Proyecto Final - Minería de Datos

Aplicación del Algoritmo K Means a un dataset sobre la depresión.

Arely Luis
Tecnologías para la Información en Ciencias
UNAM Campus Morelia
arelyluis@comunidad.unam.mx

ABSTRACT

El siguiente documento hace una implementación del algoritmo K-Means a un conjunto de datos sobre la información personal, situación económica, flujos financieros y nivel educativoto de pacientes con depresión. Además realiza análisis estadistico de variables de interés. El conjunto de datos fue tomado de la plataforma Kaggle. Es un dataset reducido de un estudio de 2015 realizado por El Busara Center en el condado rural de Siaya, cerca del lago Victoria en el oeste de Kenia. El dataset original incluye más de 70 caracteristicas, entre ellas información sobre la composición de los hogares, la actividad económica y la salud.

A continuación se describe los pasos de prepocesamiento y limpieza de datos, asi como la interpretación de los resultados obtenidos.

ACM Reference Format:
Arely Luis. 2022. Proyecto Final - Minería de Datos: Aplicación del Algoritmo K Means a un dataset sobre la depresión.. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 EXPLORACIÓN DE DATOS

Como ya se menciono anteriormente el dataset fue obtenido de Kaggle, el conjunto de datos reducido contiene 23 variables las cuales son:

- 'SurveyId'
- 'VilleId'
- 'sex'
- 'Age'
- 'Married'
- 'NumberChildren'
- 'EducationLevel'
- 'TotalMembers'
- 'GainedAsset'
- 'durableAsset
- 'saveAsset'
- 'livingExpenses'
- 'otherExpenses'
- 'incomingSalary'
- 'incomingOwnFarm'
- 'incomingBusiness'
- 'incomingNoBusiness'
- 'incomingAgricultural'
- 'farmExpenses'
- 'laborPrimary'
- 'lastingInvestment'
- 'nolastInginvestment'
- 'depressed'

Se redujo a 13 las variables a utilizar en el análisis, ya que el usar todas abarcaba aspectos que no se quieren tomar en cuenta y que no quedan muy claro a que se refieren. Los atributos que se dejaron son : 'sex', 'Age', 'Married', 'NumberChildren', 'educationLevel','totalMembers', 'gainedAsset', 'durableAsset', 'saveAsset','livingExpenses', 'otherExpenses', 'incomingSalary', 'depressed'. Con estas 13 variables se busca encontrar patrones frecuentes que los analizados tengan, y clasificarlos en grupos que compartan similitudes.

Continuando con la exploración se verificó el tipo de dato de los atributos, todos son int64. Además se obtuvo la dimensión del dataframe la cual es (1429, 13).Un paso importante a la hora de indagar en los datos es verificar si existen datos faltantes, convenientemente todos los atributos estaban completos.

2 ANALISIS ESTADISTICO

Para esta sección se realizaron unos gráficos que muestran los contrastes en porcentaje y cantidad de distintos atributos. El primero de ellos es un histograma que muestra como esta distribuida la edad en los analizados, se encontro que la media de la edad es de 34.77 años.

El segundo de ellos fue un histograma que arroja la distribución de nivel educativo, cabe resaltar que se mide por número de años de estudio siendo 19 el máximo de años cursados. La media en cuanto nivel educativo es de 8.68 años.

Por último se realizó un diagrama de pastel que muestra el contraste de género en los registros. Se observa que el 91.8 de los registros son de personas del género masculino mientras que el 8.19 femeninos.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

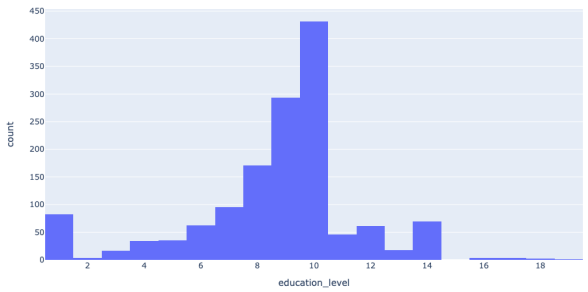

Figure 1: Histograma respecto a la edad en los registros del dataset. Los niveles de educación se miden por años cursados.

