

**DAVID SUÁREZ NIETO**

EDA: “STUDENT ALCOHOL CONSUMPTION”

**ÍNDICE:**

1. INTRODUCCIÓN
   1. Tema y fuente
   2. Problemática y objetivo
   3. Metodología inicial
2. EXPLORACIÓN DE LOS DATOS – METODOLOGÍA
3. LIMPIEZA DE LOS DATOS
4. ANÁLISIS DE LOS DATOS
   1. Estudio de la muestra
   2. Hipótesis y explicaciones
5. CONCLUSIONES Y POSIBLES SOLUCIONES A LA PROBLEMÁTICA
6. FORTALEZAS, LIMITACIONES, FUTURAS CONSIDERACIONES DEL EDA
   1. Fortalezas
   2. Limitaciones control de variables contaminadoras o extrañas, el investigador como variable contaminadora, la deseabilidad social
      1. Principales dificultades encontradas
   3. Consideraciones futuras
7. AGRADECIMIENTOS
8. ANEXO
9. **INTRODUCCIÓN**

Hoy en día, cada vez más estudiantes consumen una excesiva cantidad de alcohol con una frecuencia significativamente alta. Esto se puede traducir en un incipiente deterioro de sus funciones cognitivas que puede ocasionar dificultades de concentración y de toma de decisiones ocasionando un detrimento en su rendimiento escolar conllevando problemas familiares.

* 1. TEMA Y FUENTE

El presente análisis exploratorio de datos trata sobre el consumo de alcohol, sus variables precipitantes, predisponentes y consecuencias sobre una muestra inicial de 395 jóvenes de un colegio. Para ello, se ha trabajado con un dataset obtenido de la siguiente fuente: Kaggle: https://www.kaggle.com/uciml/student-alcohol-consumption

* 1. PROBLEMÁTICA Y OBJETIVO

Pertenecemos a la junta directiva del gabinete educativo de un colegio en el que ha habido creciente consumo de alcohol en los alumnos que ha ocasionado una serie de consecuencias traducidas en un descenso de las puntuaciones de las notas finales. Las familias de los estudiantes no han tardado en notar estar consecuencias y, ante esta situación, una junta extraordinaria ha sido convocada con el objeto de analizar el malestar de las familias.

* 1. METODOLOGÍA INICIAL

Tras reunirnos en la junta directiva, se ha tomado la decisión de realizar una encuesta a 395 estudiantes de diferentes cursos en la que se abordan una serie de preguntas que analizan numerosas variables que puedan explicar la conducta de consumo de los alumnos para buscar soluciones ante esta problemática.

La metodología a seguir ha sido: ex post facto (porque todas las variables han tenido lugar cuando se analiza la problemática), encuesta (ha sido la técnica empleada), analítica (ya que se buscar analizar un problema para estudiar sus posibles causas y consecuencias) y transversal (porque la muestra se analiza en este mismo momento).

1. **EXPLORACIÓN DE LOS DATOS – METODOLOGÍA FINAL**

Una vez que se ha aplicado la encuesta, todos los datos han sido recogidos y exportados en un fichero csv que consiste en un dataset con una serie de columnas y filas (que representan variables y estudiantes respectivamente) cuyo significado se adjunta en el anexo 1.

Como científico de datos del centro educativo he recogido el mandato de continuar la presente investigación a partir del dataset. En este caso, he realizado un estudio con herramientas como Python, Visual Studio Code y librerías que he importado tales como Pandas, Numpy, Matplotlib, Seaborn y Plotly.

A continuación, he procedido a realizar una primera exploración de los datos en los que se observan, inicialmente, 395 filas con 33 columnas. Las variables son, mayormente, categóricas y de tipo object, aunque también se pueden observar algunas de tipo int64.

