



**UNIVERSIDAD
AUTÓNOMA
METROPOLITANA**



**Unidad Iztapalapa
División de Ciencias Básicas e Ingeniería
Posgrado en Ciencias (Ciencias y Tecnologías de la Información)**

**“MODELO AVANZADO DE RECONOCIMIENTO DE PATRONES
NO SUPERVISADO Y APLICACIONES”**

TESIS

Que para obtener el grado de
DOCTOR EN CIENCIAS (CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

PRESENTA:

M.C. Daniel Edahi Urueta Hinojosa
Matrícula: 2181800351
ORCID: 0000-0002-8741-6978
Correo electrónico: deurueta@outlook.com

DIRECTOR:

Dr. Pedro Lara Velázquez

CODIRECTOR:

Dr. Miguel Ángel Gutiérrez Andrade

JURADO:

Presidente:

Dr. José Francisco Gómez Aguilar

Secretario:

Dr. Pedro Lara Velázquez

Vocales:

Dra. Bibiana Obregón Quintana

Dra. Hérica Sánchez Larios

Dr. Jorge Enrique Lavín Delgado

RESUMEN

La Inteligencia Artificial (IA) es un área de estudio que se dedica al desarrollo de sistemas con la capacidad de ejecutar funciones que generalmente demandan la inteligencia humana, tales como: el aprendizaje, el razonamiento, la percepción y la comunicación. En el ámbito de la IA, existen diversas clases de algoritmos que tienen la habilidad de aprender a partir de los datos y efectuar tareas particulares. Algunos de estos algoritmos requieren que los datos estén etiquetados previamente, es decir, que se les asigne una categoría, estos algoritmos se denominan de clasificación supervisada y se basan en el aprendizaje a partir de ejemplos. Por otro lado, existen otros algoritmos que no necesitan que los datos estén previamente etiquetados, sino que los clasifican por medio de las características inherentes de cada elemento, estos algoritmos se denominan de clasificación no supervisada o clustering. Esta aproximación es especialmente útil cuando no se tiene un etiquetado previo o se desconoce cómo puede organizarse un conjunto de datos y es necesario clasificarlos. No obstante, también se emplea con fines de comparación y retroalimentación entre los resultados que arroja el modelo y aquellos que determina un especialista, de esta manera, se puede evaluar la calidad y la validez del modelo, así como obtener nuevos hallazgos a partir de una base de datos.

GPT es un modelo de inteligencia artificial creado por OpenAI y lanzado a finales del 2022. Este modelo está basado en el aprendizaje profundo para el procesamiento de lenguaje natural (PLN), que es una rama de la IA que se ocupa de analizar, comprender y generar lenguaje humano. GPT utiliza una arquitectura de redes neuronales llamada transformadores, que le permite capturar las relaciones entre las palabras y las oraciones de un texto. GPT es capaz de producir textos que simulan la reacción humana, es decir, que son coherentes, relevantes y creativos. Actualmente se han explorado sus diversas aplicaciones en diferentes ámbitos culturales, sociales y científicos, como la generación de resúmenes, la traducción automática, la creación de contenidos, la respuesta a preguntas, entre otras.

En este trabajo se detallará el uso de un marco de trabajo que permite encontrar patrones ocultos por medio de GPT cuando se hace una clasificación no supervisada; validándolo en diferentes conjuntos de datos que provienen de diversas fuentes y áreas de conocimiento, determinando así su calidad e innovación.

AGRADECIMIENTOS

"Ni en mis más oscuros sueños imaginé tener que postergar aquello que más anhelaba."

Gracias al Dr. Pedro Lara Velázquez y al Dr. Miguel Ángel Gutiérrez Andrade, por su guía incansable y por sostenerme cuando yo creía todo perdido.

Al Posgrado en Ciencias en Tecnologías de la Información y su cuerpo docente por brindarme las herramientas necesarias para todo el porvenir.

Al SECIHTI, por el apoyo que hizo posible este recorrido.

A mi familia, por ser el rompeolas en mi vida.

A la vida, por derribarme solo para recordarme lo que es importante.

"A lo que se perdió, se quebró, renació, decidió quedarse... y a lo nuevo que, sin buscarlo, trajo sentido a todo."

