INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

Evidencia 2: Proyecto de aprendizaje no supervisado

Por

Daniel Eduardo Arana Bodart, A01741202

Jose Manuel Guerrero Arellano, A01747623

Valeria García Hernández, A01742811

Modelación del aprendizaje con inteligencia artificial (Grupo 201)

Dra. Maria Valentina Navárez Terán

30 de abril del 2024

INTRODUCCIÓN A LA PROBLEMÁTICA

También conocido como machine learning no supervisado, de acuerdo con IBM el aprendizaje no supervisado emplea algoritmos de machine learning para analizar y agrupar conjuntos de datos no etiquetados, es decir, datos sin una clase designada. Estos algoritmos funcionan por medio de descubrir patrones ocultos y agrupan los datos sin la necesidad de que exista intervención humana alguna. Debido a que estos algoritmos son muy buenos para descubrir similitudes y diferencias entre la información hace que sean muy útiles a la hora de realizar el análisis de exploración de datos, realizar estrategias de venta cruzadas, segmentar a los posibles clientes, también se pueden emplear incluso en el ámbito de reconocimiento de imágenes.

Algunas de las aplicaciones, así como áreas en donde se pueden aplicar algoritmos de aprendizaje no supervisado son en las secciones de noticias, en donde ya existen ejemplos puntuales como lo es el caso de Google News, el cual emplea aprendizaje no supervisado para catalogar las diferentes noticias con alguna etiqueta específica. También se emplean dentro del área de visión artificial, ya que como se explicó anteriormente, los modelos de aprendizaje no supervisado se desempeñan muy bien para reconocer objetos. Otra área es la detección de anomalías, ya que los modelos de aprendizaje no supervisado son capaces de analizar grandes cantidades de datos y descubrir puntos atípicos dentro de un conjunto de datos. También sirven para detectar perfiles de personas, es decir, ubicar y reconocer ciertas características que tienen en común un grupo de personas para agruparlos juntos, esta aplicación del aprendizaje no supervisado es en la que estaremos trabajando en este reporte más adelante.

De acuerdo con la Escuela de Salud Pública de México, dentro del ámbito de salud el avance tecnológico en herramientas de inteligencia artificial, como lo es el machine learning ha adquirido aún más importancia con el paso del tiempo. Dentro del área de salud el machine

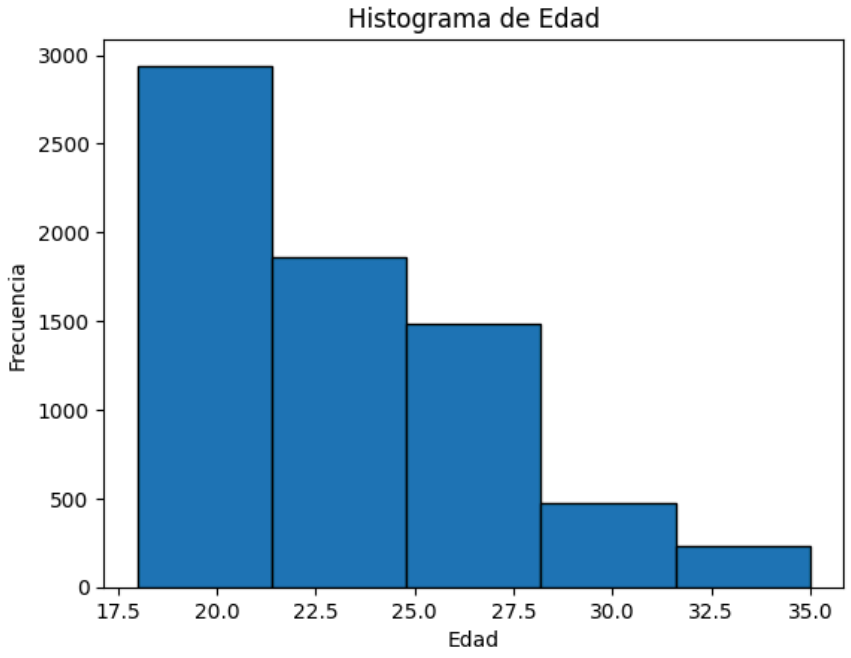
learning se puede emplear para la interpretación de resultados clínicos, así como el apoyo para la toma de decisiones médicas, e incluso para proporcionar información más clara y concisa tanto para los pacientes como para los médicos permitiendo así una atención médica mucho más personalizada que antes. Para este reporte se optó por realizar un modelo de aprendizaje no supervisado basado en la salud mental con la intención de catalogarla en diferentes grupos de acuerdo con las características de los individuos. Esto debido a que consideramos que la salud mental es un tema muy sensible y que solo recientemente se comenzó a tratar de manera correcta, por lo que necesita mucha más difusión y es necesario conocer este problema aún mejor, por lo que emplear un modelo que nos ayude a agrupar a las personas de acuerdo con su nivel de salud mental nos podría ayudar a comprender mejor sus antecedentes.

El Instituto Mexicano del Seguro Social define a la salud mental como el estado de equilibrio que debe de existir entre las personas y el entorno socio-cultural que los rodea. Esto incluye el bienestar emocional, psíquico y social y dependiendo del estado de la salud mental de una persona puede tener repercusiones tanto positivas como negativas en cómo piensa, siente, actúa y reacciona una persona ante situaciones de estrés. La salud mental es de suma importancia, pues se considera la base para el bienestar y correcto funcionamiento de una persona y su comunidad.

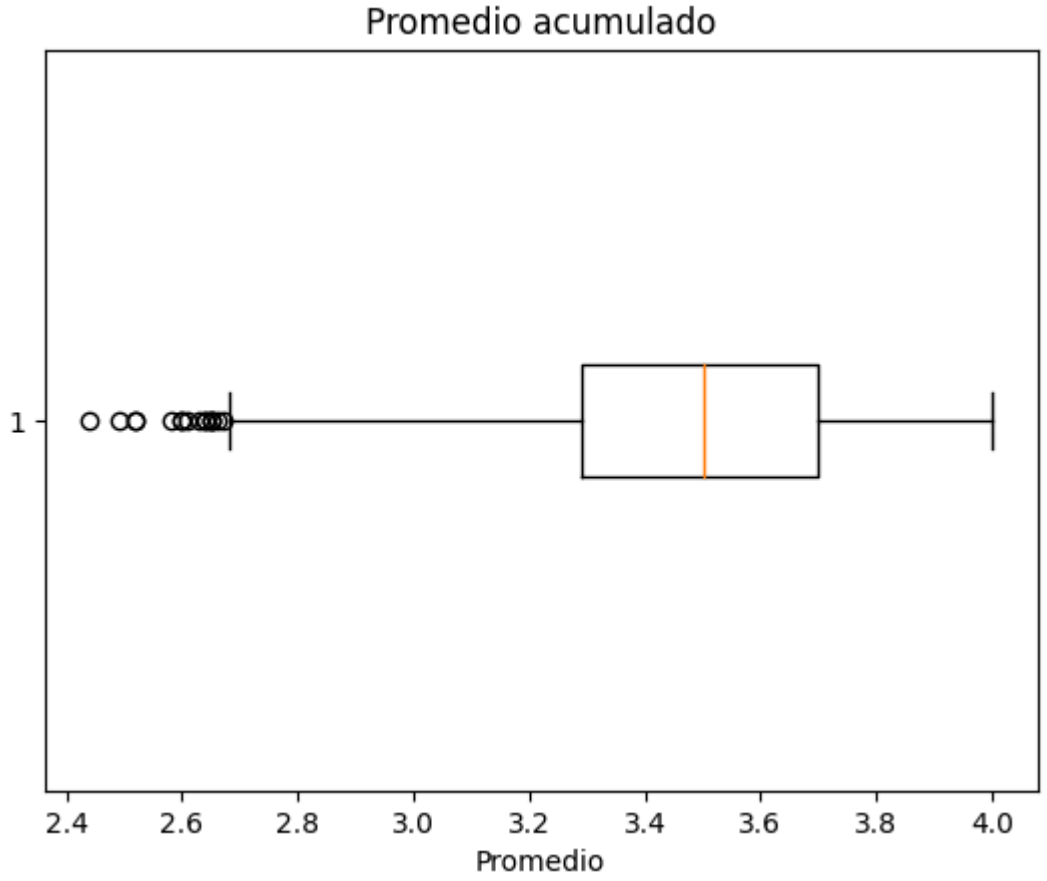
Para este entregable se optó por emplear una base de datos de Kaggle nuevamente, pero esta vez enfocada en la salud mental, la base de datos se obtuvo a través del siguiente enlace (https://www.kaggle.com/datasets/sonia22222/students-mental-health-assessments?resource=download). Este dataset tiene una forma inicial de 7022 registros, con un total de 20 columnas en donde se emplean variables como edad, género, promedio acumulado, carrera de estudios, nivel de estrés, ansiedad y depresión dentro de una persona, estrés financiero, el número de créditos que está cursando actualmente, entre otras variables relacionadas con la

