Consumo de alcohol en estudiantes de preparatoria

José Ignacio Esparza Ireta
Licenciatura en Tecnologías para la Información en Ciencias
Escuela Nacional de Estudios Superiores, Unidad Morelia
ignacio.ireta@outlook.com

Miriam Guadalupe Valdez Maldonado
Licenciatura en Tecnologías para la Información en Ciencias
Escuela Nacional de Estudios Superiores, Unidad Morelia
mirluvams@gmail.com

RESUMEN

El objetivo del presente es hacer uso de métodos estadísticos y modelos de aprendizaje automático para analizar y predecir el consumo de alcohol en estudiantes entre las edades de 15 a 19 años, tomando en cuenta las siguientes características: escuela, edad, zona rural o urbana, cantidad de miembros que componen su familia, estado civil de los padres, escolaridad de los padres, trabajo de los padres, tiempo de traslado a la escuela, tiempo que pasan estudiando, reprobaciones, días en los que consumen alcohol, clasificados como entre semana y fines de semana, estado de salud, asistencias, y calificaciones de 0-20 en cada uno de los tres grados de estudio serían equiparables a aquellos de un bachillerato o preparatoria (G1, G2, G3).

KEYWORDS

precisión, recall, exactitud, f1-score, entropía, sensibilidad

I. INTRODUCCIÓN

Cada año, aproximadamente 5,000 jóvenes menores de 21 años mueren como resultado del consumo de alcohol entre menores; esto incluye unas 1,900 muertes por accidentes automovilísticos, 1,600 como resultado de h incluye unas 1,900 muertes por accidentes automovilísticos, 1,600 como resultado de homicidios, 300 por suicidio, así como cientos por otras lesiones como caídas, quemaduras y ahogamientos, según el National Institute on Alcohol Abuse and Alcoholism (NIAAA) de Estados Unidos. [2]

Muchos adolescentes comienzan a beber a edades muy tempranas. En 2003, la edad promedio del primer consumo de alcohol era de unos 14 años, en comparación con los 17.5 de 1965. Las personas que informaron haber comenzado a beber antes de los 15 años tenían cuatro veces más probabilidades de confirmar que también cumplieron los criterios de dependencia del alcohol en algún momento de sus vidas. [3] Por eso, nos hemos dado a la tarea de analizar algunas de las características sociales y de escolaridad en

estudiantes de nivel secundaria, para encontrar relaciones, probabilidades, hacer estimaciones y lograr entrenar modelos que predigan qué tipo de perfil tienen los estudiantes adolescentes más propensos a consumir alcohol de forma excesiva, y cómo ésto se ve relacionado con su desempeño académico o su entorno familiar y social.

II. MATERIALES Y MÉTODOS

Obtención de datos

Los datos utilizados han sido tomados de la plataforma Kaggle, dicho conjunto de datos es *Student Alcohol Consumption,* y cuentan con licencia de dominio público.

Limpieza de datos

Es indispensable asegurarse de qué tipo de datos se tienen y cuáles son las instancias presentadas, para evitar confusiones al momento de realizar el análisis exploratorio de los datos, también, es importante observar tanto los posibles datos atípicos que puedan existir, como la existencia de valores nulos.

Análisis exploratorio de datos

Antes de comenzar a aplicar métodos de Aprendizaje Automático, se requiere de estadística descriptiva para organizar y presentar el conjunto de datos con el propósito de facilitar su uso, con apoyo de tablas, gráficas, medidas de tendencia central y de dispersión. Para esto, fueron utilizadas librerías de Python, tales como *numpy* , *pandas* , *matplotlib* y *seaborn.*

Métodos de clasificación de aprendizaje automático

Se consideraron los siguientes algoritmos de clasificación:

- Categorical Naive Bayes
- Decision Trees
- k-Nearest neighbors (kNN)
- Support Vector Machines

II.A. OBTENCIÓN Y LIMPIEZA DE DATOS

De los dos conjuntos de datos ofrecidos, se tomaron únicamente datos de estudiantes de la clase de matemáticas. Se observó que no existen datos faltantes en ninguna de las instancias, y que, a pesar de que se ofrecen datos de estudiantes con edades de 15 a 22 años, solamente hay 5 estudiantes en el rango de 20 a 22 años, por lo que se decidió separarlos de nuestro estudio, por consiguiente, nuestro análisis se reduce a estudiantes en edades de 15 a 19 años. Después de haber eliminado dichos datos, nuestra muestra está compuesta por 390 estudiantes.