ÍNDICE

INTRODUCCIÓN	7
OBJETIVOS	10
Objetivo general.....	10
Objetivos específicos.....	10
METODOLOGÍA	11
1. MARCO DE REFERENCIA	12
1.1. APRENDIZAJE COMPUTACIONAL.....	12
1.1.1. Aprendizaje supervisado	12
1.1.2. Aprendizaje no supervisado	12
1.1.3. Aprendizaje semi-supervisado	12
1.2. OPENAI	13
1.2.1. OpenAI.....	13
1.2.2. GPT	13
1.2.3. ChatGPT.....	13
1.2.4. API GPT	14
1.2.5. Prompt.....	14
1.2.6. Tokens	14
1.3. AGRUPAMIENTO/CLUSTERING	15
1.3.1. Algoritmo de k -medias	16
1.3.2. Modelo de Mezclas Gaussianas	18
1.3.3. Agrupamiento espectral.....	20
1.3.4. Análisis de componentes principales	21
1.3.5. Métodos para determinar el número de clusters	23
1.3.5.1. Método del codo	23
1.3.5.2. Resiliencia generalizada	23
2. IMPLEMENTACIÓN DE LOS ALGORITMOS	25
2.1. DESCRIPCIÓN DE LAS HERRAMIENTAS UTILIZADAS	25
2.2. PREPROCESAMIENTO DE LAS BASES DE DATOS.....	25
2.2.1. Transformación de variables categóricas.....	25
2.2.2. Limpieza de las bases de datos	26

2.2.3.	Normalización de las bases de datos	26
2.3.	EVALUACIÓN DEL MODELO.....	26
2.3.1.	Coefficiente de Correlación de Pearson	26
2.3.2.	Coefficiente de silueta.....	27
2.3.3.	Matriz de confusión.....	28
2.3.4.	Validación cruzada	30
2.3.4.1.	Validación cruzada <i>k</i> -fold	30
2.3.4.2.	Validación cruzada repetida de <i>k</i> -fold.....	31
2.3.4.3.	Selección del número <i>k</i> de pliegues	31
2.3.5.	Modelo matricial propuesto para 60 Universidades.....	31
3.	MODELO PROPUESTO DE CLUSTERING AVANZADO	33
4.	INSTANCIAS	35
4.1.	SALUD.....	35
4.1.1.	Hepatitis dataset	35
4.1.2.	Mamografías de cáncer de mama.....	36
4.1.3.	Biopsias de cáncer de mama	37
4.2.	NUTRICIÓN	37
4.2.1.	Cereales.....	37
4.2.2.	Crema comestibles	38
4.3.	ECONÓMICAS	40
4.3.1.	Wine dataset	40
4.4.	INVESTIGACIÓN	41
4.4.1.	60 universidades mexicanas.....	41
4.5.	CLASIFICACIÓN HUMANA PREVIA	43
4.5.1.	Cuadernos	43
4.6.	CLASIFICACIÓN DE RESULTADOS DE CUESTIONARIOS.....	45
4.6.1.	Implementación de nuevos modelos educativos.....	45
5.	RESULTADOS.....	47
5.1.	SALUD.....	47
5.1.1.	Hepatitis	47
5.1.1.1.	Análisis GPT de resultados de Hepatitis de la mejor técnica (GMM).....	49
5.1.2.	Mamografías de cáncer de mama.....	50

5.1.2.1.	Análisis GPT de resultados de Mamografías de cáncer de mama de la mejor técnica (Spectral).....	51
5.1.3.	Biopsias de cáncer de mama	53
5.1.3.1.	Análisis GPT de resultados de Biopsias de cáncer de mama (Gauss).....	55
5.2.	NUTRICIÓN	56
5.2.1.	Cereales	56
5.2.1.1.	Análisis con GPT	57
5.2.2.	Crema comestibles	59
5.2.2.1.	Análisis de los resultados reportados con GPT	60
5.3.	ECONÓMICAS	61
5.3.1.	Vinos.....	61
5.4.	INVESTIGACIÓN	67
5.4.1.	60 universidades mexicanas.....	67
5.4.2.	Resultados de <i>k</i> -medias.....	69
5.4.3.	Resultados de <i>GMM</i>	70
5.4.4.	Resultados de agrupamiento espectral.....	72
5.4.5.	Análisis de resultados con GPT.....	74
5.5.	Clasificación humana previa.....	78
5.5.1.	Cuadernos	78
5.5.2.	Análisis de clasificación con GPT	80
5.6.	Clasificación de resultados de cuestionarios	82
5.6.1.	Implementación de nuevos modelos educativos.....	82
5.6.2.	Determinación de etiquetado con GPT	84
6.	CONCLUSIONES Y APORTACIONES.....	86
6.1.	<i>Conclusiones</i>	86
6.2.	<i>Aportaciones</i>	87
	REFERENCIAS	88

INTRODUCCIÓN

Dentro de la IA, una de su amplia cantidad de ramas es la del reconocimiento de patrones; este campo está relacionado con el descubrimiento automático de regularidades a través del uso de algoritmos de computadora y cómo esas regularidades son útiles para tomar acciones como la clasificación de los datos en diferentes categorías [1]. El reconocimiento de patrones se ha utilizado para efectuar diagnósticos de enfermedades [2][3], identificar objetos con sonares submarinos [4], determinar cuáles clientes pueden ser más fieles [5], entre muchas otras. Los sistemas de clasificación son una forma específica de identificación de patrones que asignan una categoría a un objeto basándose en la similitud de sus atributos, ya sea que se trate de personas, otros organismos vivos u objetos inanimados.