salud mental, después de un análisis de las columnas encontramos que todas las variables podrían tener relación y ser útiles para definir grupos de salud mental, por lo que optamos por mantenerlas a todas. Este dataset no venía en perfectas condiciones, pues con más de 7000 registros era evidente que algunos de ellos iban a tener que ser limpiados, por lo que para la parte de limpieza del dataset se optó por encontrar los valores nulos de la base de datos y contarlos por variable, para nuestra sorpresa encontramos que de los 7022 registros solamente 12 de promedio acumulado contenían valores nulos, y solamente 15 registros de uso de sustancias dañinas contenían valores nulos, por lo que al representar solamente al 0.38% de los datos totales optamos por eliminarlos y no complicarnos con otro proceso más sofisticado. Al eliminarlos terminamos con 6995 registros y 20 columnas en nuestro dataset, pero con la base de datos completamente limpia y lista para empezar la fase de exploración.

Para la fase de exploración optamos por dividir el dataset en dos, un dataset que contuviera las columnas con valores numéricos, y un dataset que contuviera las columnas con valores categóricos, esto debido a que cada tipo de dato puede ser representado mejor con diferentes gráficos. Una vez separados optamos por realizar estos diferentes gráficos, los resultados obtenidos se muestran a continuación:

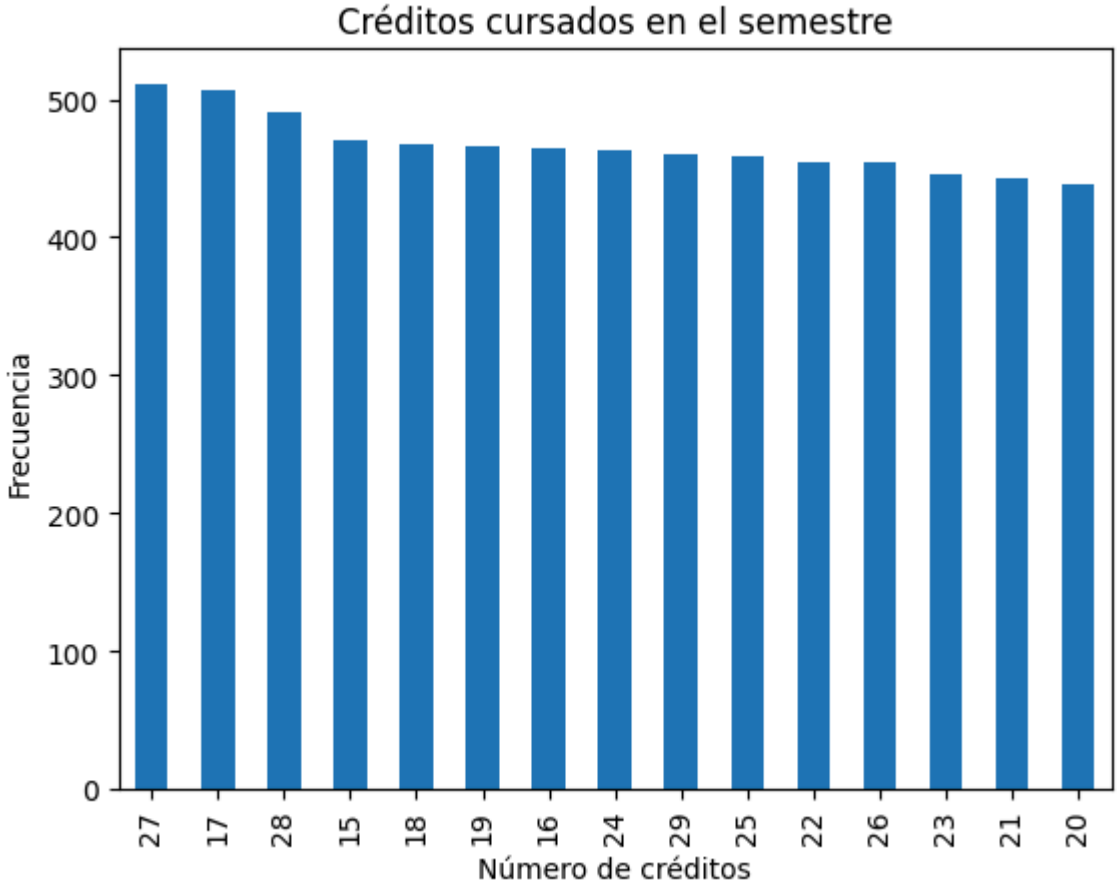


Para este histograma de edad se puede observar que la gran mayoría de las personas encuestadas en este dataset se encontraban en un rango de 18 hasta 25 años, esto tiene sentido, debido a que se menciona que están cursando una carrera universitaria y ese rango de edad es el normal para un universitario, también se puede observar que se tienen bastantes registros desde 25 hasta 28 años, lo cual es un poco sorprendente, pero suponemos que puede deberse a aquellas personas que se cambian de carrera una vez iniciada alguna, o que han reprobado años, el grupo menos común de personas se encuentra entre 32 y 35 años.