II.B. ANÁLISIS EXPLORATORIO DE DATOS

Ahora que conocemos lo anterior mencionado, proseguimos a visualizar descriptivamente los valores del conjunto de datos.

Index	*count*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
age	390	16.65	1.201	15.0	16.0	17.0	18.0	19.0
Medu	390	2.75	1.096	0.0	2.0	3.0	4.0	4.0
Fedu	390	2.53	1.089	0.0	2.0	3.0	3.0	4.0
traveltime	390	1.45	0.700	1.0	1.0	1.0	2.0	4.0
studytime	390	2.04	0.838	1.0	1.0	2.0	2.0	4.0
failures	390	0.31	0.713	0.0	0.0	0.0	0.0	3.0
famrel	390	3.93	0.894	1.0	4.0	4.0	5.0	5.0
freetime	390	3.22	0.992	1.0	3.0	3.0	4.0	5.0
goout	390	3.10	1.115	1.0	2.0	3.0	4.0	5.0
dalc	390	1.46	0.865	1.0	1.0	1.0	2.0	5.0
walc	390	2.28	1.278	1.0	1.0	2.0	3.0	5.0
health	390	3.56	1.391	1.0	3.0	4.0	5.0	5.0
absences	390	5.69	8.026	0.0	0.0	4.0	8.0	75.0
G1	390	10.90	3.309	3.0	8.0	11.0	13.0	19.0
G2	390	10.71	3.758	0.0	9.0	11.0	13.0	19.0
G3	390	10.40	4.583	0.0	8.0	11.0	13.8	20.0

Tabla 1. Valores descriptivos de los datos

De la Tabla 1., se destaca que la mayoría de los estudiantes de la muestra se encuentra entre los 16 y los 17 años; la mayoría de sus padres cuenta con educación superior o media superior; se muestra que la mayoría vive cerca del instituto, y su tiempo de estudio está dentro de la media; en cuanto a la relación familiar, ésta arroja un porcentaje arriba de la media, lo que permite concluir que, al menos en promedio, los estudiantes cuentan con un ambiente familiar bueno; los días en los que es más común que consuman alcohol son los fines de semana, mientras que entre semana es un poco menos común, aunque aún consumen; la cantidad de faltas a clases está muy por debajo de la media; el promedio de calificaciones por grados está dentro de la media; su estado de salud está por encima de la media.

Adicionalmente, el porcentaje de reprobaciones se considera bajo, ya que solamente 79 estudiantes de 390, han reprobado, es decir, el 20.25% de ellos; el porcentaje de faltas es considerado bajo, ya que de 390 estudiantes, 260 de ellos tienen menos de 10 faltas.

En la Fig. 1., que se encuentra a la derecha, se observa que existe una correlación lineal alta entre las variables “Medu” y “Fedu” y entre “Dalc” y “Walc”, lo cual significa que el nivel de escolaridad de los padres está altamente relacionado con el consumo de alcohol de los estudiantes, sin embargo, no podemos permitirnos concluir que eso sea del todo cierto, ya que anteriormente se observó que la mayoría de los padres cuentan con un nivel de escolaridad alta, lo que podría estar causando la correlación. También se presenta una correlación, aunque mucho menos alta, entre “failures” y “age”, “walc” y “goout”, “freetime” y “goout”. Por otro lado, “failures” muestra una correlación lineal negativa con las variables “Medu” y “Fedu”, lo cual significa que cuando los valores de “Medu” y “Fedu” son más altos, los valores de “failures” son más bajos, y viceversa. De la misma manera, sucede que mientras más tiempo pasen estudiando, consumen menos alcohol, tanto en entre semana como en fines de semana, esto se concluye después de observar una correlación lineal negativa entre la variable “studytime” con las variables “Walc” y “Dalc”.