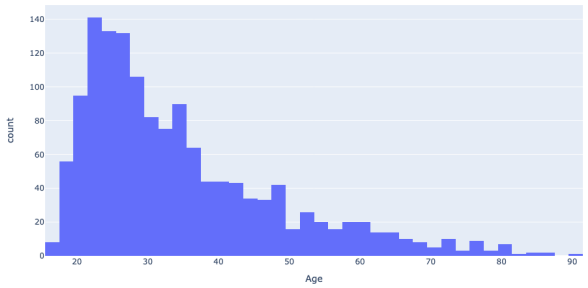

Figure 2: Histograma que muestra la cantidad de personas que comparten la misma edad en los registros del conjunto de datos. Se observa que la mayor cantidad se encuentra en el intervalo de 20-30 años

3 K-MEANS

El método que elegí para el analisis de mis datos es K-Means, antes que todo comencé por normalizar mis datos, use el tipo min-max normalization. Me pareció importante realizar este paso ya que a pesar de que mis datos son númericos en todos los casos se encuentran en escalas demasiado separadas. Como métodos predecesores a K-Means utilice clustering con ayuda de dendrogramas, índice de Silhouette y método del codo como guía para elegir el valor K.

Comenzando por el custering con dendrograma en un principio lo realice con los datos sin normalizar y me arrojaba un grafico que indicaba la presencia de 6 clusters aunque a decir verdad no los considere muy bien definidos, después de normalizar el dataframe y realizar nuevamente el dendrograma se pueden distinguir la formación de 4 cluster mucho mejor definidos, por lo que para este caso se tomo una K = 4.

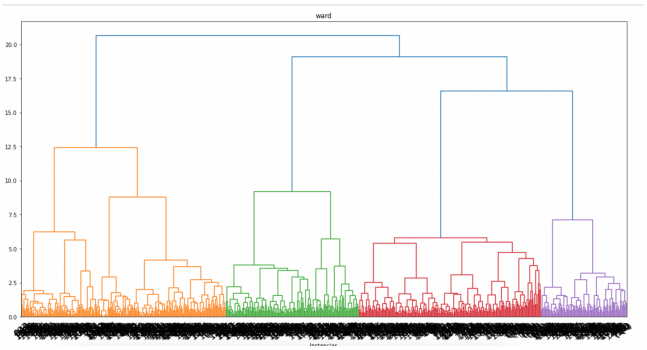

Figure 3: Dendrograma a 10 niveles usando el criterio ward y la distancia Euclidiana. Se observa la formación de 4 clusters bien definidos.

Como otra técnica para encontrar la K, elegí el índice de Silhouette hice graficos para 2-9 clusters. Los resultados me parecieron muy confusos, ya que no me ayudaron a poder encontrar un K, las proporciones de los conjuntos eran demasiado diferentes, el resultado que mas me convencio fue el de K = 6.

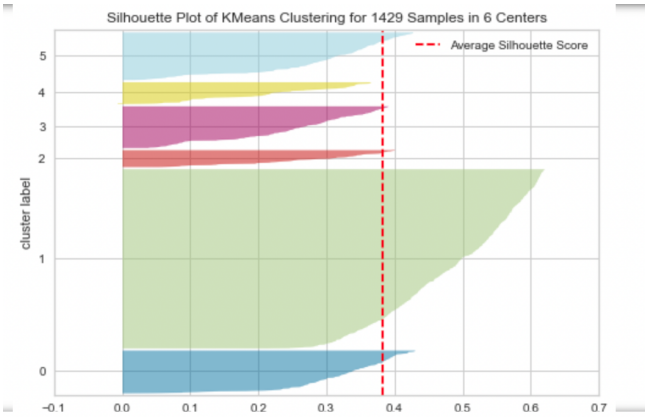

Figure 4: Grafico utilizando el índice Silhouette. Se eligió a K = 6, ya que era el que mejor se ajustaba a los parámetros.

La última técnica que apliqué fue el método del codo, realicé graficos para 2-10 clusters, como resultados pude ver que los datos dejaron de cambiar drasticamente a partir del cluster 4, por lo que para este método elegí K = 4.