Cuando se tiene un problema, algunas mediciones son denominadas como características, y esa información alimenta a una máquina de reconocimiento de patrones conocida como clasificador. Los clasificadores de reconocimiento de patrones son en la mayoría de los casos entrenados, es decir, se les enseña a ver la característica deseada a partir de datos de entrenamiento etiquetados (aprendizaje supervisado), pero cuando no existen datos etiquetados, otros algoritmos pueden ser usados para clasificar una base de datos que previamente no haya sido estudiada (aprendizaje no supervisado).

La aparición de GPT y chatGPT a finales del año 2022, los cuales son modelos de inteligencia artificial creados por OpenAI basados en el aprendizaje profundo entrenados a partir de millones de datos en Internet y cuya finalidad es el procesamiento de lenguaje natural para producir textos que simulen la reacción humana, estas tecnologías han marcado un hito y revolución para la clasificación supervisada, así como en distintos ámbitos de la sociedad que abarcan desde la ciencia, educación, salud pública, programación y hasta la medicina. No obstante de tener un periodo relativamente corto de vida en el mercado, en la literatura existen numerosos trabajos donde proponen la creación de nuevos modelos de clasificación usando la tecnología GPT, ya sea para mejorar la clasificación dado un conjunto de datos [6], generar un clasificado completo a partir de descripciones generadas de productos [7] o artículos periodísticos [8]. Sin embargo, en los trabajos citados anteriormente es necesario conocer a priori las etiquetas o tener que validar la clasificación por medio de una verificación manual, limitando así la explicación y las relaciones que pueden guardar los elementos de un mismo grupo a lo que saben los expertos. Y aunque actualmente existen algunos esfuerzos por desarrollar modelos que combinen ambas aproximaciones, estas suelen ser muy limitadas [9] o para casos muy específicos [10].

Por otro lado, el acercamiento del aprendizaje no supervisado es distinto, debido a que estos plantean un método de agrupamiento con una menor intervención humana que los clasificadores supervisados y para ello es necesario perfeccionar la técnicas de agrupado bajo el menor filtrado de datos [11]. De esta forma, si se toma en cuenta que la mayoría de los datos del mundo real no tienen una etiqueta de clasificación previamente establecida, este campo de estudio es de vital importancia para clasificar tales bases de datos, porque aquellos elementos que pertenezcan a un mismo grupo será debido a que comparten características en común, las cuales pueden ser obvias o temas de estudio y comparación para los expertos; es

así como se tiene una problemática y se debe a que si los expertos no encuentran las relaciones entre estos elementos, simplemente se omitirán; no porque no existan, sino porque no fueron encontradas.

Lo anterior permite reconocer un claro sesgo en la clasificación no supervisada, la cual no está determinada por el algoritmo de clasificación, sino por el análisis derivado del mismo; creando así la necesidad de desarrollar un modelo de aprendizaje no supervisado avanzado que permita, en lo posible, solucionar esta problemática.

En este trabajo se desarrolla un modelo avanzado para analizar los resultados de una clasificación hecha por métodos de clustering y un análisis para encontrar patrones ocultos por medio del uso de GPT.

La creación de este modelo avanzado para la clasificación no supervisada, brinda múltiples beneficios comprobados en los resultados, tales como:

- Facilitar las comparaciones entre las clasificaciones establecidas por humanos y las generadas por el modelo.
- Permitir hacer reconsideraciones en la clasificación de algunos elementos, particularmente en las bases de datos con elementos previamente clasificados.
- Disminuir el tiempo y el esfuerzo que implica la clasificación e interpretación en cualquier base de datos.
- Aminorar costos al no necesitar implicar un grupo de expertos para generar clasificaciones.
- Descubrir relaciones entre los elementos de un mismo conjunto previamente no consideradas por los expertos humanos.
- Generar de posibles líneas de investigación afines.