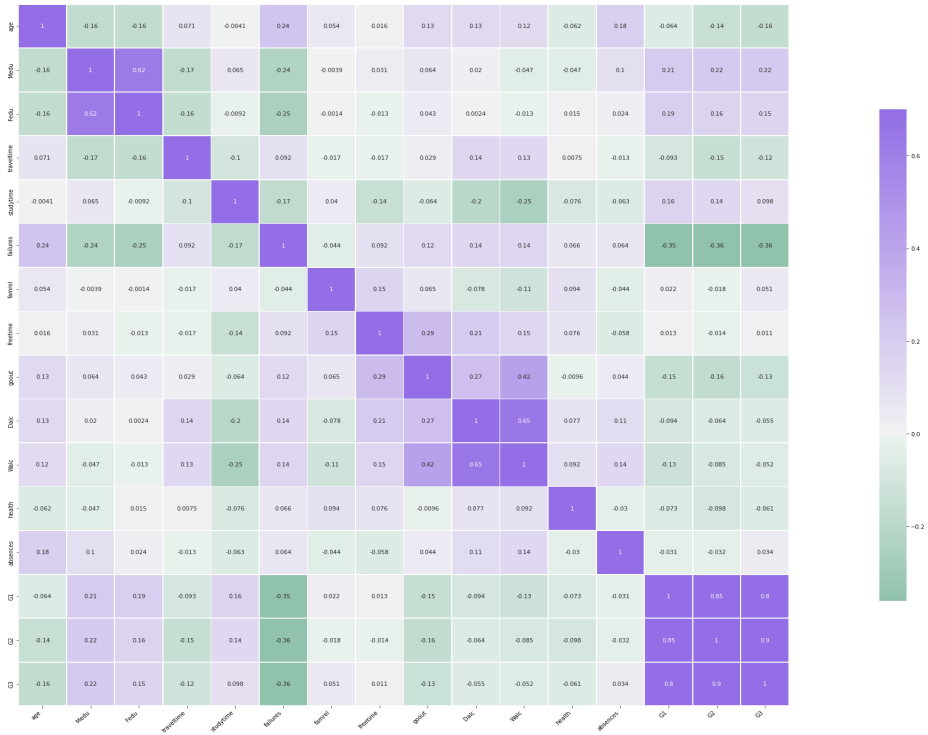

Fig. 1. Correlación lineal entre variables

III. EXPERIMENTOS Y RESULTADOS

III.A. APLICACIÓN DE MÉTODOS DE CLASIFICACIÓN DE APRENDIZAJE AUTOMÁTICO

Con el fin de ayudar a una mejor clasificación, se decidió no hacer una diferenciación tan marcada entre el consumo de alcohol durante los días laborales y el fin de semana. Aunque tenemos presente que el consumo de alcohol durante estos dos grupos de días representa, al menos socialmente, connotaciones distintas de aquellos que lo realizan, nos enfocaremos únicamente en el efecto que los parámetros planteados tienen en el grado relativo de consumo de alcohol en los alumnos. Con el mismo propósito, definimos el consumo de alcohol como moderado para aquellas respuestas en las que los estudiantes identificaban como 0, 1 ó 2, y sin moderación a aquellos que respondieron 4 ó 5.

En lo que respecta a la categorización de los alumnos que tienen un consumo nulo a moderado de alcohol la gran mayoría de los algoritmos pudieron hacer una clasificación correcta, con puntajes que rondan entre el 88% y el 90% de las instancias. En el caso de los alumnos que presentan un consumo sin moderación, los resultados variaron más.

III.A1. Clasificador de Naïve Bayes

En el primer algoritmo, Naïve Bayes categórico, se obtuvo una F1 de 0.81 sugiriendo un desempeño adecuado del mismo. Sin embargo, su precisión y sensibilidad con respecto al grupo de interés alcanzó apenas 0.58 y 0.41 respectivamente. De lo anterior podemos notar que, aunque el modelo es considerablemente bueno en identificar a aquellos alumnos con un consumo bajo de alcohol, estos resultados no nos permiten de manera directa poder identificar a aquellos alumnos que pudieran presentar problemas por su consumo de alcohol.

III.A2 Clasificador de Árboles de Decisión

Por su parte, los árboles de decisión alcanzaron un puntaje f1 ligeramente mayor de 0.83 para la f1; aunque ésta diferencia es relativamente pequeña, representando a penas una mejora de 2.46% con respecto al anterior, la precisión y sensibilidad concerniente al grupo de interés se encontraba bastante por encima del anterior en 0.61 y 0.65 lo que significa un aumento en las mismas de un 5.17% y 58.53%.

III.A3 Clasificador de k Vecinos más cercanos y máquinas de soporte vectorial

Por otro lado, los k vecinos más cercanos obtuvieron un puntaje mucho mayor a los últimos dos con 0.9348 en la f1. Finalmente, las máquinas de soporte vectorial alcanzaron apenas una precisión promedio de 0.83. Por lo que nos decidimos enfocar en el contraste entre los k vecinos más cercanos y los árboles de decisión.