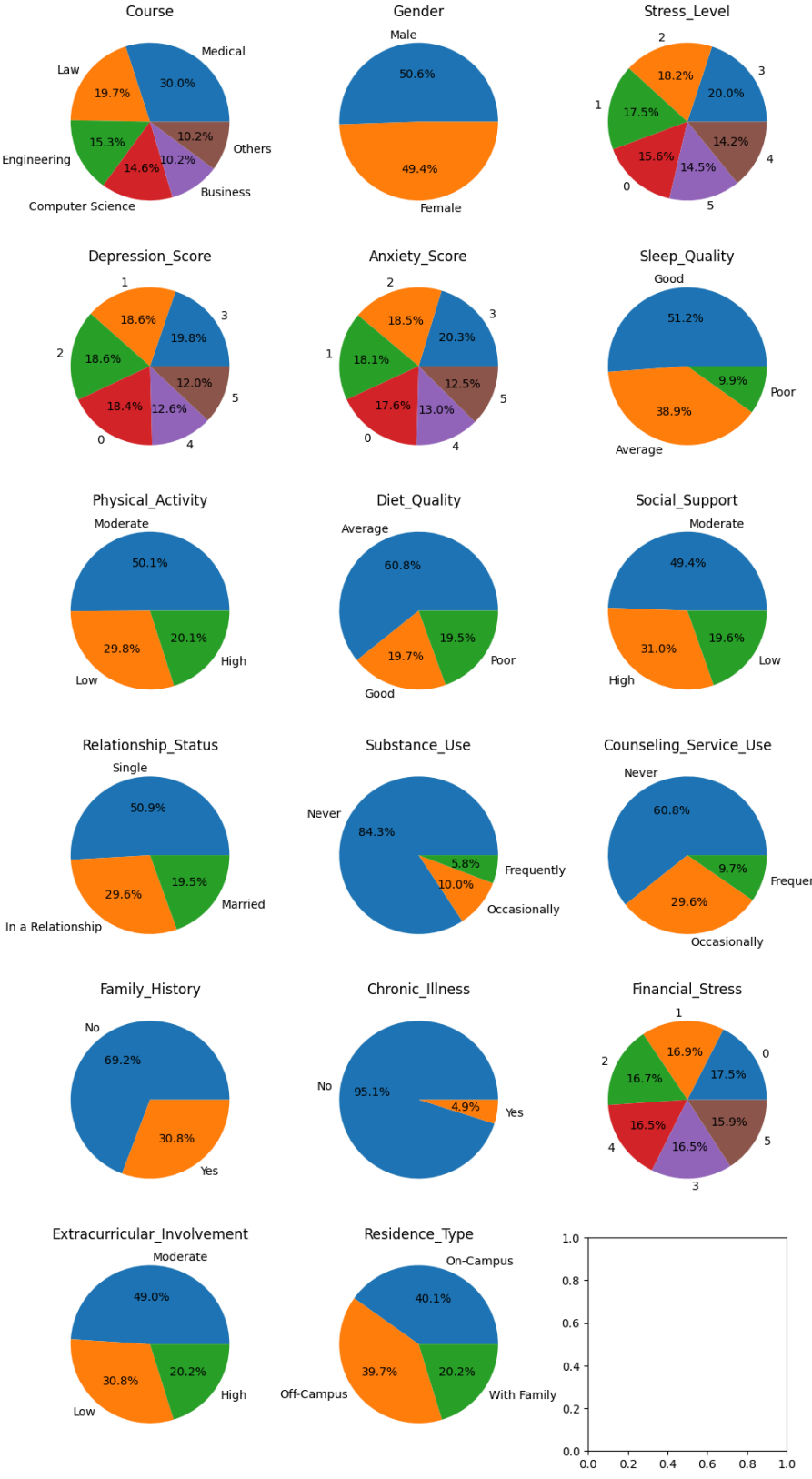
En este diagrama de cajas y bigotes se puede observar que aparentemente la escuela tiene un valor máximo de calificación de 4, ya que ningún registro de alguno de los alumnos sobrepasa este límite superior, mientras que el límite inferior se encuentra aproximadamente en 2.7, con una mediana de 3.5 aproximadamente y se tiene una gran cantidad de valores

atípicos inferiores, es decir, existen muchos registros de personas que se están desarrollando muy por debajo de lo esperado en su escuela.



Como última variable numérica dentro del dataset tenemos la cantidad de créditos que están cursando los alumnos en el semestre actual, esta gráfica de barras fue sin duda una de las más interesantes, ya que se encuentra bastante balanceada, tanto que en un dataset de casi 7000 registros el rango de esta variable es de aproximadamente 50 personas, donde el número de créditos que más veces se repite es el 27, seguido del 17, mientras que el número de créditos que menos veces se repite son el 21 y el 20. El número mayor de créditos cursados es de 29, mientras que el número menor de créditos cursados es de apenas 15, casi la mitad que el mayor.

Posterior a este análisis de variables numéricas, procedimos con el análisis de las variables categóricas, donde optamos por realizar diagramas de pastel para observar las proporciones de las respuestas, un resumen de los resultados obtenidos se muestra a continuación:



Donde se puede apreciar una gran cantidad de información valiosa, por ejemplo:

- Los cursos que más se repiten dentro de los alumnos (30%) son los de medicina, seguidos de derecho (20%), ingeniería (15%), ciencias de la computación (15%) y negocios y otros son los cursos que menos alumnos tienen de acuerdo con este dataset (10% cada uno).
- La proporción de hombres y mujeres está casi completamente balanceada, solamente habiendo 0.8% más hombres que mujeres.
- El nivel de estrés más común en los alumnos es el 3 (en una escala de 0 a 5), seguido de 2, 1, 0, 5 y finalmente 4.
- El nivel de depresión más común en los alumnos también es el 3 (20%), seguido de 1, 2, 0, 4 y finalmente 5.
- El nivel de ansiedad más común en los alumnos es igualmente el 3 (también con 20%), seguido de 2, 1, 0, 4 y 5.
- La calidad de sueño de los alumnos no es para nada mala ya que el 90% de los alumnos registraron que tienen una calidad de sueño buena o promedio (51% y 39% respectivamente), mientras que solamente el 10% de los alumnos consideró que su calidad de sueño era mala.
- La actividad física realizada por los alumnos es en su mayoría moderada (50%), mientras que un 30% tiene malos hábitos de actividad física y el 20% restante tiene buenos hábitos de actividad física.
- En un 80% de los alumnos la calidad de su dieta es por lo menos aceptable, pues un 61% de las personas dice que la calidad de su dieta es promedio, mientras que un 19% comenta que tienen buena dieta, esto es interesante, pues un porcentaje muy parecido (20%) se le atribuye a aquellas personas que tienen buenos hábitos de actividad física. Por otro lado, el 20% restante admite que la calidad de su dieta es mala.

- En cuestión de soporte social, un 49% de los alumnos considera que su soporte social es moderado, mientras que el 31% piensa que es alto y el 20% restante piensa que es bajo.
- Más de la mitad de los alumnos (51%) se encuentran completamente solteros, mientras que un 30% se encuentran en una relación y sorprendentemente un 19% de los alumnos se encuentran incluso casados.
- En términos de consumo de sustancias dañinas, un increíble 84% de los alumnos nunca han consumido ningún tipo de sustancia dañina para el cuerpo, mientras que un 10% de los estudiantes lo hace de manera ocasional, con un 6% de alumnos que consume sustancias dañinas frecuentemente.
- En uso de servicios de apoyo, un 61% de los alumnos nunca los han usado, mientras que un 30% de los alumnos los emplea de manera ocasional y el 9% restante los usa frecuentemente.
- Al preguntarles si en su familia existía historial de una mala salud mental un 69% de los alumnos contestaron que no, mientras que el 31% restantes dijeron que sí.
- Un 95% de los alumnos no presentan ningún tipo de enfermedad crónica, mientras que el 5% restante, sí que tienen algún padecimiento crónico.
- En cuestión de un estrés financiero los alumnos presentaron respuestas sumamente balanceadas, ya que el mayor porcentaje se encuentra en un nivel 0 (17.5%), mientras que el menor porcentaje de estrés financiero se encuentra en el nivel 5 (15.9%), es decir, solamente una diferencia de 1.6% total en las respuestas para este ámbito.
- En cuestión de involucramiento en actividades extracurriculares, un 49% considera que se involucra moderadamente, un 31% considera que su involucramiento es bajo, mientras que un 20% considera que se involucra bastante en actividades extracurriculares.

