Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Área de Ingeniería e Innovación

Informe Final

Grupo 6: Efectos del alcohol en estudiantes

Integrantes: Nicolás Arancibia
Catalina Bórquez
Matías Miranda
Profesor: Carlos Flores
Auxiliar: Francisco Santibáñez

Fecha de entrega: 14 de Diciembre de 2022
Santiago de Chile

Índice de Contenidos

Índice de Tablas

1. Contexto

Los datos trabajados provienen de una encuesta realizada a 395 estudiantes de dos escuelas secundarias de Brasil y Portugal. Estos basados en su situación social y académica (edad, sexo, educación de los padres, consumo de alcohol, tiempo de estudio, repruebos, entre otras). El objetivo de este modelos es responder a las preguntas: ¿Cuál será el rendimiento de un alumno en la nota final? ¿Hay relación del consumo de alcohol de los estudiantes con su rendimiento académico?

Algunas de las variables importantes que consideramos para la elaboración del modelo de regresión fueron:

1. **Reprobaciones**: Corresponde a la cantidad de asignaturas reprobadas por el estudiante, este dato se relaciona con la problemática directamente con respecto al rendimiento del mismo en la escuela, con la probabilidad de que haya sido un efecto del consumo de alcohol, o de otra forma, que la reprobación genere un malestar o desmotivación derivando en el consumo de alcohol en el transcurso del año escolar. Ordenada de 0 a 3, donde 0 significa que un estudiante no ha reprobado ninguna materia. 1, es que un estudiante ha reprobado una sola materia, y así sucesivamente.

2. **Ausencias a clases**: Esta variable, está enumerada por la frecuencia con la que los estudiantes no han ido a clases, con un mínimo de 0 días a 75 días como máximo. Además, se relaciona con el rendimiento académico, puesto que las ausencias a clases suponen que los y las estudiantes se atrasan con los contenidos.

3. **Consumo de alcohol**: Esta variable se definió al juntar las variables "consumo de alcohol en la semana" y "consumo de alcohol en el fin de semana". Ambas, clasificadas del 1 al 5, con 1 consumo muy bajo y 5 muy alto. Esta variable es clave para poder predecir la nota final.

4. **Eficiencia de estudio**: Se definió la variable de eficiencia de estudio como el cociente entre la cantidad de tiempo de estudio y de tiempo libre de los y las estudiantes después de clases, variables que eran parte de la matriz y que tenían valores numéricos naturales del 1 al 5. Esta variable se relaciona con el rendimiento académico dado que un o una estudiante con un buen manejo de su tiempo de estudio muestra más disposición en el aprendizaje.

5. **Apoyo adicional**: Se creó esta variable para la matriz de características unificando las variables existentes de apoyo escolar adicional con la variable de apoyo adicional de la familia, mediante el promedio simple de sus valores numéricos naturales, que iban de 1 a 5. Esto influye en el rendimiento dado que el apoyo académico adicional sirve como método de retroalimentación para el proceso de aprendizaje de los y las estudiantes.

6. **Notas**: Las notas recopiladas de los estudiantes corresponden a dos periodos diferentes, las notas del primer periodo y las notas del segundo periodo, además, con estos datos se tiene otra variable llamada nota final, correspondiente al promedio de notas entre el primer y segundo periodo escolar. Estos datos son claves dentro del modelo, puesto que son el parámetro para medir el rendimiento académico de cada uno de los estudiantes, parámetro cuyo valor varía entre los 0 y 20 puntos.

7. **Tiempo fuera**: Esta variable corresponde al tiempo que los alumnos dedican en salir con sus amigos, variable que se mide en valores numéricos del 1 al 5, siendo 1 muy bajo y 5 muy

alto. Esta variable tiene implicancias en el rendimiento académico del alumno en el aspecto del tiempo disponible para estudiar o tiempo disponible para consumir alcohol.

Además de estas variables descritas, se agregaron nuevas variables para caracterizar el consumo de alcohol de los y las estudiantes desde su carácter social, variables que posteriormente se eliminaron por la imposibilidad de recopilar o representar los datos en sí (consumo de alcohol de los padres y consumo de alcohol de los amigos). De esta forma se creó una matriz de características base para proyectar el modelo para cumplir con el objetivo planteado.

2. Matriz de características

Inicialmente, la matriz de características estaba compuesta por datos personales de cada estudiante, que permitieran caracterizar a cada uno, con como por ejemplo, la escuela, edad, sexo, dirección, materias reprobadas, frecuencia al salir con amigos, alcohol consumido durante la semana y en el fin de semana, etc. formando una llave de la matriz.

