Analisis Exploratorio de Datos usando el dataset de Consumo de Alcohol de Estudiantes

Cesar Arcos
Licenciatura en Tecnologías para la Información en Ciencias
ENES Morelia, UNAM
racec9999@gmail.com

Eduardo Ceja
Licenciatura en Tecnologías para la Información en Ciencias
ENES Morelia, UNAM
lalitoceja@gmail.com

Figure 1: El problema del alcohol

ABSTRACT

A menudo se ha culpado al consumo de alcohol por el bajo rendimiento de las personas, especialmente los estudiantes. El objetivo de este artículo es realizar un análisis exploratorio de un conjunto de datos utilizando varios métodos de aprendizaje automático sobre el consumo de bebidas alcoholicas en los estudiantes y múltiples factores familiares y personales, así como su rendimiento académico, y ver qué grupos y patrones, para ver si están relacionados. o no. Además, al final del análisis, continuaremos la exploración de datos con una visualización realizada en una aplicación denominada Streamlit.

All authors contributed equally to this work.

Minería de Datos, 26-01-2021, ENES Morelia

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

KEYWORDS

clustering, exploratory data analysis, data visualization, unsupervised learning

ACM Reference Format:
Cesar Arcos and Eduardo Ceja. 2021. Analisis Exploratorio de Datos usando el dataset de Consumo de Alcohol de Estudiantes. In *Proceedings of Minería de Datos*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCCIÓN

El consumo de alcohol ha sido el principal culpable del bajo rendimiento académico pero de hecho, hay más factores que ese como relaciones familiares, contexto social, interés de los estudiantes, situación económica, entre otros. En este artículo intentaremos ver cuáles son los otros factores que contribuyen al mal desempeño académico de los estudiantes. Como se trata de un análisis exploratorio, llevaremos a cabo métodos de aprendizaje no supervisados como KMeans, agrupamiento jerárquico como también la implementación de visualizaciones de los datos y la información que producimos.

Este artículo está organizado de la siguiente manera: Sección 2 descripción del conjunto de datos que exploramos. Sección 3 Descripción de la tarea de aprendizaje no supervisado que se llevó

a cabo.Sección 4 Descripción de experimentos., Sección 5 Análisis y discusión de los resultados.Sección 6 Conclusiones.

2 DESCRIPCIÓN DEL CONJUNTO DE DATOS

Los datos fueron obtenidos de una encuesta de estudiantes de los cursos de matemáticas y lengua portuguesa a nivel secundaria, pueden consultar el conjunto de datos directamente de kaggle Dataset, El conjunto de datos contiene datos relacionados con el interés social, relación familiar, género,nivel económico, etc.El conjunto de datos cuenta con 395 registros y 33 columnas para la clase de matemáticas, para la clase de portugués cuenta con 649 registros y 33 columnas.
En cuanto a la información del dataset podemos condensar la más importante en estas categorías

(1) Familiares.
- Pstatus: Estado de convivencia de los padres.
- Medu: Educación de la Madre.
- Fedu: Educación del Padre.
- Mjob: Trabajo de la Madre.
- Fjob: Trabajo del padre.
- famsize: Tamaño de la familia.
- Famrel: Relación con la familia.

(2) Escolares
- traveltime: Tiempo de traslado a la escuela.
- studytime: Horas de estudio semanal.
- failures: Numero de clases reporbadas.
- schoolsup: Educación extra.
- activities: Actividades extracurriculares.
- higher: Desea estudiar educación superior

(3) Ocio.
- romantic: Cuenta con relación amorosa
- freetime: Tiempo libre después de la escuela
- goout: Salir con amigos
- Dalc: Consumo diario de alcohol
- Walc: Consumo semanal de alcohol

3 DESCRIPCIÓN DE LA TAREA DE APRENDIZAJE NO SUPERVISADO

El método de aprendizaje no supervisado fue Kmeans el cual es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Una breve descripción de cómo funciona Kmeans.

(1) Elegir el número k de clusters
(2) Seleccionar al azar k puntos, los baricentros
(3) Asignar cada punto al baricentro más cercano
(4) Calcular y asignar el nuevo baricentro de cada cluster
(5) Reasignar cada punto de los datos a su baricentro más cercano.Si ha habido nuevas asignaciones, ir al paso 4, si no terminar.

