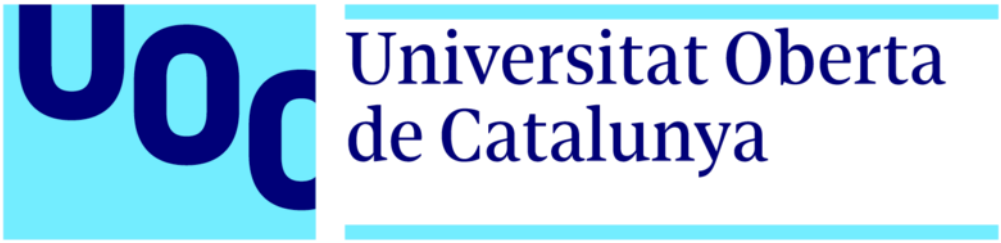

Máster universitario de Ciencia de Datos

Práctica 1 – PRA1

Visualización de Datos – Selección del conjunto de datos

Autor:

Mario Ubierna San Mamés

Índice de Contenido

1. Enunciado

Esta actividad, primera parte de la práctica final, consiste en la selección por parte del estudiante de un conjunto de datos de su interés que será usado en el proyecto de creación de la visualización de datos, de acuerdo con unos criterios establecidos. Básicamente, la temática es libre, pero se valorarán los aspecto siguientes:

1. *Justificad brevemente vuestra selección, ya sea por motivos personales o profesionales. (10%)*

2. *La relevancia del conjunto de datos en su contexto. ¿Son datos actuales? ¿Tratan un tema importante por algún colectivo concreto? ¿Se ha tenido en cuenta la perspectiva de género? (10%)*

3. *La complejidad (medida, variables disponibles, tipos de datos, etc.). ¿Tiene del orden de centenares o miles de registros? ¿Tiene del orden de decenas de variables? ¿Combina datos categóricos y cuantitativos? ¿Incluye otros tipos de datos? Evitad los conjuntos excesivamente simples. (25%)*

4. *La originalidad. No repetid los conjuntos de datos clásicos. Podéis, por ejemplo, combinar o mejorar visualizaciones existentes. ¿Hay otras visualizaciones basadas en este conjunto de datos? ¿Es una evolución o actualización de un conjunto anterior? ¿Habéis enriquecido un conjunto de datos ya existente? (25%)*

5. *Las cuestiones que responderéis con la visualización de datos. ¿Tienen en cuenta los puntos anteriores? ¿Están bien planteadas? ¿Son adecuadas por el conjunto de datos elegido? (30%)*

2. Solución

2.1. Justificación de la selección

El *dataset* elegido es "*Student Performance Data Set*" [1] [2], éste contiene información sobre el rendimiento de los estudiantes en dos institutos de Portugal. Para medir el rendimiento hay variables del tipo: calificaciones de los estudiantes, características demográficas y sociales…

He elegido este *dataset* porque me parece a nivel personal muy interesante el cómo diferentes factores pueden influir en un estudiante, aunque esto es interesante de por sí, me parece aún más interesante las historias ocultas que cuentan los datos, es decir, este *dataset* se suele usar para predecir el rendimiento académico (hacernos preguntas sobre qué factor o factores influyen a nivel académico), pero otro punto para mi de mayor interés es el cómo el consumo de alcohol de la población joven (adolescentes) es afectado por características sociales (nivel de educación de los padres, qué trabajos tienen ellos, si la unidad familiar vive junta o no, cómo es la relación del alumno con sus padres…).

Ésta es una de las historias ocultas que pueden contener estos datos, la cual me parece de gran interés el ver el qué tipo de estudiante puede ser más propenso a ingerir alcohol dependiendo de las características sociales-demográficas.

Como bien he mencionado esta es una línea de estudio, pero puede haber más, por lo que en la segunda parte de la práctica las preguntas que nos hacemos podrían ser diferentes con el fin de obtener información valiosa.