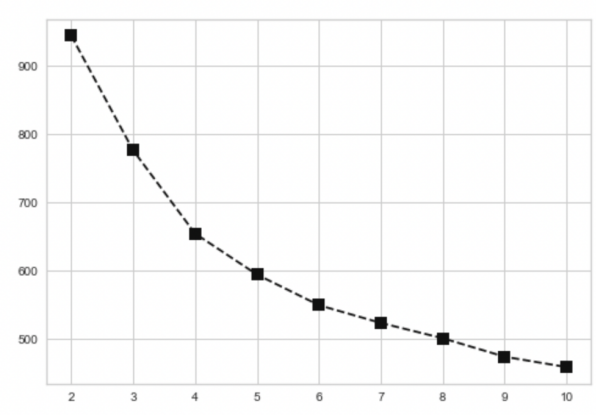

Figure 5: Diagrama de codo para 2-10 clusters. Se observa que el pico de camibio sucede a partir del 4 clúster.

Una vez realizado el análisis exploratorio aplicando los recursos anteriormente mencionados, elegí K = 4 para aplicar el algorítmo K-Means. Hice uso de la reducción de dimensionalidad lineal para comprimir los datos a dos dimensiones. Posterior apliqué K-Means a este nuevo conjunto de datos de dos dimensiones que condensaba la imformación del dataset entero. Por último grafiqué los resultados y obtuve el diagrama de la figura . Se puede observar el como los datos se separan en 4 secciones aunque estas secciones no son resultado propiamente del algoritmo ya que se puede notar que estas secciones pertenecen a más de 1 cluster. Tras analizar este grafico por un tiempo llegué a la conclusión de que una posible razón es la dispersión de los datos, que no son predilectos a una situación creada con datos sintéticos en los cuales la desviación estándar 1 y los datos se presentan de una manera estética y conveniente para el análisis.

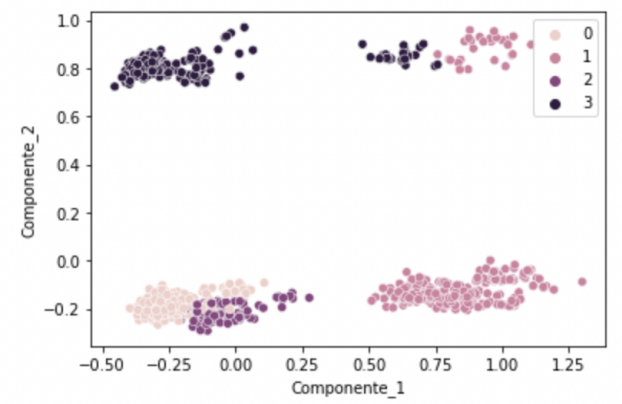

Figure 6: Diagrama de dispersión utilizando el algóritmo K-Means tomando en cuenta una K=4, se observan datos bastante dispersos no propiamente por el clustering, sin embargo 2 de los 4 clusters se definen de buena manera.

4 TABLERO INTERACTIVO

Para la creación del tablero interactivo hicé uso de la biblioteca de Python Dash y los graficos de Plotly. En el tablero muestro graficos que se hicieron en el análisis estadistico que cambian conforme se actualiza la edad en el tablero.

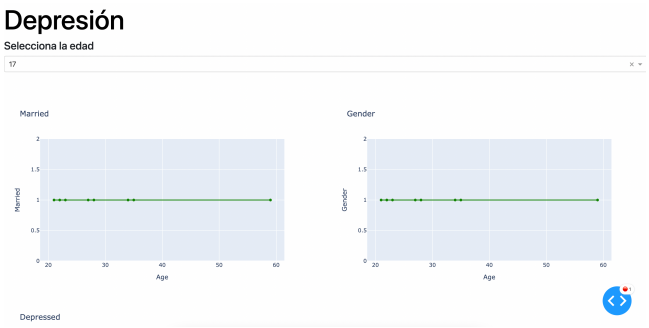

Figure 7: Tablero interactivo utilizando las bibliotecas de Python Dash y Plotly. Se crean 3 gráficos que cambian conforme se modifica la edad en el tablero. Los graficos integrados son: Estado Civil, Género, Diagnóstico.

5 CONCLUSIONES

En conclusión se encuentra la formación de 4 clústers en el conjunto de datos, que comparten caracteristicas encontradas en los atributos. A pesar de contar con datos muy variados y de distintas índoles, es posible encontrar patrones de reconocimiento.

6 BIBLIOGRAFÍA

Plotly. (s. f.). https://plotly.com/python/ Adams, R. P. (s/f). K-means clustering and related algorithms. Princeton.edu. Recuperado el 8 de diciembre de 2022, de https://www.cs.princeton.edu/courses/archive/fall18/cos324/