- Por último, el tipo de residencia de los alumnos se distribuye en un 40% para aquellos alumnos que residen dentro del campus, un 40% de los alumnos igualmente reside fuera del campus, y el 20% de los alumnos restantes reside con sus familiares.

Estos son los análisis que se pueden obtener por medio de las gráficas de las variables en el dataset. Se puede observar que todas las variables pueden llegar a guardar algún tipo de relación con la salud mental de los alumnos e inclusive se pueden notar ciertas correlaciones entre variables, como lo pueden ser la calidad de la dieta y el nivel de actividad física realizada.

METODOLOGÍA

Una vez realizada la limpieza de los datos y el análisis exploratorio de los datos procedemos a realizar los modelos de agrupamiento. Sin embargo, fue necesario realizar un paso previo que a nuestro parecer podría ser beneficioso para los modelos. La forma en la que los algoritmos de aprendizaje no supervisado funcionan es por medio de la agrupación de datos basados en similitud, y la forma en la que miden la similitud es por medio de la distancia, como lo puede ser la distancia euclidiana, es por eso que optamos por convertir todas las variables con datos categóricos en datos numéricos ordinales que mantuviera el significado, consideramos que esto podría ayudar a los modelos a definir mejor una distancia adecuada, para eso se empleó la librería pandas con ayuda de la función replace() y a cada respuesta de una columna se le asignó un valor numérico. La variables del dataset quedaron de la siguiente manera:

```
Age: Edad del individuo
Course: Carrera de estudio del individuo (1 = Business, 2 = Computer Science, 3 = Engineering, 4 = Law, 5 = Medical, 99 = Others)
Gender: Género autoasginado del individuo (1 = Male, 2 = Female)
CGPA: Promedio acumulado del individuo
Stress Level: Nivel de estrés del individuo
Depresion Score: Nivel de depresión experimentado por el individuo
Anxiety Score: Nivel de ansiedad experimentado por el individuo
Sleep Quality: cálidad del sueño de los individuos (1 = Poor, 2 = Average, 3 = Good)
Physical Activity: Nivel de actividad física (1 = Low, 2 = Moderate, 3 = High)
Diet Quality: Cálidad de la dieta del individuo (1 = Poor, 2 = Average, 3 = Good)
Social Support: Nivel de apoyo social recibido por el individuo (1 = Low, 2 = Moderate, 3 = High)
Relationship Status: Estado Civil del individuo (1 = Single, 2 = In a Relationship, 3 = Married)
Substance Use: Uso de sustancias como alcohol, cigarros o drogas (1 = Never, 2 = Occasionally, 3 = Frecuently)
Counseling Service Use: Uso de los servicios de asesoramiento (1 = Never, 2 = Occasionally, 3 = Frecuently)
Family History: Si el individuo tiene familiares con historial en problemas de salud mental o no (0 = No, 1 = Yes)
Chronic Illness: Si el individuo tiene alguna enfermedad crónica o no (0 = No, 1 = Si)
Financial Stress: Nivel de estrés financiero del individuo
Extracurricular Involvement: Nivel de involucramiento en actividades extracurriculares (1 = Low, 2 = Moderate, 3 = High)
Semester Credit Load: Número de créditos que el individuo toma en el semestre
Residence Type: Tipo de residencia del individuo (1 = Off-Campus, 2 = On-Campus, 3 = With Family)
```

Una vez este paso fue realizado se puede proceder a la creación de los modelos de aprendizaje no supervisado, en total se aplicaron dos modelos, esto con el objetivo de hacer una comparativa entre las agrupaciones de diferentes modelos. Como método visto en clase se optó por emplear k-means, el cual se encarga de seleccionar un número k de centroides mediante los cuales se calcula la distancia entre ellos y todos los datos, aquellos datos más cercanos a cada centroide pertenecen al imss grupo, después se calcula la distancia promedio de cada grupo y se mueve el centroide dependiendo de esa distancia, repitiendo el proceso, el centroide se deja de mover después de determinadas iteraciones o cuando ya no se presenten cambios en los grupos. Por otro lado, como método ajeno a la clase se optó por utilizar k modes, el cual es otro método de aprendizaje no supervisado que se encarga de la clasificación de grupos de datos por medio de elegir k observaciones de forma aleatoria y se usan como clusters, posterior a esto se calcula la desemejanza de los datos a estos clusters y se asignan al cluster más cercano, finalmente se deciden nuevas modas y se repiten los pasos anteriores hasta que ya no haya ningún reasignamiento, es decir, funciona de forma muy similar a k-means, la diferencia radica en que k modes se especializa más en datos categóricos.