La división de este trabajo se detalla a continuación:

En el Capítulo 1, se explica el marco de referencia, abordando los tipos de aprendizaje, las técnicas para determinar el número de clases a usar para clasificar una base de datos, así como los algoritmos de clasificación no supervisada usados como lo son: *k*-medias, modelo de mezclas Gaussianas, clasificación espectral, técnicas para determinar el número recomendado de grupos para clasificar una instancia; del mismo modo, se describen las características de GPT, sus aplicaciones, aspectos a tomar en cuenta, así como conceptos clave para entender su funcionamiento y actualidad.

En el capítulo 2, se detalla la implementación de los algoritmos, así como las herramientas para evaluar el modelo propuesto, representar los datos, mejorar su visualización y analizar los resultados.

En el Capítulo 3, se describe por medio de un diagrama el funcionamiento del modelo propuesto de clasificación no supervisada avanzado.

En el Capítulo 3, se describen los orígenes y características de las instancias empleadas para evaluar el modelo, así como sus formatos de entrada y salida.

En el Capítulo 4, se hace un análisis de los resultados obtenidos del modelo propuesto para cada problema mediante la comparación con trabajos reportados en la literatura.

Finalmente, se presentan las conclusiones, trabajo a futuro y aportaciones derivadas de este trabajo de investigación.

OBJETIVOS

Objetivo general

Desarrollar e implementar un modelo avanzado de reconocimiento de patrones no supervisado que permita resolver aplicaciones diversas.

Objetivos específicos

- Estudiar las características generales de los modelos de clasificación no supervisada.
- Proponer una metodología para el desarrollo y análisis de los resultados del modelo que permita encontrar patrones no reportados anteriormente.
- Evaluar la calidad del modelo en instancias de investigación bien conocidas como: iris, vinos, cáncer de mama y hepatitis; instancias de investigación reportadas recientemente en la literatura, como lo son aquellas derivadas de estudios de PROFECO; así como aquellas generadas por estudios propios.
- Comparar resultados de este modelo con resultados reportados en la literatura.
- Analizar los resultados obtenidos para complementar los reportados anteriormente.

METODOLOGÍA

Para el estudio del estado del arte se realizó un estudio de técnicas y modelos relevantes en la literatura que se enfocan a resolver el problema de reconocimiento de patrones no supervisado. Asimismo, se investigaron los antecedentes y trabajos reportados en la literatura referentes al uso de GPT en el área de clasificación, de esta manera se pudo formular una propuesta que abordó las áreas de oportunidad considerando las carencias que poseen las técnicas y metodologías existentes.

Posteriormente al estudio del arte, se exploraron e implementaron diferentes algoritmos, así como diversas técnicas para resolver el problema de reconocimiento de patrones no supervisado; una vez implementados estos algoritmos se analizaron diferentes métodos para la representación y evaluación de los resultados provistos por el modelo. A partir de los resultados de estas investigaciones se concibieron, adaptaron y desarrollaron los conceptos base necesarios para diseñar un modelo avanzado de reconocimiento de patrones no supervisado.

Para obtener las instancias a usar en la validación del modelo se hizo una investigación a fondo para encontrar aquellas que fueran adecuadas y representativas de diferentes áreas. Derivado de lo anterior, se obtuvieron instancias de económicas: vinos; instancias médicas como lo son: cáncer de mama y hepatitis; instancias de nutrición como: cereales y cremas comestibles; instancias financieras: cuadernos; instancias de investigación como: 60 universidades e instancias generadas: encuestas estudiantiles.

Finalmente, con el fin de comparar la validez de los resultados del modelo desarrollado, se analizó la información reportada en diversos trabajos científicos. A partir de los resultados obtenidos en estos apartados se elaboraron las conclusiones correspondientes.

1. MARCO DE REFERENCIA

La construcción de un sistema clasificador consiste en diseñar un algoritmo que, a partir de un conjunto de objetos, este sea capaz de asignarles una etiqueta a partir de un número de clases preestablecidas o, que a partir del conjunto de objetos, se determine el número de clases recomendables para agruparlos tomando en cuenta sus características comunes [12].

1.1. APRENDIZAJE COMPUTACIONAL

El Aprendizaje Computacional es una especialización de la IA cuyo objetivo radica en la creación de modelos matemáticos que representen los procesos cognitivos y analizando el comportamiento de los algoritmos bajo diversas modalidades de datos de entrada o entrenamientos. Asimismo, se distinguen tres tipos de aprendizaje computacional.

1.1.1. Aprendizaje supervisado

Consiste en encontrar una aproximación de un problema o función objetivo usando datos etiquetados [13], es decir, los ejemplos proporcionados ya tienen clases asociadas previamente. La meta es conseguir que los demás datos que no se usen como ejemplo puedan ser clasificados dentro de las clases de ejemplo.