2.2. Relevancia del conjunto de datos

Los datos que se van a usar a priori no son actuales, es decir, según el *paper* [3] que se publicó en su día, los datos se recogieron durante los años 2005 y 2006, pero la publicación de los mismos se hizo en el 2014. Por lo tanto, podemos ver que los datos originales tienen unos 15 años respecto a día de hoy (2021).

Aunque es verdad que a nivel escolar 15 años puede ser una pequeña diferencia, es decir, podría haber mejorado la evolución de los estudiantes en Portugal a nivel escolar durante todo ese tiempo. Sin embargo, el enfoque con el que se usan los datos es completamente diferente, no se busca ver qué factores influyen a nivel escolar, sino que como bien se mencionó en el punto anterior se busca saber cómo los factores sociales-demográficos influyen en el consumo de alcohol en los adolescentes.

Por lo tanto, aunque los datos tienen 15 años vemos que siguen siendo de gran utilidad, ya que este problema lo encontramos fácilmente en la sociedad, sigue habiendo muchos adolescentes bebiendo alcohol, se siguen haciendo botellones, fiestas para menores… Es decir, es un tema importante ya que sigue afectando a una parte de la sociedad, al colectivo de los adolescentes.

Para concluir con este punto, destacar que sí que se tuvo en cuenta la perspectiva de género a la hora de capturar los datos, ya que por cada "registro" de estudiante se captura el sexo, con ello podemos realizar diferentes análisis según si el o la estudiante es hombre o mujer.

2.3. La complejidad de los datos

En este apartado se va a hacer un resumen del conjunto de datos, es decir, del número de registros que hay, qué variables pertenecen al dominio del problema, de qué tipo son éstas…

Lo primero de todo comentar que tenemos 395 registros, por lo que estamos en un *dataset* del orden de centenares, es decir, tenemos una muestra lo suficientemente grande como para poder extrapolar resultados.

En segundo lugar, vamos a realizar un estudio de las diferentes variables que hay. En nuestro caso hay 33 variables, y se combinan tanto datos categóricos como cuantitativos. Las variables que contiene el *dataset* son las siguientes [1]:

- *School:* escuela a la que pertenece el estudiante. Variable categórica binaria.
- *Sex:* sexo del estudiante. Variable categórica binaria.
- *Age:* edad del estudiante. Variable numérica [15,22]
- *Address:* dónde vive el estudiante. Variable categórica binaria.
- *Famsize:* tamaño de la familia. Variable categórica binaria.
- *Pstatus:* estado de convivencia de los padres. Variable categórica binaria.
- *Medu:* nivel educativo de la madre. Variable numérica 5 valores posibles.
- *Fedu:* nivel educativo del padre. Variable numérica 5 valores posibles.
- *Mjob:* trabajo de la madre. Variable categórica 5 valores posibles.
- *Fjob:* trabajo del padre. Variable categórica 5 valores posibles.