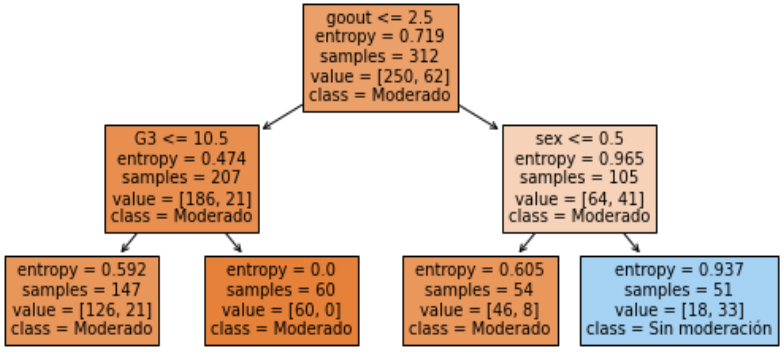

Fig. 2 Representación gráfica del árbol de decisión

III.B RESULTADOS FINALES Y COMPARATIVOS ENTRE K VECINOS MÁS CERCANOS Y ÁRBOLES DE DECISIÓN

Para ambos modelos se buscó utilizar los mejores parámetros y se encontró que para los k Vecinos más cercanos, la k con mejores resultados era k = 3, mientras que la mejor profundidad del árbol era igual a 2.

	k Vecinos	Árboles
f1	0.9348	0.8333
S	0.81	0.65
P	0.69	0.61

Tabla 2. Comparación de métricas entre kNN y árboles de decisión

Por lo tanto, concluímos que el mejor modelo a utilizar en éste caso son los k vecinos más cercanos.

IV. CONCLUSIONES

De todos los modelos seleccionados, k vecinos más cercanos y árboles de decisión tuvieron los puntajes más altos para la f1. Sería entonces razonable esperar utilizar aquel con mayor puntaje; empero, contrario a lo que desearíamos, todos los modelos parecieran ser más sensibles a aquellos alumnos que sí moderan su consumo de alcohol. Durante nuestro análisis procuramos alternativas que nos dieran modelos más sensibles al grupo de interés; considerando relaciones entre aquellos atributos que:

- Dependen en su mayoría del alumno v. gr. Tiempo de estudio, número de salidas, relaciones amorosas, etc.
- Estaban fuera del control del alumno v. gr. La escolaridad de los padres, el tiempo de traslado, ubicación de la vivienda, etc.
- Tenían una mayor correlación entre el atributo y el consumo de alcohol v. gr. número de salidas, ausencias en la clase, etc.
- El conjunto de datos como lo presentan los autores de este.

A pesar de ello, nuestros análisis indican que los mejores modelos se produjeron a partir del conjunto de datos como lo presentan los autores, con las alteraciones mínimas que detallamos anteriormente. Incluso ajustar la f beta para darle más peso a la sensibilidad terminó produciendo modelos que a pesar de que aumentaban marginalmente la sensibilidad, causaban una caída drástica en el desempeño del modelo perdiendo entre el 20% y el 40% de su exactitud. Por lo tanto, consideramos que la mejor opción es utilizar kNN para identificar aquellos alumnos que no están en riesgo de caer en un consumo excesivo de alcohol y procurar apoyar a aquellos que no se encuentren en éste grupo; ya que incluso cuando identificáramos a algún alumno que no requiriera apoyo, la inversión de recursos en ellos no sería tan detrimental como dejar escapar los casos que sí lo requieren.

Finalmente, también recomendamos ampliar el conjunto de datos para que la sensibilidad del modelo alcance a ser lo suficientemente buena para poder identificar directamente a los alumnos de interés.

REFERENCIAS

[1]"Student Alcohol Consumption", *Kaggle.com*, 2016. [Online]. Available: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption. [Accessed: 20- Jun- 2022].

[2]"Alcohol Facts and Statistics | National Institute on Alcohol Abuse and Alcoholism (NIAAA)", *Niaaa.nih.gov*, 2022. [Online]. Available: https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-facts-and-statistics. [Accessed: 20- Jun- 2022].

[3]"Underage Drinking-Why Do Adolescents Drink, What Are the Risks, and How Can Underage Drinking Be Prevented?", *Pubs.niaaa.nih.gov*, 2022. [Online]. Available: https://pubs.niaaa.nih.gov/publications/AA67/AA67.htm. [Accessed: 20- Jun- 2022].