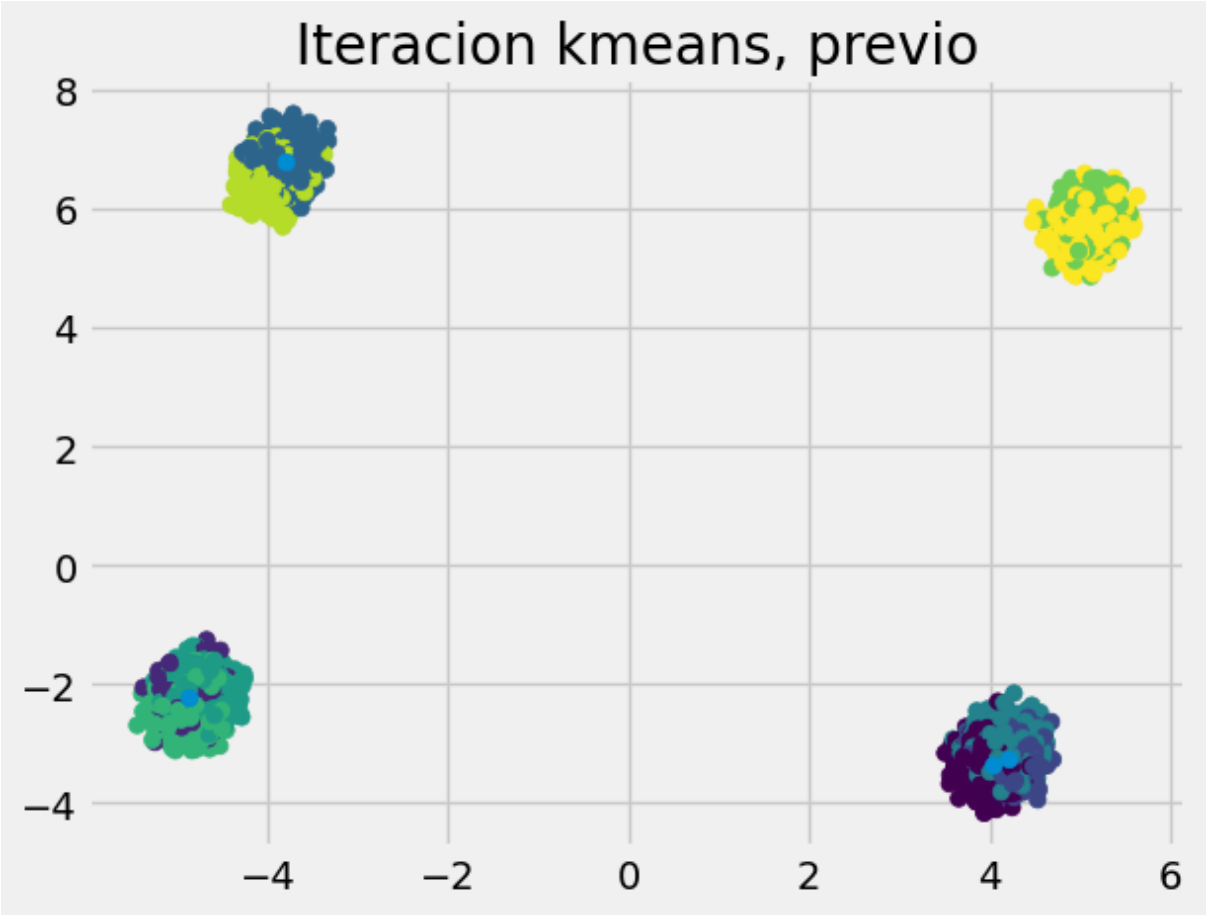
Para realizar el modelo k means fue necesario importar las librerías y módulos necesarios, siendo el más importante de todos KMeans de sklearn. Una vez importadas las librerías se definió una función que se encargaría de graficar los clusters para cada iteración en donde los centroides se iban ajustando dependiendo del cálculo de las distancias obtenidas en la iteración anterior (el código se puede revisar en el anexo). Posterior a esto se prueba el modelo KMeans con 11 clusters con el objetivo de tener la suficiente información para definir el número óptimo de clusters que deben existir para optimizar el modelo, se grafican los clusters con la función creada anteriormente y se observa que el resultado está muy lejos de ser adecuado. Esto se debe a que el número de clusters creados fue erróneo, para conocer el número correcto fue necesario emplear el método del codo y respaldarlo con el coeficiente de siluetas, en donde se descubrió que el número adecuado de clusters es 4, se modificó el modelo para que solamente presente estos clusters y los resultados obtenido superaron con creces a los anteriores, con esto el modelo k means está completado.

Por otro lado, para realizar el modelo k modes se tomó un enfoque similar, antes que nada fue necesario importar las librerías necesarias, en donde la más importante para este caso fue KModes de sklearn. De igual manera que en el modelo anterior se definió una función encargada de graficar los clusters para cada iteración y de forma muy parecida se calcula los cambios en los centroides dependiendo del promedio de distancias de los datos a los centroides en la iteración anterior, y para evitar modificar el dataset original se optó por crear una copia de este en otra variable y en base a ella obtener los clusters. Esta vez iniciamos con los 4 clusters directamente porque la morfología de los datos es la misma y en el método anterior se demostró que es el número de clusters adecuado, se entrena el modelo con este hiperparámetro y se obtiene el resultado, el cual medimos su desempeño con el método del coeficiente de siluetas, para este modelo, el desempeño resultó bastante malo, pues no logró designar los grupos adecuadamente con el método anterior, pero más sobre eso se verá en el

apartado de resultados en el que vamos a comparar ambos modelos y explicar porqué el desempeño de k modes deja mucho que desear.

RESULTADOS:

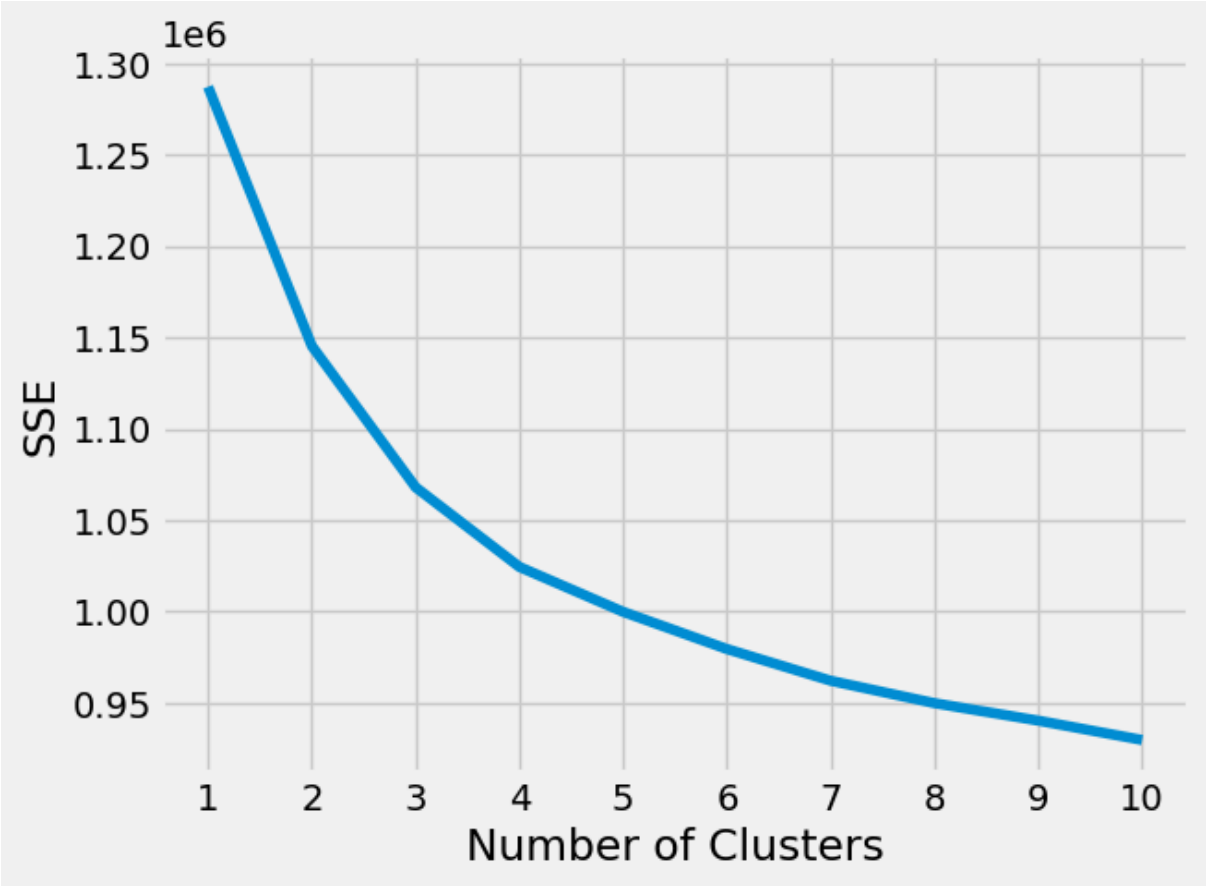
Posteriormente, tras haber realizado el entrenamiento preliminar para el modelo K Means para poder determinar cuál número de clusters es el indicado para este modelo, se graficó el resultado con 11 clusters diferentes.