4 DESCRIPCIÓN DE EXPERIMENTOS

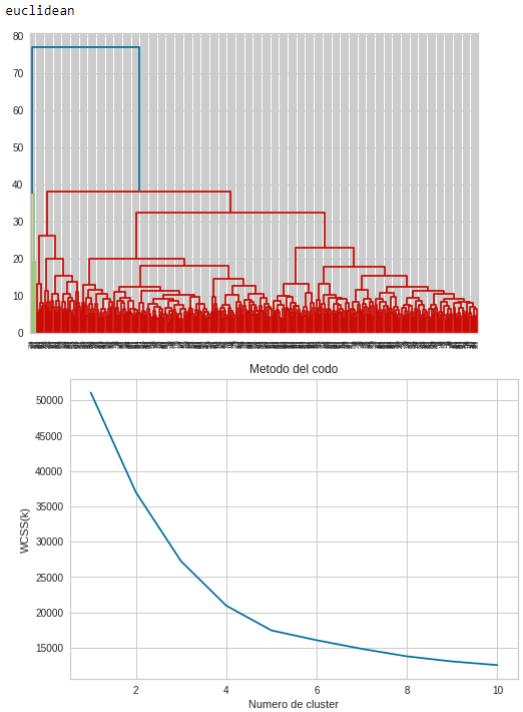
En cuanto a los experimentos para analizar nuestros datos realizamos muchas gráficas para ver si había alguna relación con las calificaciones finales de ambas materias. Estos son algunos de los atributos que comparamos para ver si existía correlación con la calificación:

- Edad
- Educación de la Madre
- Educación del Padre
- Relación amorosa
- Internet
- Tamaño de la familia

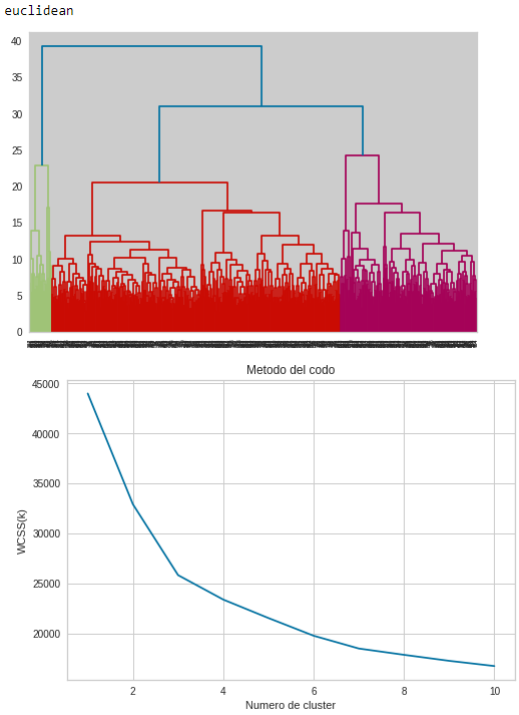
En cuanto a las gráficas para ver que podía llevar al consumo de bebidas alcohólicas comparamos los siguientes atributos:

- Que tanto salen
- Relación con la familia
- Estatus parental
- Salud

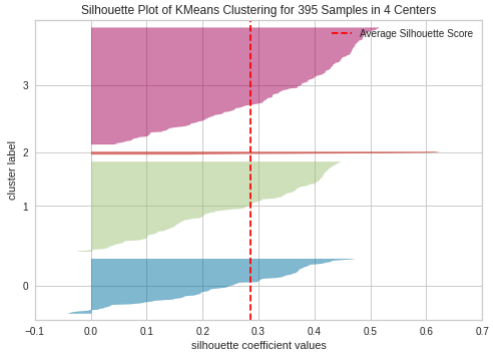
En cuanto a los clusters utilizamos el método de partición Kmeans el cual describimos en la sección 3, para tener un conocimiento más exacto decidimos usar dendogramas con diferentes métricas de distancia, em lo cual nos llevamos una gran sorpresa que analizaremos más a fondo en la sección 5. Para corroborar los dendogramas decidimos también aplicar el método del codo con el cual llegamos a la conclusión de utilizar 4 clusters para la materia de matemáticas y 3 clusters para la materia de portugués.Ahora presentaremos las gráficas obtenidas para la clase de Matemáticas



