Consumo de Alcohol en Estudiantes

Angélica Nayeli Rivas Bedolla
Licenciatura en Tecnologías para la Información en Ciencias
ENES Morelia, Unam
angelica.nayeli@comunidad.unam.mx

Figure 1: Estudiantes en el Campo de Refugio Tierkidi, Región de Gambella

ACM Reference Format:
Angélica Nayeli Rivas Bedolla. 2019. Consumo de Alcohol en Estudiantes. In *Proceedings of Minería de Datos.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Como proyecto final de la materia en Mineria de datos se busca crear un modelo que pueda categorizar estudiantes de acuerdo a su consumo de alcohol semanal. Estos conjuntos de datos son obtenidos de la plataforma de concursos Kaggle donde se describen dos conjuntos de datos con registros repetidos de estudiantes de preparatoria por medio de temas sensibles como atributos.

2 OBJETIVO

Unificar los dos conjuntos de datos correspondientes a estudiantes de matemáticas y portugués de alumnos de escuela preparatoria del estudio de Consumo de Alcohol en Estudiantes y clasificar a los alumnos de acuerdo a los datos sociales, de género y estudio para obtener su consumo de alcohol semanal. Asímismo, hacer uso de librerías como bokeh, sklearn, pandas y graphvizen por medio de 2 archivos con extensión ipynb.

Minería de Datos, 06 de enero del 2020, ENES Morelia, UNAM

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 METODOLOGÍA

Los registros se deben unificar a un solo conjunto de datos tomando en cuenta la existencia de registros que pertenecen a los dos conjuntos de datos y con el resultado crear un clasificador.

Los atributos que componen a los estudiantes son:

(1) school - escuela perteneciente (binario: 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira)
(2) sex - sexo del estudiante (binario: 'F' - femenino or 'M' - masculino)
(3) age - edad del estudiante (numerico: desde 15 hasta 22)
(4) address - tipo de vivienda del estudiante (binario: 'U' - urbano o 'R' - rural)
(5) famsize - tamaño de la familia (binario: 'LE3' - menor o igual a 3 o 'GT3' - mayor que 3)
(6) Pstatus - estado de convivencia de los padres (binario: 'T' - viven juntos o 'A' - apartados)
(7) Medu - educacion de la madre (numerico: 0 - ninguno, 1 - educacion primaria (4to grado), 2 – 5to a 9no grado, 3 – educacion secundaria o 4 – educacion superior)
(8) Fedu - educacion del padre (numerico: 0 - ninguno, 1 - educacion primaria (4to grado), 2 – 5to a 9no grado, 3 – educacion secundaria o 4 – educacion superior)
(9) Mjob - trabajo de la madre (nominal: 'teacher' - profesor, 'health' - relacionados a salud, 'services' - sevicios civiles (e.g. administrativo o policia), 'at_home' - en casa o 'other' - otro)
(10) Fjob - trabajo del padre (nominal: 'teacher' - profesor, 'health' - relacionados a salud, 'services' - sevicios civiles (e.g. administrativo o policia), 'at_home' - en casa o 'other' - otro)
(11) reason - razon por la cual escogió la escuela (nominal: 'home' - cercano a casa, 'reputation' - repputacion de la escuela, 'course' - preferencia del curso o 'other' - otro)
(12) guardian - tutor del estudiante (nominal: 'mother' - madre, 'father' - padre o 'other' - otro)
(13) traveltime - tiempo de transporte de casa a escuela (numerico: 1 - <15 min., 2 - 15 hasta 30 min., 3 - 30 min. hasta 1 hora, or 4 - >1 hora)
(14) studytime - tiempo semanal de estudio (numerico: 1 - <2 horas, 2 - 2 hasta 5 horas, 3 - 5 hasta 10 horas, o 4 - >10 horas)
(15) failures - numero de clases fallidas (numerico: n si 1<=n<3, entonces 4)
(16) schoolsup - soporte extra educacional (binario: yes - si o no)
(17) famsup - soporte educacional familiar (binario: yes - si o no)
(18) paid - asesorias extras pagadas para la respectiva materia. (binario: yes - si o no)
(19) activities - actividades extra-curriculares (binario: yes - si o no)
(20) nursery - asistió a la guarderia (binario: yes - si o no)
(21) higher - quiere estudiar nivel superior educacional (binario: yes - si o no)
(22) internet - acceso a Internet en casa (binario: yes - si o no)
(23) romantic - en una relación romantica (binario: yes - si o no)
(24) famrel - calidad de la relación familiar (numerico: desde 1 - muy mala hasta 5 - excelente)
(25) freetime - tiempo libre después de la escuela (numerico: desde 1 - muy baja hasta 5 - muy alta)
(26) goout - salir con amigos (numerico: desde 1 - muy baja hasta 5 - muy alta)
(27) Dalc - consumo de alcohol entre semana (numerico: desde 1 - muy baja hasta 5 - muy alta)
(28) Walc - consumo de alcohol en fin de semana (numerico: desde 1 - muy baja hasta 5 - muy alta)
(29) health - estado de salud actual (numerico: desde 1 - muy mala hasta 5 - muy buena)
(30) absences - numero de faltas en la escuela (numerico: desde 0 hasta 93)
- Estas calificaciones están relacionadas con la asignatura del curso, Matemáticas o Portugués:
(31) G1 - calificacion de primer periodo (numerico: desde 0 hasta 20)
(32) G2 - calificacion de segundo periodo (numerico: desde 0 hasta 20)
(33) G3 - calificacion del periodo final (numerico: desde 0 hasta 20)