Como se puede observar, la forma de los datos después de haber aplicado la reducción de dimensiones por PCA, se divide en 4 grupos principales, sin embargo, tras haber realizado las iteraciones, se está mostrando un modelo entrenado con 11 clusters, lo que hace que los grupos de datos están demasiado mezclados entre sí. Esto vuelve la interpretación del modelo

insignificante, pues no hay una manera clara de verificar la separación de los datos de manera clara.

Posteriormente, estos resultados del entrenamiento, como lo son los valores de las distancias euclidianas, son guardados en un conjunto de datos, para posteriormente procesarlos y evaluar el desempeño del modelo según van cambiando el número de clusters.

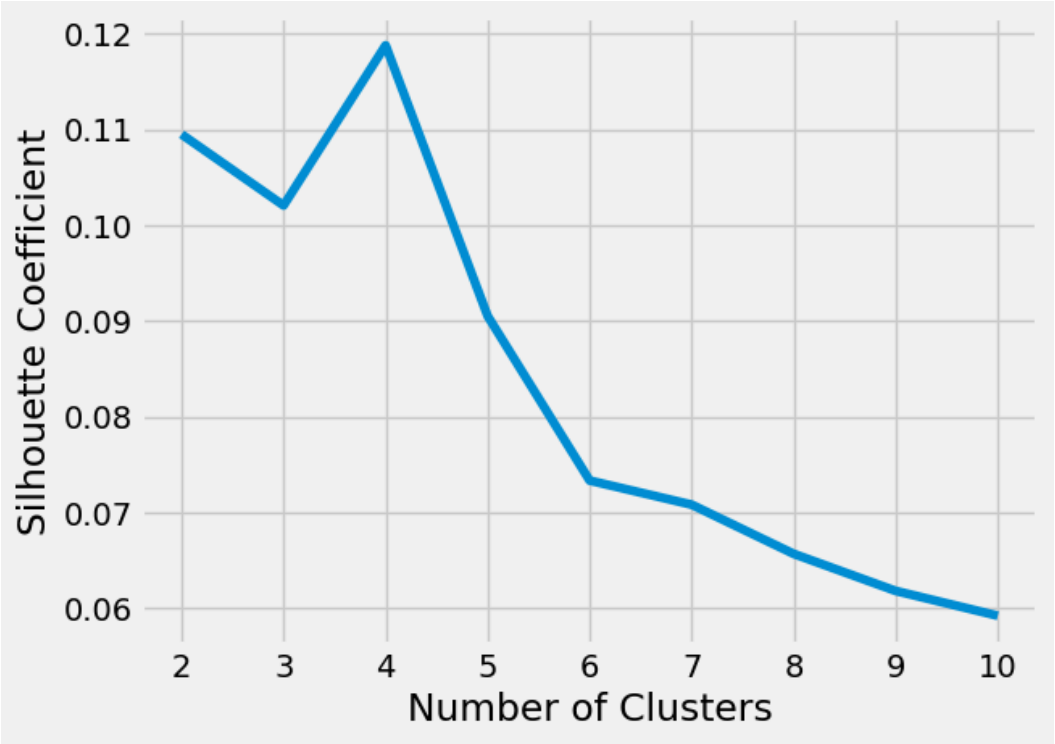


Como se muestra en la gráfica, se compara la suma del cuadrado de las distancias euclidianas, con respecto al número de clusters, como se puede observar, se puede aplicar el método del codo, para determinar cuál es el número de clusters que más se ajusta al modelo.

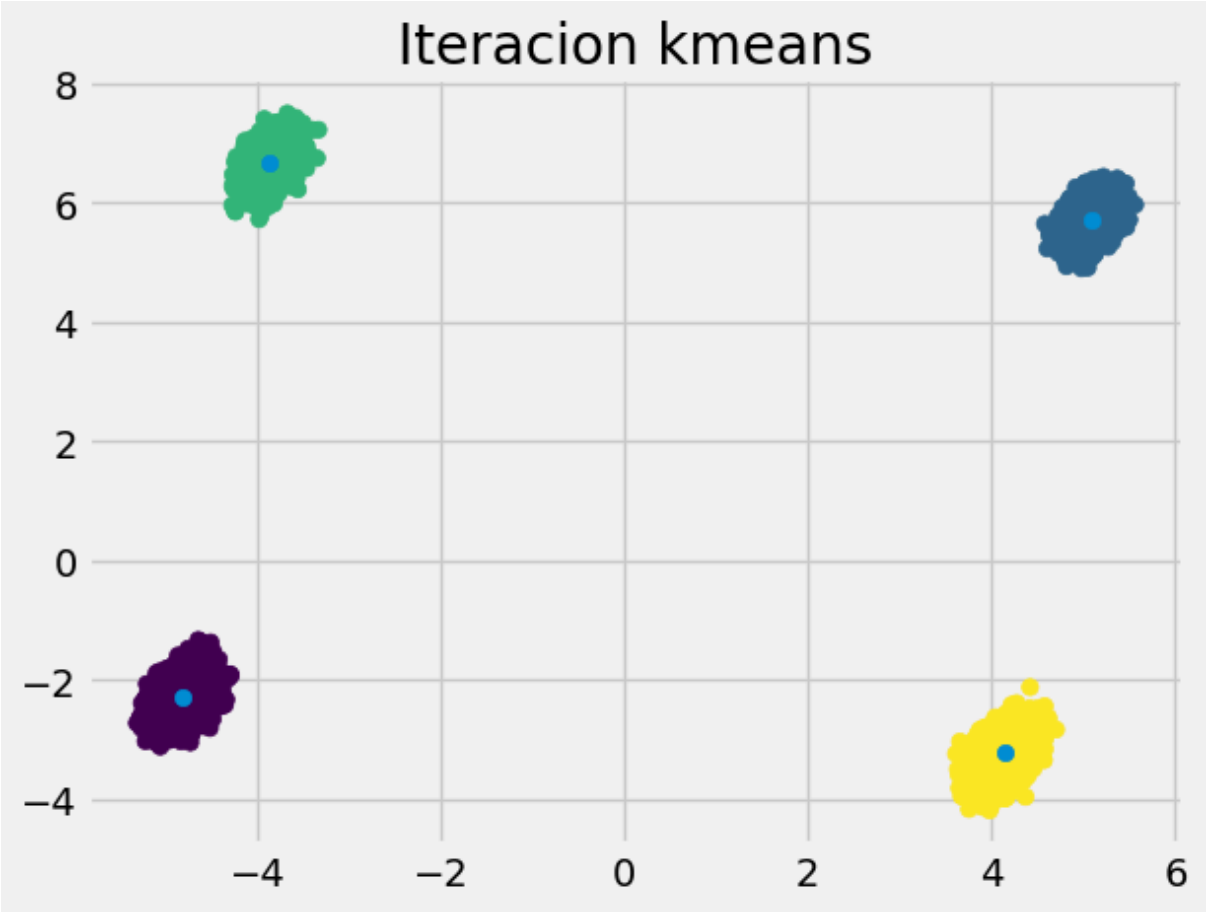
```
[ ]  kl = KneeLocator(range(1, 11), sse, curve="convex", direction="decreasing")
     kl.elbow
```

```
4
```

En este caso, a través del método del codo, se obtuvo que el número de clusters ideal es 4. Si bien esto era algo que se podía ver a simple vista revisando la morfología de los datos, el confirmar esta idea es crucial para sustentar la eficiencia del modelo.