Tabla 1: Variables preliminares de la matriz de características

Escuela	Consumo de alcohol de los amigos	Consumo de alcohol en la semana
Sexo	Salir con amigos	Consumo de alcohol el fin de semana
Edad	Asignaturas reprobadas	Consumo de alcohol de los padres
Dirección	Clases extra pagadas	Apoyo adicional
Ausencias a clases	Nota segundo periodo	Eficiencia de estudio
Nota primer periodo	Nota final	

Luego, se trabajan estos datos pasándolos todos a variables numéricas. De igual manera, se descartaron las variables con menor relación a la pregunta y se crearon nuevas características tales como la eficiencia de estudio y el apoyo adicional.

Quedando finalmente la siguiente matriz:

Tabla 2: Matriz de características

Nº	Escuela	Sexo	Edad	Zona	Familia	Repruebos	Clases part.	t fuera	Alc s	Alc fds	Ausencias	N1	N2	NF	Apoyo adicional	Eficiencia
0	0	0	18	0	1	0	1	4	1	1	6	5	6	6	1	0.66666
1	0	0	17	0	1	0	1	3	1	1	4	5	5	6	1	0.66666
2	0	0	15	0	1	3	0	2	2	3	10	7	8	10	1	0.66666
3	0	0	15	0	1	0	0	2	1	1	2	15	14	15	1	1.5
4	0	0	16	0	1	0	0	2	1	2	4	6	10	10	1	0.66666

En esta matriz se ven las siguientes variables:

1. **Escuela**: Escuela del estudiante, “GP” de Gabriel Pereira o “MS” Mousinho da Silveira.
2. **Sexo**: Sexo del estudiante, “F” si es mujer y “M” si es hombre.
3. **Edad**: Edad del estudiante, desde los 15 a 22 años.
4. **Zona**: Domicilio del estudiante, “U” si vive en zona urbana y “R” en una zona rural.
5. **Familia**: Tamaño familiar de la familia, “GT3” representa más de 3 integrantes y “LT3” si es menor o igual de 3.
6. **Repruebos**: Cantidad de asignaturas reprobadas.
7. **Clases particulares**: Clases extras pagadas, dos opciones “sí” o “no”.
8. **Tiempo fuera**: Tiempo libre con amigos, 1 muy bajo a 5 muy alto.

9. **Alc s**: Consumo de alcohol durante la semana, 1 muy bajo a 5 muy alto.
10. **Alc fds**: Consumo de alcohol durante el fin de semana, 1 muy bajo a 5 muy alto.
11. **Ausencias**: Número de ausencias escolares, de 0 a 93.
12. **N1**: Nota del estudiante del primer periodo, de 0 a 20.
13. **N2**: Nota del estudiante del segundo periodo, de 0 a 20.
14. **NF**: Nota final del estudiante, de 0 a 20.
15. **Apoyo adicional**: Promedio del apoyo adicional familiar y apoyo adicional escolar.
16. **Eficiencia**: Coeficiente del tiempo de estudio sobre el tiempo libre.

Algunas variables eran de tipo string, por lo que se hizo un cambio a int para una mejor lectura de los datos. Estas fueron las siguientes:

1. **Escuela**: “GP” a 0 y “MS” a 1.
2. **Sexo**: “F” a 0 y “M” a 1.
3. **Zona**: “U” a 0 y “R” a 1.
4. **Familia**: “LT3” a 0 y “GT3” a 1.
5. **Clases particulares**: “si” a 0 y “no” a 1.

Siendo la variable NF (nota final) la variable objetivo, en otras palabras, la variable a predecir.

3. Modelos

Se generaron los modelos de árbol de decisión y random forest para 4 variaciones de la matriz de características utilizada, esto para evaluar el peso de algunas variables o conjuntos de variables de la matriz de características sobre la predicción de la nota final de los alumnos. Esto se determinó según las métricas de error para regresión lineal de los modelos generados para las correspondientes matrices. Estos modelos se entrenaron con un 75 % de los datos y se testearon con el 25 % restante de los datos del dataset.