Para poder trabajar con los datos se necesita unirlos en un solo conjunto de datos tomando en cuenta que existen varios estudiantes que pertenecen a ambos. Para ello, se hace uso de la librería Pandas de Python3 y se lleva a cabo en la libreta "crear_csv.ipynb".
Primero se pasa por la función combinar_datos que funciona de la siguiente manera:

(1) Agregar una columna de 1's en cada conjunto de datos para indicar que cursan una materia.
(2) Unir y asignarlo a un tercer conjunto de datos así se obtendrán registros que estuviesen en los dos conjuntos de datos. Se crea una lista de los atributos que no dependen de la asignatura que se esté cursando y todas aquellas que si dependen incluyendo la del inciso anterior serán distinguidas por el sufijo correspondiente a la asignatura.
(3) A los primeros dos conjuntos se les agregan columnas indicando las materias que no cursas y sus equivalentes a valores negativos.
(4) Renombrar las columnas del paso anterior.
(5) Concatenar los primeros dos conjuntos de datos que creará registros repetidos y guardar en un cuarto conjunto de datos.
(6) Borrar registros repetidos, todas las coincidencias.
(7) Se juntas el tercer y cuarto conjunto de datos. El de la union y el de la concatenación que sus registros repetidos serán los de la unión y reindexar.
(8) Crear la variable de salida "Alc" que será la media de "Dalc" y "Walc" y borrar las últimas dos mencionadas, así solo predecir un numerito que es la media de consumo de alcohol a la semana.

De la combinación de los datos se obtienen columnas que su contenido no son números que causarán problemas en el procesamiento de los datos. Por lo tanto, se decide pasar cada columna categórica por un label_encoder, que es codificar atributos categóricos a valores numéricos, y el resto asegurar que sean enteros y se guardan los datos en el archivo estudiantes.csv.

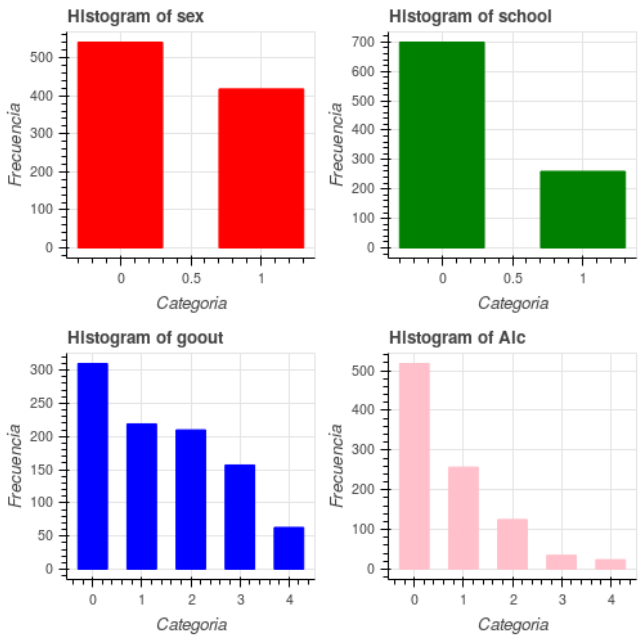
Se hace observación de los datos por medio de histogramas para ver su distribución y una matriz de correlación visual para observar las variables con mayor relación proporcional e inversamente proporcional con el consumo de alcohol semanal.

Para hacer la predicción se opta por un modelo de árboles de decisión disponible en sklearn para Python donde se hace búsqueda del moelo óptimo por medio de búsqueda en malla. Los parámetros empleados en la primera iteración son:

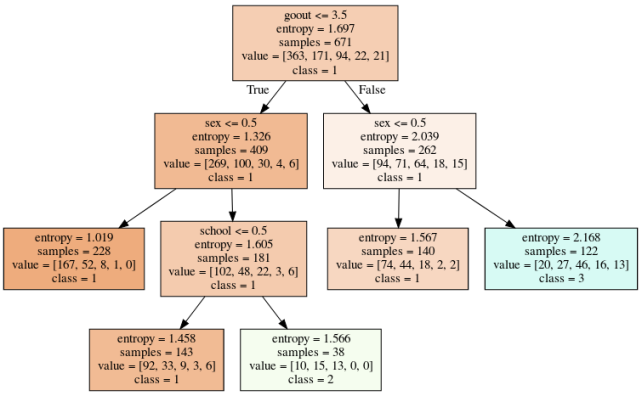
Parámetro	Valor
divisor	'best', 'random'
criterio	'giny', 'entropy'
profundidad máxima	[None, 3, 4, 5, 6, 7, 8, 9]
mínimo número de muestras para hacer división	[20, 30, 40, 50, 60, 70, 80]
mínimo número de muestras por hoja	[1, 6, 11, 16, 21, 26, 31, 36, 41, 46]
máximo número de hojas	[2, 3, 4, 5, 6, 7, 8 , 9]
presort	'Verdadero', 'Falso'

4 RESULTADOS

En el apartado de "Visualizando los Datos" se puede observar la casi nula correlación entre el atributo principal y el resto de los atributos al igual que la desproporción de frecuencias de los valores en los atributos principales.