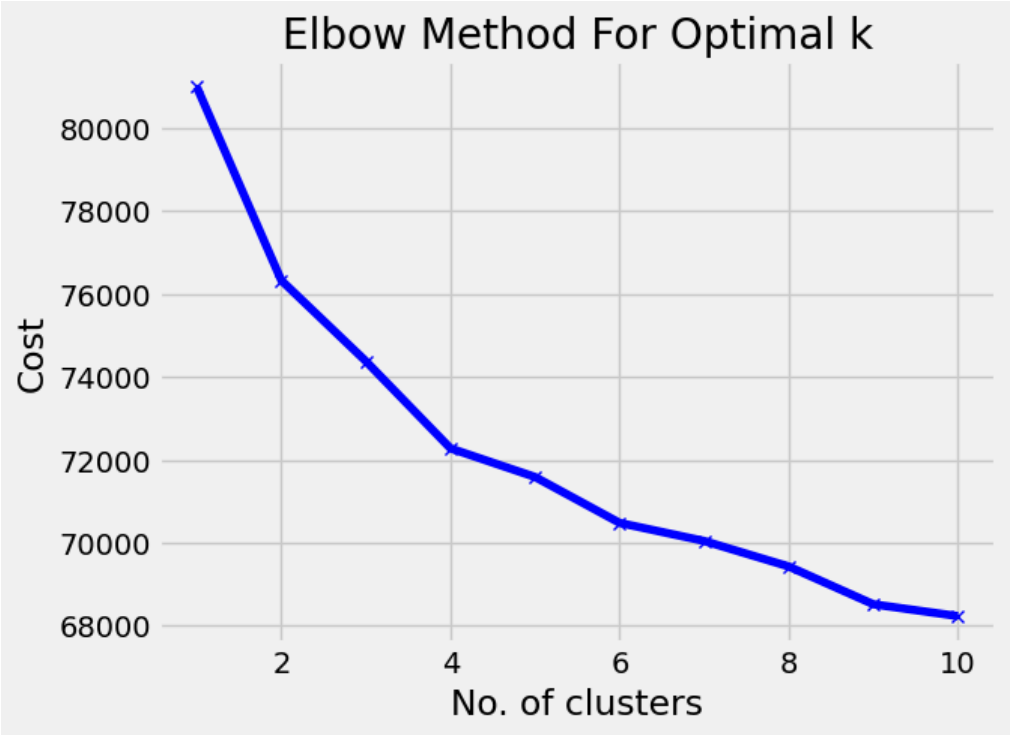
Como se puede observar, de igual manera, la cantidad ideal de clusters es 4, sin embargo, este tiene un aumento en el coeficiente de siluetas para 4 clusters, por lo que se sustenta la idea de utilizar ese número de agrupaciones, así como corrobora lo analizado en el método del codo.



Este gráfico representa el modelo entrenado K-means con 4 clústers, en donde los puntos azules representan los centroides y cada cluster está identificado con un color diferente. Como se puede observar, las agrupaciones están representadas de manera clara, por lo que se puede determinar que a partir de los datos analizados por el modelo, que existen 4 diferentes tipos de alumnos basados en las características de su salud mental. Posteriormente, se podrían realizar diferentes análisis para determinar cuáles son las similitudes y diferencias entre cada tipo de estudiantes.

Como segundo método, se decidió implementar el modelo K-Modes, el cual, a diferencia de K-Means, toma la estadística descriptiva de la moda para encontrar las similitudes entre cada agrupación. Esto puede ser benéfico al momento de trabajar con un set de datos que trabaje con variables cualitativas. Para este modelo no se realizó una visualización previa a la

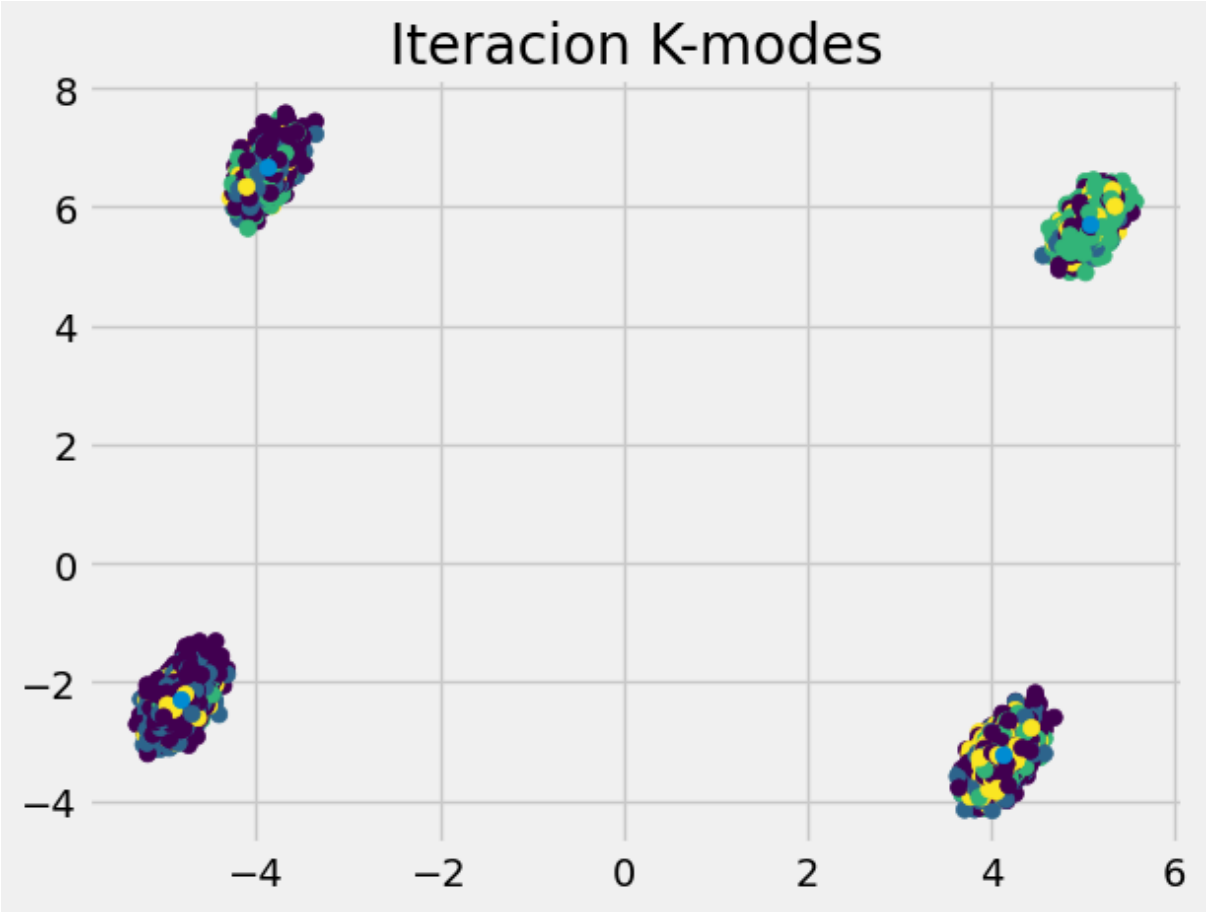
selección de clusters óptima, puesto que la morfología de los datos es la misma. Por lo que después de entrenar el modelo con las 10 primeras iteraciones, se graficó la relación de cambio entre el costo, que para este tipo de clustering, se utiliza la distancia Hamming, y el número de clusters.




```
kl_modes = KneeLocator(range(1, 11), cost, curve="convex", direction="decreasing")
kl_modes.elbow
```

```
4
```

Posteriormente, con ayuda tanto del gráfico se puede observar cómo es que existe un punto de inflexión llegando a 4 clusters, sin embargo, la interpretación visual no puede ser un resultado definitivo, por lo que se añade a la toma de decisión el comando anexado, que determina de igual manera cómo es que 4 es el número de clusters indicado.