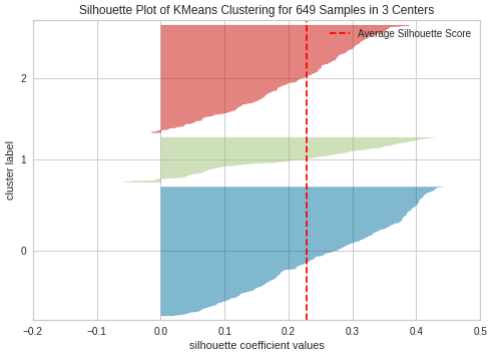
Para la clase de Portugués nos encontramos con estas gráficas:



Se resalta la importancia de las pruebas de bondad de ajuste en la selección de la cantidad idónea de clusters que mejor representen los datos, nuestra prueba de bondad y ajuste fue la visualización de Silhouette. Se puede observar que para la materia de matemáticas se tiene un cluster muy delgado el cual es algo interesante que vamos a analizar en la siguiente sección



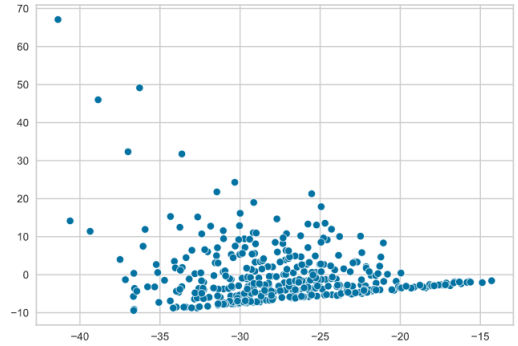
Para la materia de portugués hay un cluster muy delgado pero no tan marcado como en la clase de matemáticas.



Nuestro último experimento y más importante fue hacer una factorización SVD de rango para poder graficar nuestros datos y ver su comportamiento el cual nos dio un nuevo panorama para nuestro análisis.

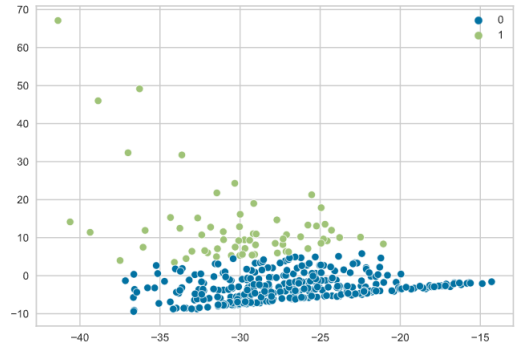
5 ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS

Lo que se descubrió del análisis de datos es que los datos no se podían separar en grupos bien definidos, por lo que al realizar la factorización SVD y utilizar su aproximacion de rango 2 para poder visualizarla en una gráfica, vimos que estaban distribuidos de la siguiente manera. Primero para el dataset de los que tomaron matemáticas:

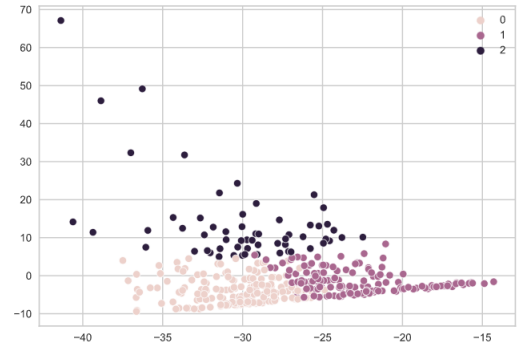


Ahora para cada grupo que encontramos:

