.03744
1	traveltime	-0.00840
12	sex	0.01048
3	failures	0.18108
10	romantic	0.19952
4	famrel	0.21112
13	famsup	0.22336
5	freetime	0.22701
2	studytime	0.41357

Regresión Ridge

En el caso de Ridge para alpha=0, con cross validation de 4 pliegues obtuvimos una media de aprox. 0.4702312 y una desviación estándar de 0.0434727, en el score tenemos una precisión del 0.4941935 aprox. y con el intercepto tuvimos 0.85506, dándonos esta tabla donde la calidad de la relación familiar es lo que menos influye.

	0	1
4	famrel	-0.11396
2	studytime	-0.10762
5	freetime	-0.04885
3	failures	-0.04359
13	famsup	-0.04090
10	romantic	-0.02690
12	sex	0.00000
0	age	0.01016
1	traveltime	0.01640
11	Pstatus	0.03295
8	health	0.08720
9	absences	0.10494
6	goout	0.30050
7	Dalc	0.72329

Regresión Lasso

En el caso de Lasso tomamos alpha=0.1, cross validation de 4 pliegues y la media es de 0.47578, con desviación estándar de 0.046214 e intercepto de 0.134216, para esta regresión vemos que lo que menos influye es el tiempo de estudio dedicado y llama la atención que hay bastantes variables que aparentemente no inflyen en nada.

	0	1
2	studytime	-0.02255
13	famsup	-0.00259
0	age	0.00000
1	traveltime	0.00000
3	failures	0.00000
4	famrel	-0.00000
5	freetime	0.00000
9	absences	0.00000
10	romantic	-0.00000
11	Pstatus	0.00000
12	sex	0.00000
8	health	0.00011
6	goout	0.19125
7	Dalc	0.61481

Regresión Red Elástica

En este caso elegimos alpha=0.001 y ratio=0.07, se hizo el ajuste y con cross validation de 4 pliegues se obtuvo una media y una desviación estándar de 0.4709033 y 0.0421295 respectivamente, además de un intercepto=0.14487131, esta empieza a tomar forma y notamos que lo único que no afecta es el género.

	0	1
4	famrel	-0.11124
2	studytime	-0.10694
5	freetime	-0.04531
13	famsup	-0.04054
3	failures	-0.03885
10	romantic	-0.02600
12	sex	0.00000
0	age	0.00795
1	traveltime	0.01647
11	Pstatus	0.03189
8	health	0.08594
9	absences	0.10254
6	goout	0.29755
7	Dalc	0.71110

Regresión Lineal

Al igual que en los casos anteriores se utilizó un cross validation de 4 pliegues obteniendo una media de 0.4702312, una desviación estándar de 0.043473, un intercepto de 0.144112 y la siguiente tabla de valores.

	0	1
4	famrel	-0.11396
2	studytime	-0.10762
5	freetime	-0.04885
3	failures	-0.04359
13	famsup	-0.04090
10	romantic	-0.02690
12	sex	0.00000
0	age	0.01016
1	traveltime	0.01640
11	Pstatus	0.03295
8	health	0.08720
9	absences	0.10494
6	goout	0.30050
7	Dalc	0.72329

Regresión Bayesiana

Finalmente se utilizó regresión bayesiana y se consiguió un promedio de 0.470516, con desviación de 0.0414945 y un intercepto de 0.1445433, de esta tabla se sabe que el al igual que, a diferencia de la regresión logística, lo que más influye en el consumo de alcohol durante los fines de semana en los jóvenes es cuánto consumen entre semana.

	0	1
4	famrel	-0.11220
2	studytime	-0.10781
5	freetime	-0.04535
13	famsup	-0.04066
3	failures	-0.04056
10	romantic	-0.02634
12	sex	0.00000
0	age	0.01146
1	traveltime	0.01815
11	Pstatus	0.03255
8	health	0.08669
9	absences	0.10381
6	goout	0.29749
7	Dalc	0.70550

Comparación de Modelos

En resumen:

Regresión	Promedio	Desviación Estándar	Score	Intercepto
Logística	0.203835560	0.032658000	0.502008000	
Ridge	0.470231198	0.043472730	0.494193500	0.855059990
Lasso	0.475780379	0.046214039		0.134216610
Red Elástica	0.470903350	0.042129522	0.494029293	0.144871310
Lineal	0.470231198	0.043472738	0.494193545	0.144112150
Bayesiana	0.470515867	0.041494480	0.493598180	0.144543346

Resultado

Se toma como mejor modelo la regresión de Ridge y note que están escalados los valores, por lo cual son equiparables los atributos.