Seguidamente, he decidido solicitar una breve descripción de estadísticos descriptivos de las columnas con la función df.describe()

1. **LIMPIEZA DE LOS DATOS**

Todo científico de datos ha de realizar una adecuada limpieza de datos ante de proceder a su análisis. En consecuencia, he procedido a comprobar:

* La existencia de valores nulos que pudieran perjudicar la obtención de conclusiones: en este caso, no he observado ningún valor nulo.
* Los valores duplicados que pudieran desvirtuar el sentido del análisis: tampoco se ha detectado ningún valor duplicado.

A continuación, y, siguiendo con la limpieza y preparación del dataframe, he eliminado con la función .drop() aquellas columnas que no iban a formar parte de mi análisis, quedándome, en principio, con las que sí pretendía analizar.

He proseguido renombrando la columna sex por gender ya que considero que, utilizando un lenguaje más purista, es más adecuado hablar de género en lugar de sexo.

Los valores outlier o extremos, según lo que se esté analizando, pueden ser susceptibles de ser eliminados. En mi caso, decidí eliminar aquellos correspondientes a la franja de edad de 21 y 20 años pues consideraba que los estudiantes con esas edades estaban por encima del promedio de edad de los alumnos de educación obligatoria.

1. **ANÁLISIS DE LOS DATOS**
   1. ESTUDIO DE LA MUESTRA

Tras efectuar la respectiva limpieza de los datos, he procedido a analizar las características demográficas de mi muestra con funciones como value conts, max, mean, min, mode, goupby y flitros de pandas utilizando máscaras. En consecuencia, he obtenido las siguientes características:

* Distribución de mujeres: 207
* Distribución de hombres: 183
* Edad media: 16 años
* Edad máxima: 19 años
* Edad mínima: 15 años

Por otro lado, es destacable que no ha habido ningún estudiante que no consumiera alcohol, tanto diario, como en fin de semana (únicamente).

* 1. HIPÓTESIS Y EXPLICACIONES

Tras este primer análisis, ya me encuentro en situación de poder plantear hipótesis que puedan arrojar luz a la problemática del consumo de alcohol en la escuela.

**Consideraciones previas:**

* Los datos han sido extraídos de la aplicación de una encuesta. En consecuencia, cuando realice alusión al consumo de alcohol, me referiré a la frecuencia de consumo de alcohol en una escala Likert, ordinal, de 1 a 5 siendo: 1(muy poco consumo), 2(poco consumo), 3(consumo moderado), 4(alto consumo), 5(consumo excesivo). En ningún caso hablaré de mg por litro de aire espirado, número de copas, o grados de alcohol.
* Al abordar las hipótesis, haré siempre referencia a correlación y nunca a causa-efecto. Puede haber una gran cantidad de variables que modulen, contaminen o expliquen esa relación y que no esté tomando en consideración.

**Hipótesis:**

1. *El consumo de alcohol decrece con la edad.*

Tras realizar un scatterplot, puedo señalar que, en mi muestra, se aprecia como, efectivamente, va descendiendo. Esto es relevante, especialmente, durante los fines de semana. La explicación puede ser debido a que, con la edad, crecen las responsabilidades, los adolescentes maduran y ya no necesitan tanta excesiva aprobación social derivada de la continua comparación social con el grupo de iguales. Las prioridades cambian, el grupo de referencia y de pertenencia pasa a ser más reducido.

1. *Las chicas beben más que los chicos.*

Inicialmente por lo que podemos escuchar en los medios, podemos pensar que las mujeres consumen más alcohol que los hombres pero el gráfico countplot nos muestra que no es así. Podemos observar que, tanto en el consumo diario, como en el consumo de fin de semana la frecuencia con la que consumen más alcohol es mayor en los hombres que en las mujeres. Esto puede ser debido a que los ganglios basales de los hombres generan mayor cantidad de Dopamina que los de las mujeres y, en consecuencia, mayor búsqueda de ese efecto placentero.