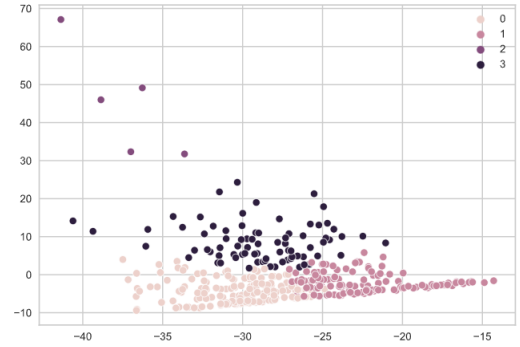
$k = 2$



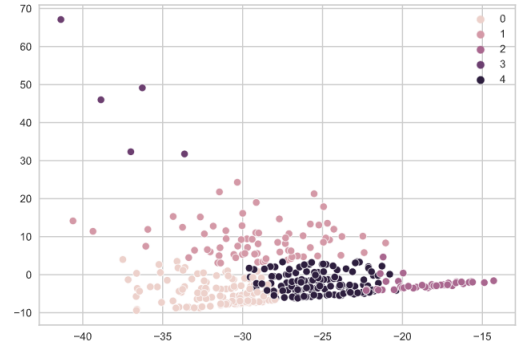
$k = 3$



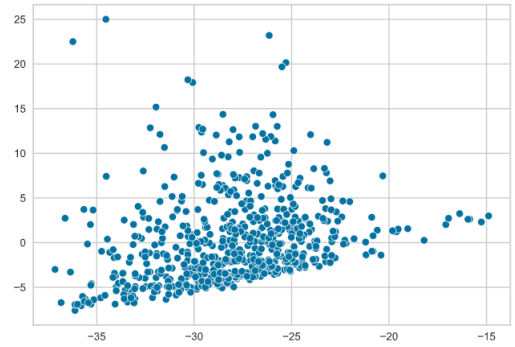
$k = 4$



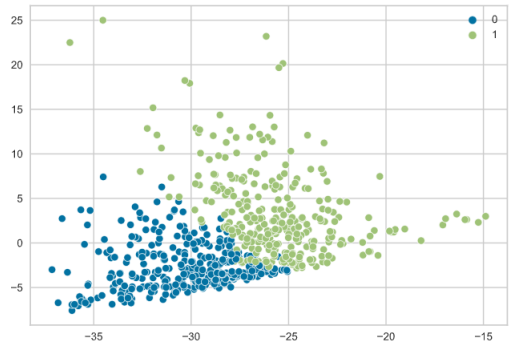
$k = 5$



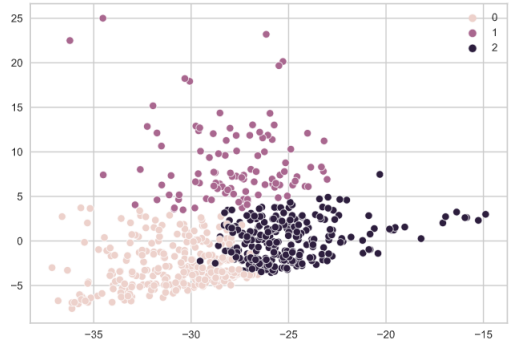
Dataset de los que tomaron Portugués:



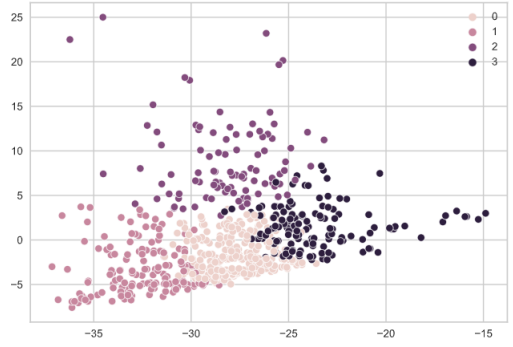
$k = 2$



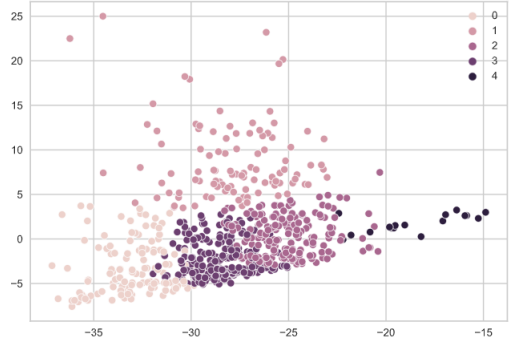
$k = 3$



$k = 4$



$k = 5$



Lo que sí encontramos de esto, es que dentro de estos datos, hay entre uno o dos grupos de personas que si están bien definidos. En el caso de matemáticas, descubrimos que el grupo mejor definido es de los estudiantes que sí aprobaron el curso o que sacaron una calificación mayor al promedio y en el caso de los de portugués, son

las personas que también tienen el promedio más alto de la clase además que sus padres tienen mayor educación, pero tienen la peor salud de todos, aunque casi no toman alcohol y realizan actividades extracurriculares y casi no tienen faltas. Sobre los demás grupos que encontró el algoritmo son un más complicadas de clasificar aunque se puedes rescatar unas características como: En la clase de portugués los dos cluster que quedan se clasifican en estos dos, el primero son las personas que más alcohol consumen en todos los clusters también los que tienen más falta pero no faltan por salud ya que tienen un 4, de calificación tienen el promedio y es el cluster con menor calificaciones de la escuela Gabriel Pereira, el segundo cluster restante se caracteriza por: Su escuela es Mousinho da Silveira a diferencia de los otros dos,tienen un consumo de alcohol normal ya que es un dos, casi no tiene faltas es un grupo muy similar a los dos anteriores con una gran excepción la cual es que tienen el peor promedio, tienen 9.