Tras haber obtenido el resultado anterior, se graficó el modelo K-modes. Como se puede apreciar, a diferencia del modelo K-means, este tiene una segregación de datos mucho menos clara. Tras un breve análisis, se buscó obtener el coeficiente de siluetas para procurar saber más al respecto.

```
[102] # Calcular el silhouette score para este modelo
      modes_silhouette_score = silhouette_score(data, clusters)
      # Mostrar los resultados
      print('El silhouette score para el modelo de Kmodes  es', modes_silhouette_score)
```

Como se puede observar, el coeficiente de siluetas es muy cercano a 0, esto quiere decir que en promedio, los datos se sitúan en la frontera entre 2 o más centroides de los clusters. Esto es algo que puede llegar a ser contraproducente, pues no permite separar de manera clara aquellas agrupaciones de alumnos, en este caso. Para poder solucionar este problema, se descubrió que el algoritmo K-modes es utilizado para variables cualitativas, por lo que

elementos como el promedio acumulado de los estudiantes, así como la edad o la deuda escolar pueden llegar a afectar en el desempeño del modelo. Una posible solución sería eliminar estas variables, quedando únicamente con variables del tipo categórica o cualitativa, Sin embargo, para mantener la integridad del conjunto de variables y así poder tomar todas estas características en cuenta, se puede implementar en su lugar el modelo K-prototype, que es capaz de realizar una distinción entre las variables numéricas y las variables cualitativas.

CONCLUSIONES INDIVIDUALES

Daniel Eduardo: En mi opinión esta evidencia resultó un poco más demandante que la anterior enfocada meramente en aprendizaje supervisado. Esto debido a que los métodos de aprendizaje supervisados fueron más sencillos de comprender y explicar ya que tener una clase que predecir los hace más entendibles. Por otro lado, la particularidad de los métodos de aprendizaje no supervisados es que crean los clusters sin intervención humana, por lo que la manera en la que se dividen los clusters no siempre queda del todo clara. Sin embargo, he de decir que este tipo de aprendizaje me parece mucho mejor a la hora de enfocarlo en el área de la salud, ya que como tal no arroja un resultado que pueda ser evaluado como verdadero o falso, sino más bien asigna a un dato a un cierto grupo. Esto puede llegar a ser de verdadera utilidad para la comprensión de patrones que incluso pueden llegar a ser imprevisibles para los humanos. A pesar de que mi postura sea más conforme para este tipo de machine learning, me veo en la necesidad de recalcar el mismo punto que di para los métodos de aprendizaje supervisado: la información que se obtiene de ellos no deja de ser una predicción o estimación realizada por una máquina. Si, sustentada en estadística y procedimientos que funcionan bastante bien, pero la inteligencia artificial, en especial en ámbitos tan delicados como lo es la salud no deberían usarse como fuentes de información, sino más bien como herramientas de verificación.

Manuel:

A través de la implementación inicial de los modelos de aprendizaje automático no supervisados fue posible detectar la cantidad de agrupaciones que existen de alumnos con diferentes tipos de salud mental en este set de datos, basados en sus características. Fue enriquecedor el poder haber comprendido el orígen del modelo al haber realizado un algoritmo K-means sin el uso de librerías como Scikit Learn, ya que logré una comprensión más profunda acerca de cómo funciona el modelo, así como la manera en la que sus componentes pueden ir cambiando para obtener mejores resultados. Por otra parte, al ser un modelo de aprendizaje supervisado, el resultado obtenido al haber implementado el modelo de inteligencia artificial abre paso a poder realizar otros trabajos de ciencia de datos para conocer a mayor profundidad el cómo es la distribución de los alumnos, cuales pueden ser estrategias para mejorar la salud mental de los estudiantes basados en la agrupación a la que pertenecen, etc. A diferencia del modelo supervisado, encontré la estrategia de clustering mucho más interesante, pues involucra un tema visto anteriormente, como lo es el análisis de componentes principales. Cambiando de tema a la implementación de herramientas de inteligencia artificial en el área de salud, en especial el uso de algoritmos de clasificación no supervisada, considero que tiene el potencial, la IA, de ser una herramienta sumamente útil para el personal de la salud, sin embargo, hay que ser cuidadosos en este tema, puesto que para que exista un buen modelo, además de las comprobaciones estadísticas y la selección adecuada del modelo, la confiabilidad, estructura e integridad de los datos juega un papel crucial al momento de poder generar modelos de calidad.

Valeria:

Mi experiencia con este proyecto ha sido reveladora. Al explorar este tipo de métodos por primera vez, me enfrenté a desafíos y aprendizajes. Uno de los aspectos más importantes de este proyecto es la selección de los algoritmos adecuados según el tipo de datos que estamos

manejando. Inicialmente seleccionamos los métodos de K- means y K-modes, suponiendo que estos métodos tendrán resultados satisfactorios. Sin embargo al observar los bajos coeficientes de silueta y la falta de coherencia en los clusters generados, me percate de la necesidad de un selección más rigurosa y fundamentada de las técnicas de clustering.

Otro aspecto fundamental fue la comprensión de las limitaciones asociadas con la aplicación del Análisis de Componentes Principales (PCA) a conjuntos de datos predominantemente categóricos. Aunque inicialmente consideraba que la reducción de dimensionalidad mediante PCA simplificará nuestros datos, rápidamente reconocí que la falta de variación de distancia entre las categorías podría distorsionar los resultados.

En resumen, este proyecto ha reforzado la noción de que el análisis de datos es un proceso multidisciplinario que requiere una comprensión profunda de los datos y una selección rigurosa de técnicas de análisis

Anexos:

Notebook:

https://colab.research.google.com/drive/1cgZ1ezkdVXi_3qEUXlhkc4nZRmMdQEMn?usp=sharing

Base de datos:

https://drive.google.com/file/d/1CKvM020-eHGErHYtVX1WSGN2mVtz1gy1/view?usp=sharing

Video:

https://youtu.be/7B4k-1XPIVY

Referencias:

Aprilliant, A. (2023, 7 enero). The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical). *Medium*. https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

Bonthu, H. (2021, June 13). *KModes Clustering Algorithm for Categorical data*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/

Estrada, M. I. D. M. C., MASS Paola Victoria López. (n.d.). *Explorando el aprendizaje automático: perspectivas para una enfermería innovadora*. Revista.espm.mx. Retrieved April 30, 2024, from https://revista.espm.mx/nota-explorando-el-aprendizaje-automatico-perspectivas-para-una-enfermeria-innovadora-46

IBM. (n.d.). *¿Qué es el aprendizaje no supervisado? | IBM*. Www.ibm.com. https://www.ibm.com/mx-es/topics/unsupervised-learning

IMSS. (2022). *Salud Mental*. Imss.gob.mx. https://www.imss.gob.mx/salud-en-linea/salud-mental