El mejor árbol de clasificación obtenido es:
Logrando una clasificación de los primeros 3 de 5 clases que se le atribuye a los escasos registros de las otras 2 clases.
Los Mejores parámetros encontrados para el árbol de clasificación son:



Parámetro	Valor
divisor	'best'
criterio	'entropy'
profundidad máxima	3
mínimo número de muestras para hacer división	3
mínimo número de muestras por hoja	2
máximo número de hojas	5
presort	'Verdadero'

Por la falta de información y la desproporción de los datos el modelo obtuvo una precisión del 55.2% . Se recomienda obtener nuevos datos tomando en cuenta la necesidad de nuevos atributos que estén relacionados y un mayor número de registros con frecuencias similares entre los valores para poder obtener resultados más realistas.

5 CONCLUSIÓN

Los datos proporcionados no son suficientes para crear un modelo preciso que se adapte a la naturaleza de los datos. Por ello, no se puede concluir como un modelo exitoso pero si lo más adecuado para esto.

6 REFERENCIAS

- Kaggle.com. (2017). Student Alcohol Consumption. [online] Available at: https://www.kaggle.com/uciml/student-alcohol-consumption.
- Alvarez, C. (2017). 6 tips para crear objetivos inteligentes. [online] Blog.wearedrew.co. Available at: https://blog.wearedrew.co/6-tips-para-crear-objetivos-inteligentes.
- P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
- Fabio Pagnotta, Hossain Mohammad Amran. Email: fabio.pagnotta @studenti.unicam.it, mohammadamra.hossain '@' studenti.unicam.it University Of Camerino.
- Vallejo, N. (2017). Cómo redactar Objetivos de aprendizaje perfectos. [online] OJÚLEARNING. Available at:

https://ojulearning.es/2017/06/como-redactar-los-objetivos-de-aprendizaje-perfectos/ [Accessed 6 Jan. 2020].
- Sinonimos Online. [online] Available at: https://www.sinonimosonline.com.
- Shaikh, R. (2017). Choosing the right Encoding method-Label vs OneHot Encoder. [online] Medium. Available at: https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b.
- Ortiz, J. (n.d.). Cómo empezar una introducción: consejos, ejemplos - Lifeder. [online] Lifeder. Available at: https://www.lifeder.com/como-empezar-introduccion/.
- Molina, L. (2007). Apuntes de Latex. [ebook] Available at: http://metodos.fam.cie.uva.es/ latex/apuntes/apuntes2.pdf.
- Scikit-learn.org. (2020). scikit-learn: machine learning in Python — scikit-learn 0.22.1 documentation. [online] Available at: https://scikit-learn.org/stable/.
- Stack Overflow. (2020). python bokeh, how to make a correlation plot?. [online] Available at: https://stackoverflow.com/questions/39191653/python-bokeh-how-to-make-a-correlation-plot.
- Pandas.pydata.org. (2020). Python Data Analysis Library — pandas: Python Data Analysis Library. [online] Available at: https://pandas.pydata.org.
- Rodríguez, D. (2019). ¿Cómo cambiar el nombre de las columnas en Pandas? - Analytics Lane. [online] Analytics Lane. Available at: https://www.analyticslane.com/2019/05/06/como-cambiar-el-nombre-de-las-columnas-en-pandas/.
- Rodríguez, D. (2019). Seleccionar filas y columnas en Pandas con iloc y loc - Analytics Lane. [online] Analytics Lane. Available at: https://www.analyticslane.com/2019/06/21/seleccionar-filas-y-columnas-en-pandas-con-iloc-y-loc/.
- Rodríguez, D. (2018). Eliminar registros duplicados en pandas - Analytics Lane. [online] Analytics Lane. Available at: https://www.analyticslane.com/2018/06/20/eliminar-registros-duplicados-en-pandas/.
- RSTOPUP. Suma de dos columnas en una pandas dataframe. Available at: https://rstopup.com/suma-de-dos-columnas-en-una-pandas-dataframe.html.
- RIP Tutorial. Pandas - Añadiendo una nueva columna. Available at: https://riptutorial.com/es/pandas/example/5958/anadiendo-una-nueva-column.
- Chris Albon. Join And Merge Pandas Dataframe. Available at: https://chrisalbon.com/python/data_wrangling/pandas_join_merge_dataframe/.