Para los 3 agrupamientos que quedan en la clase de matemáticas. Este cluster presenta la característica de que sus padres tienen menor educación que los otros clusters, el trayecto a la escuela es de 2 horas el cual puede afectar al rendimiento, ya que por lo menos han reprobado una materia lo cual no parece mejorar ya que cuentan con la calificación más baja que es 6. Lo que diferencia al tercer cluster es que los estudiantes que pertenecen a este cluster tienen menos tiempo libre y no participan en actividades extracurriculares, el dato más asombroso es que sus faltas a clases ascienden a 53 puede que sean las personas que trabajan en cuanto a sus calificaciones es un promedio de 9.Este cluster se diferencia por ser muy equilibrado ya que en la mayoría tiene lo mismo que los otros cluster con diferencia de que han reprobado por lo menos una clase y han faltado en promedio 15 veces,además su calificación final está a la mitad, un 10.

En la aplicación de streamlit hay una visualización interactiva donde se pueden pedir graficar los grupos en 2D o 3D con los grupos con los que corrimos los experimentos, los cuales son para $k \in \{2, 3, 4, 5\}$.

Algo a tomar en cuenta al hacer cualquier análisis es tomar en cuenta la localización geográfica de la persona ya que puede variar mucho en cuanto a países e incluso en cuanto a regiones, este conjunto de datos tiene la particularidad que es de portugal, si indagamos más en los datos encontramos el nombre de las escuelas Gabriel Pereira y Mousinho da Silveira las cuales están situadas en el distrito de évora. No conseguimos una correlación muy marcada en cuanto a la cantidad de alcohol consumida con las calificaciones, pero esto puede variar si lo vemos en otro país de habla portuguesa como brasil el cual tiene un porcentaje mayor de consumo de alcohol [Review 2021]

6 CONCLUSIONES

Lo que se puede concluir de este trabajo, es que las tareas de aprendizaje no supervisado no siempre son sencillas ya que al no haber una especie de "respuesta correcta" como etiquetas, se tiene que hacer una análisis más profundo de lo que el grupo está clasificando. Eso sumado con el caso de que nuestros datos no tenían muchos grupos definidos, fue un gran reto extraer información del clustering que se hizo. Este trabajo trata de responder si hay una especie de correlación con el consumo de alcohol con la vida académica tomando en cuenta varios factores extras no solo si toma o no. ya que hay que tener en mente que la correlación no es necesariamente causación. Lo que encontramos demuestra que no hay una correlación clara entre las calificaciones y el consumo de alcohol. Eso no significa que no hayamos aprendido y descubierto cosas acerca de los datos, pero lo mas importante son las preguntas que hemos ido haciéndonos acerca de los datos. Estas son las que nos hicimos: ¿Si añadimos todavía mas clusters, se empezarían a definir mas los grupos? ¿Como se podrían limpiar los datos para mejorar la definición de los clusters?

ACKNOWLEDGMENTS

Los autores de este artículos quisieran dar las gracias a la Dra. Marisol Flores y a Armando Ortíz por impartir el curso de Minería de Datos así como proporcionar varios materiales complementarios que nos ayudaron a llevar a cabo un análisis y visualización de datos mas completa.

REFERENCES

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

Numpy. 2020. Numpy documentation. https://numpy.org/doc/stable/.

Pandas. 2020. Panadas documentation. https://pandas.pydata.org/docs/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Leskovec Rajaraman and Ullman. 2013. *Mining of Massive Datasets.*

World Population Review. 2021. Alcohol Consumption by Country 2021. https://worldpopulationreview.com/country-rankings/alcohol-consumption-by-country.

Seaborn. 2020. Seaborn: Statistical data visualization Documnetation. https://seaborn.pydata.org/api.html.

Stremlit. 2020. Streamlit documentation. https://docs.streamlit.io/en/stable/.

Yellowbrick. 2020. Yellowbrick: Machine Learning Visualization. Visualizers and API.