- *Reason:* razón por lo que se eligió ese centro. Variable categórica 4 posibles valores.
- *Guardian:* tutor del estudiante. Variable categórica 3 valores posibles.
- *Traveltime:* tiempo de casa al centro. Variable numérica 4 posibles valores.
- *Studytime:* tiempo de estudio semanal del estudiante. Variable numérica 4 posibles valores.
- *Failures:* número de clases pasadas falladas. Variable numérica 4 posibles valores.
- *Schoolsup:* apoyo educacional extra. Variable categórica binaria.
- *Famsup:* apoyo educacional familiar. Variable categórica binaria.
- *Paid:* clases particulares. Variable categórica binaria.
- *Activities:* actividades extracurriculares. Variable categórica binaria.
- *Nursery:* si fue a la guardería. Variable categórica binaria.
- *Higher:* si quiere continuar con estudios superiores. Variable categórica binaria.
- *Internet:* si tiene acceso a internet en casa. Variable categórica binaria.
- *Romantic:* si tiene pareja sentimental. Variable categórica binaria.
- *Famrel:* calidad de la relación entre la familia y el estudiante. Variable numérica 5 valores posibles.
- *Freetime:* si tiene tiempo libre después del colegio. Variable numérica 5 valores posibles.
- *Goout:* si pasa tiempo con sus amigos. Variable numérica 5 valores posibles.
- *Dalc:* consumición de alcohol entre semana. Variable numérica 5 posibles valores.
- *Walc:* consumición de alcohol los fines de semana. Variable numérica 5 valores posibles.
- *Health:* salud actual del estudiante. Variable numérica 5 valores posibles.
- *Absences:* número de ausencias en clase. Variable numérica [0,93].
- *G1:* nota durante el primer trimestre. Variable numérica [0,20].
- *G2:* nota durante el segundo trimestre. Variable numérica [0,20].
- *G3:* nota final del curso. Variable numérica [0,20].

En resumen, vemos que el *dataset* es complejo ya que nos proporciona mucha información a partir de sus 33 variables, es por ello que podemos jugar con él para obtener información oculta, esto es lo que se pretenderá en la siguiente práctica.

2.4. Originalidad

Respecto a este apartado podemos ver que el *dataset* no es el clásico que se utiliza para visualización de datos, es más el objetivo de este conjunto de datos es poder clasificar los alumnos según su nota final o predecir directamente dicha nota final, es decir, tiene un enfoque más de minería de datos que de visualización. Sin embargo,

debido a que el *dataset* tiene un enfoque diferente al que vamos a usar se podría considerar que sí que es original. Además, con este conjunto de datos vamos a ver qué historias ocultas nos cuentan los datos, es decir, no buscamos dar respuesta a la típica pregunta de cómo es el rendimiento de los estudiantes, sino que se quiere buscar respuestas sobre el consumo del alcohol a partir de datos sociales-demográficos.

En cuanto a si existen visualizaciones basadas en este conjunto de datos, la respuesta es sí, pero estas visualizaciones son el resultado de aplicar diferentes algoritmos de minería de datos, es decir, visualizaciones de clasificación o de regresión, por lo tanto, son visualizaciones totalmente diferentes a lo que se busca con esta práctica.

Finalmente, no se ha enriquecido el *dataset* ya que de por sí está muy bien detallado, se podría mejorar el mismo si se buscara un enfoque más académico, como por ejemplo discretizar las variables G1, G2 y G3 para que tuvieran valores del tipo "Suspenso, Suficiente, Bien, Notable, Sobresaliente", pero como nuestro objetivo es otro no se ha hecho.

2.5. Las cuestiones que se responderán

La cuestión que se busca responder con este análisis de los datos es simple, ¿cómo afectan diferentes factores sociales-demográficos al consumo de alcohol entre adolescentes?

Como podemos ver la pregunta es sencilla, pero esta a su vez implica diferentes preguntas como por ejemplo, ¿dependiendo de si eres hombre o mujer se consume más alcohol o menos?, ¿dependiendo de los estudios/trabajo de los padres hay un mayor/menor consumo?, ¿si el estudiante tiene pareja influye?...

En resumen, son múltiples preguntas las que nos podemos hacer, dependiendo de como sea la segunda parte de la práctica se buscarán resolver más o menos, pero todas ellas relacionadas con la pregunta general, es decir, ¿cómo diferentes factores sociales-demográficos influyen en el consumo de alcohol entre estudiantes?

3. Bibliografía

[1] «Student Grade Prediction». https://kaggle.com/dipam7/student-grade-prediction (accedido nov. 18, 2021).

[2] «UCI Machine Learning Repository: Student Performance Data Set». https://archive.ics.uci.edu/ml/datasets/student+performance (accedido nov. 18, 2021).

[3] P. Cortez y A. Silva, «USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE», p. 8.