1. *¿Influye si el tutor es el padre o la madre?*

El gráfico swarmplot muestra que, tanto el consumo diario como en el fin de semana de alcohol es mayor en aquellos estudiantes cuyo tutor es el padre. Podría explicarse a que los padres, tradicionalmente, de manera generalista, han ejercido estilos parentales de educación más autoritarios.

1. *Aquellos estudiantes que reciben más apoyo familiar consumen menos alcohol diario.*

En esta ocasión, vemos que el sunburst nos muestra una conclusión totalmente diferente: aquellos que reciben menor apoyo familiar, beben menos. La explicación es difícil de hallar, puede ser debido a los datos.

1. *Aquellos estudiantes que consumen con más frecuencia alcohol diariamente, obtienen peores notas.*

El gráfico catplot verifica totalmente esta hipótesis. La explicación es obvia para todo el mundo.

1. *Los alumnos que viven en un entorno rural consumen menor alcohol tanto diario como durante el fin de semana.*

Teniendo en cuenta el gráfico catplot, es evidente que el consumo de alcohol es mayor en aquellos que viven en zonas urbanas. Las explicaciones a esta conclusión pueden basarse en que en las zonas urbanas hay mayores lugares a los que acudir para comprar alcohol y a que, simplemente, hay mayor proporción de alumnos que viven en esas zonas y por lo tanto pueden reunirse con mayor facilidad para consumir alcohol. Además, los desplazamientos se pueden realizar con mayor facilidad gracias al transporte público.

1. *Los padres que viven juntos funcionan como un factor protector frente al consumo de alcohol de sus hijos a diferencia de aquellos que están separados.*

Podríamos extraer conclusiones precitadas basándonos en que aquellos estudiantes cuyos padres están separados tienen una menor vigilancia o han pasado por una situación complicada que puede llevarlos a emplear estrategias de afrontamiento desadaptativas. La realidad es que, analizando esta gráfica, he tenido la oportunidad de visionar que los datos están imbalanceados. Dicho problema se analizará en el apartado 6 de la presente memoria.

1. *Los estudiantes que consumen menos alcohol, quieren estudiar una educación superior.*

Teniendo en consideración el sunburst, podemos encontrar que esta hipótesis se verifica. La explicación es sencilla: cuanto menor consumo de alcohol, más conservada se encuentran las funciones cognitivas de la CPDorsolateral implicada en la toma de decisiones y la planificación a futuro.

1. *Las relaciones románticas desempeñan un factor protector frente al consumo de alcohol.*

Una vez más, el sunburst nos permite aceptar la hipótesis. Un argumento a favor puede consistir en que las relaciones de pareja funcionan como un soporte emocional y/o instrumental evitando que los estudiantes desempeñen estrategias de afrontamiento desadaptativas ante sucesos vitales estresantes.

1. *Aquellos estudiantes que pasan más tiempo navegando por internet, consumen más alcohol.*

En esta ocasión, el countplot nos permite confirmar la hipótesis. No obstante, antes de sacar conclusiones precipitadas, hemos de recordar que los datos están imbalanceados. En consecuencia, la mayoría de los estudiantes disponen de acceso a Internet.

El análisis del Heatmap nos muestra que no existe ninguna correlación estadísticamente significativa entre el resto de variables porque ninguna alcanza el coeficiente de correlación de 0.5 . No obstante, podríamos estudiar más a fondo y plantearnos la existencia de relación entre goout - Walc; studytime - Walc y freetime - Wal

1. *Los estudiantes que salen con más frecuencia, consumen más alcohol los fines de semana.*

El gráfico de barras explicita que puede existir relación pues podemos apreciar que, a más frecuencia en salir con los amigos, mayor frecuencia de consumo de alcohol lo que, desde un punto de vista lógico, tiene sentido. Además, la crosstab y el factorplot acompañan el sentido de esta hipótesis.