Como ya se venía estimando, los factores que más influyen en el aumento del consumo de alcohol en los jóvenes con edad entre 15 y 20 años es cuánto alcohol consumen a la semana, qué tan mala es la relación familiar y un poco qué tan seguido conviven con amigos, además notamos que el género no es un factor influyente.

	0	1
4	famrel	-0.11396
2	studytime	-0.10762
5	freetime	-0.04885
3	failures	-0.04359
13	famsup	-0.04090
10	romantic	-0.02690
12	sex	0.00000
0	age	0.01016
1	traveltime	0.01640
11	Pstatus	0.03295
8	health	0.08720
9	absences	0.10494
6	goout	0.30050
7	Dalc	0.72329

Resultado

Se toma como mejor modelo la regresión de Ridge y note que están escalados los valores, por lo cual son equiparables los atributos.

Como ya se venía estimando, los factores que más influyen en el aumento del consumo de alcohol en los jóvenes con edad entre 15 y 20 años es cuánto alcohol consumen a la semana, qué tan mala es la relación familiar y un poco qué tan seguido conviven con amigos, además notamos que el género no es un factor influyente.

	0	1
4	famrel	-0.11396
2	studytime	-0.10762
5	freetime	-0.04885
3	failures	-0.04359
13	famsup	-0.04090
10	romantic	-0.02690
12	sex	0.00000
0	age	0.01016
1	traveltime	0.01640
11	Pstatus	0.03295
8	health	0.08720
9	absences	0.10494
6	goout	0.30050
7	Dalc	0.72329

Conclusión

Estos datos fueron recabados de jóvenes habitantes de Portugal pero desafortunadamente, como lo indica el inicio de esta presentación, el consumo de alcohol es un factor de riesgo mundial, que ataca tanto a jóvenes adolescentes como a adultos mayores.

Información como la obtenida nos ayuda a entender qué comportamientos impactan de manera directa que orillan a las personas a caer en el alcoholismo y cómo evitarlos.



Conclusión

Alcoholismo en México

En México el 42.9% de los adolescentes de 12 a 17 años ha consumido alcohol alguna vez en la vida.

Por género en este rango de edad se encontró que el 17.4% de los hombres y el 11.6% de las mujeres ha consumido alcohol el último mes.

En México, el uso de alcohol es la cuarta causa de muerte de la población en el país (8.4%).

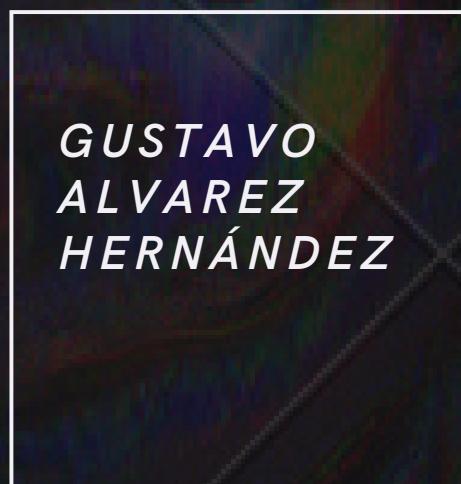
Bibliografía

- Ahumada-Cortez, Jesica Guadalupe, & Gámez-Medina, Mario Enrique, & Valdez-Montero, Carolina (2017). EL CONSUMO DE ALCOHOL COMO PROBLEMA DE SALUD PÚBLICA. *Ra Ximhai*, 13(2), 13-24.[fecha de Consulta 22 de Noviembre de 2021]. ISSN: 1665-0441. Disponible en: <https://www.redalyc.org/articulo.oa?id=46154510001>
- Marketersbyadlatina.com. (2017, 8 agosto). Portugal y España registraron el mayor consumo del alcohol en Iberoamérica - Marketers by Adlatina. 2021 Marketers by Adlatina. <http://www.marketersbyadlatina.com/articulo/2672-portugal-y-espana-registraron-el-mayor-consumo-del-alcohol-en-iberoamerica>

Colaboradores



DIEGO
GONZÁLEZ
MENDOZA



GUSTAVO
ALVAREZ
HERNÁNDEZ



HIRAM
PAULIN
ARISTA



FRANCISCO
HUERTA
LÓPEZ



MARIANO
ROBLES
GÓMEZ



ABIGAIL
JIMÉNEZ
MARTÍNEZ



GABRIEL
URRUTIA
LÓPEZ