La primera variación de la matriz de características testeada fue sin los datos relacionados al consumo de alcohol de los y las estudiantes (consumo de alcohol en la semana y consumo de alcohol en el fin de semana), se pudo concluir que la sustracción de estas 2 variables generaban un modelo más inestable, dado que tiene valores similares de error absoluto medio (MAE) cercanos a 0.5, pero con aproximadamente una diferencia de 1,2 en la métrica de error cuadrático medio (MSE) en random forest. Estas diferencias son más notables para los modelos de árbol de desición, donde llegan a una deferencia de 3 para el error cuadrático medio (MSE).

Luego, se probó una matriz de características que contuviera como columnas solo variables relacionadas al desempeño académico de los alumnos, por lo que se eliminaron las variables de educación del padre y la madre, el consumo de alcohol en la semana y el fin de semana, el tamaño de la familia y el tiempo fuera (con amigos). Los resultados de las métricas de regresión para esta matriz de características se alejaron del resultado de manera similar a la primera variación para el modelo de random forest. Para el modelo del árbol de desición, el modelo tuvo una precisión muy superior, alcanzando una diferencia de 6 puntos para el error cuadrático medio (MSE).

El último modelo se probó con la finalidad de evaluar la linealidad de la relación entre el consumo de alcohol de los y las estudiantes con su rendimiento académico en la nota final. La matriz de características modeladas solo contiene la llave y los datos de consumo de alcohol en la semana y consumo de alcohol en el fin de semana. Los índices de error para este modelo se dispararon, alcanzando un error cuadrático medio de 39 y 34 para los modelos del árbol de desición y random forest respectivamente.

Finalmente, se creó, usando la matriz de características de la Tabla 2, un modelo usando validación cruzada, con los modelos de random forest, Desicion Tree y Gradient Boosting (que eran los modelos con menor error absoluto medio) para evaluar el rendimiento del modelo utilizado anteriormente (random forest) y compararlo con otros modelos de regresión. Se pudo comprobar que el rendimiento del modelo generado por validación cruzada es ligeramente superior en cuanto a su estabilidad, dado que el error cuadrático medio del “Modelo Ensamble” generado es menor, por una diferencia de alrededor de 1 punto.

4. Resultados

El modelo que presentó los menores índices de error según las métricas de regresión utilizadas (error cuadrático medio, su raíz y error absoluto mínimo) fue el correspondiente a la matriz de características presente en la Tabla 2.

4.1. Árbol de desición

- MSE: 4.393939393939394
- RMSE: 2.0961725582450015
- MAE: 1.121212121212121

4.2. Random forest

- MSE: 2.121212121212121
- RMSE: 1.4564381625088383
- MAE: 0.6262626262626263

4.3. Validación cruzada

Luego, para el modelo creado mediante validación cruzada se lograron los siguientes índices:

- MSE: 1.95
- RMSE: 1.4
- MAE: 0.89

5. Conclusiones

De los resultados de las métricas obtenidas para la matriz de características presentados en la sección de "Resultados", podemos concluir que la manera más precisa de predecir los resultados de los y las estudiantes en base a sus características académicas es mediante el modelo de random forest, que predice los valores de las calificaciones con un error muy bajo. Este modelo tuvo un rendimiento incluso mejor que el "Modelo Ensamble" generado por la validación cruzada, que empeoró debido al error mayor que entregaban los otros modelos usados para este método, pero que entrega un resultado más estable según su error cuadrático medio (MSE) como ya fue mencionado.

Además, se puede concluir de los resultados del modelo que solo usa las variables de consumo de alcohol, que las notas finales de los y las estudiantes no tienen una relación lineal ni directa con el consumo de alcohol en la semana o en el fin de semana, esto dada que el error de este modelo fue el más alto, alcanzando un 1100 % más que para el caso de la matriz de características. A pesar de esto, los datos de consumo de alcohol se relacionan con el resto de datos presentes en la matriz de características dado que le brinda mayor estabilidad al modelo más preciso correspondiente a la matriz de características presente en la Tabla 2, lo que se refleja en que sus índice de error cuadrático medio disminuye notoriamente al agregar las variables de consumo de alcohol de los y las estudiantes, lo que índica que el modelo logra una mejor relación entre los datos, y que estos entregan más información.

Entonces, podemos concluir que el consumo de alcohol en los y las estudiantes tiene un peso en su rendimiento académico, pero este es despreciable en comparación al de otros factores relacionados al estilo de vida y hábitos de estudio. Es por ende, la calificación de las y los estudiantes un efecto multicausal que puede ser abordado por la caracterización académica y social, incluyendo entre estos factores el consumo de alcohol.