1. *Los estudiantes que beben más alcohol los fines de semana, estudian menos horas.*

Tras analizar con más profundidad las gráficas se puede observar que, tal y como señalaba el heatmap, no existe ninguna relación entre estas dos variables.

1. *Los estudiantes que disponen de más tiempo libre beben alcohol con mayor frecuencia diaria.*

Del mismo modo, tras analizar con más profundidad las gráficas se puede observar que, tal y como señalaba el heatmap, no existe ninguna relación entre estas dos variables.

1. **CONCLUSIONES Y POSIBLES SOLUCIONES A LA PROBLEMÁTICA**

Las hipótesis que han sido verificadas arrojan luz para entender el problema del consumo de alcohol en los alumnos. Sin embargo, es realmente complicado poder realizar una modificación de dichas variables. Una posible solución puede ser realizar una intervención psicosocial de carácter psicoeducativo en forma de curso formativo tanto con los estudiantes del colegio como con sus familiares.

1. **FORTALEZAS, LIMITACIONES, FUTURAS CONSIDERACIONES DEL EDA**
   1. FORTALEZAS

Una de las principales fortalezas a tener en consideración es que la muestra del estudio es bastante homogénea en cuanto a edad y género. Además, el dataset no disponía de missing values ni de valores que pudieran afectar el estudio de la problemática.

* 1. LIMITACIONES

La limitación más destacable del EDA es que la muestra presentaba datos imbalanceados. Es decir, la proporción de datos para determinadas variables era excesivamente elevada mientras que, para otros valores de la misma variable era reducida. Esto provocaba que la extracción de conclusiones tras el análisis de la hipótesis se viera, inevitablemente, sesgada hacia la confirmación de la misma.

Por otro lado, la metodología seguida cuando se realizó la encuesta presenta una serie de puntos débiles tales como la falta de control de variables contaminadoras o extrañas, el investigador como variable contaminadora y la deseabilidad social a la hora de rellenar dicha encuesta.

* + 1. Principales dificultades encontradas durante la realización del EDA

El mayor reto al que me enfrenté fue al estudio del tipo de variables de mi dataset pues ello era relevante para la correcta selección de las gráficas. En este sentido, tuve notable dificultades para encontrar aquella gráfica que mejor se ajustara a los resultados esperados así como para interpretarlas.

Otro aspecto en el que tuve que invertir tiempo y esfuerzo fue en la realización de las gráficas “sunburst” pues requerían una primeria búsqueda por internet así como la escritura de un código más complejo.

La transformación de la columna “internet” cuyos valores eran booleanos a 0 y 1 supuso un reto pues encontré dificultades en la realización del código.

Por otro lado, encontré notables dificultades para realizar un guardado automático de las gráficas (especialmente de los sunburst) con código y especificar la ruta requerida.

Asimismo, elaborar una función para realizar los sunburst en el .py y llamarla desde el main me llevó bastante tiempo.

Finalmente, interpretar los resultados del Heatmap así como estudiar la relación entre algunas de sus variables, me resultó complicado.

* 1. CONSIDERACIONES FUTURAS

Resultaría altamente recomendable realizar un procesamiento de aquellos datos imbalanceados, por ejemplo, eliminando el exceso de datos en las variables a estudiar para que la proporción pueda resultar similar.

Asimismo, sería interesante ampliar el tamaño de la muestra para poder generalizar conclusiones a la población de interés.

Por otro lado, la verificación de hipótesis podría llevarse a cabo empleando la estadística inferencial para extraer conclusiones con más fiabilidad.

1. **AGRADECIMIENTOS**

*A mi familia por apoyarme y por ser comprensivos ante la imposibilidad, por mi parte, de dedicarles el tiempo necesario.*

*A los profesores Miguel, Leo y Cristina por estar siempre disponibles para cualquier duda que pudiera formularles y por proporcionarme las herramientas y el contenido teórico-práctico necesario para la realización del EDA.*

1. **ANEXO**

**DATASET COLUMNS**

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, output target)