1.1.2. Aprendizaje no supervisado

A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado no se cuenta con etiquetas o clases predefinidas para los datos, en cambio, el propósito es descubrir la categoría más apropiada para un conjunto de datos no etiquetados. Esto se logra mediante el grado de similitud de las características entre todos los objetos (datos), pudiendo así determinar que, si dos o más objetos pertenecen a una misma categoría es porque no existen diferencias significativas entre sus características [14].

1.1.3. Aprendizaje semi-supervisado

El aprendizaje semi-supervisado contiene una combinación de los aprendizajes supervisado y no supervisado. En adición a los datos sin etiquetar, se provee al algoritmo de algo de información supervisada, no necesariamente a todos los objetos o información de todas las clases. Los métodos semiautomáticos asumen que los datos sin etiquetar y que sean similares deben ubicarse siempre en la misma clase [13].

1.2. OPENAI

1.2.1. OpenAI

De acuerdo con la página de OpenAI, OpenAI es una empresa de investigación y despliegue de IA. Su misión es garantizar que la inteligencia artificial beneficie a toda la humanidad. Entre sus productos más conocidos que han desarrollado se encuentran GPT-4, chatGPT y DALL·E.

1.2.2. GPT

Los Modelos Predictivo de Lenguaje Generativo (GPT, por las siglas en inglés de Generative Pre-trained Transformers) son modelos de inteligencia artificial desarrollados por la empresa OpenAI para el procesamiento del lenguaje natural. Aunque históricamente se han generado distintos modelos bajo diferentes términos y nombres dependiendo la cantidad de los datos con la que fueron entrenados y su novedad, como lo son: text-davinci-003, text-davinci-002, davinci, curie, babbage, ada. Actualmente, GPT se encuentra en su versión 4 bajo el nombre de GPT-4. De acuerdo con la página del desarrollador [15], esta versión es capaz de manejar más de 25,000 palabras de texto, lo que permite casos de uso como la creación de contenido extenso, conversaciones prolongadas y búsqueda y análisis de documentos. Estas capacidades permiten que GPT-4 sea más creativo y colaborativo, ya que puede generar, editar e iterar con los usuarios en tareas de escritura creativa y técnica, como componer canciones, escribir guiones o aprender el estilo de escritura de un usuario. GPT-4 está disponible en ChatGPT Plus y como una API para que los desarrolladores construyan aplicaciones y servicios.

1.2.3. ChatGPT

ChatGPT es un modelo creado por OpenAI basado en GPT-4 que utiliza técnicas de aprendizaje profundo con 175 mil millones de parámetros para producir texto que se asemeja de cerca a la escritura humana. Tiene la capacidad de generar texto que es notablemente similar a la escritura humana [16]. Hasta el momento, se ha aplicado en una gran variedad de temas, como educación [17], finanzas [18], salud pública [19] y artículos de investigación [20]. Aunque se ha observado que reconoce patrones y genera respuestas plausibles [21], dado que la forma en que genera respuestas es mediante suposiciones probabilísticas basadas en un modelo estadístico entrenado con texto extraído de toda la internet, puede dar respuestas incorrectas [22].

1.2.4. API GPT

Para utilizar un modelo GPT a través de la API de OpenAI, es necesario enviar una consulta con las entradas y una clave de API personal, de esta forma se recibirá una respuesta con la salida generada por el modelo.

El punto final de la API de completado tiene una interfaz distinta al punto final de completado de chat. En lugar de que la entrada sea una lista de mensajes, se utiliza una solicitud (*prompt*).

1.2.5. Prompt

Una *prompt* es una entrada o instrucción proporcionada al modelo para generar una respuesta, y esta puede variar en longitud y complejidad

El crear correctamente una *prompt* es de vital importancia, debido a que, frecuentemente la calidad, exactitud y pertinencia de la respuesta generada dependerá de la claridad y especificidad de la pregunta.

Una *prompt* puede variar en longitud y complejidad dependiendo de la interacción deseada. Puede ser tan breve como una sola oración o tan extensa como varios párrafos. La *prompt* puede incluir una pregunta, una afirmación o cualquier otra forma de texto que inicie la conversación.

Un ejemplo de llamada a la API desde Python con su respectiva *prompt* se presenta a continuación:

```
import openai

response = openai.Completion.create(
    model=" gpt-3.5-turbo",
    prompt="Escribe un resumen de la Cenicienta en 500 palabras."
)
```

1.2.6. Tokens

Los tokens pueden considerarse como fragmentos de palabras. Antes de que la API procese los *prompts*, la entrada se descompone en tokens. Debido a que la API de OpenAI trata las palabras según su contexto en los datos del corpus, la cantidad de tokens puede variar en función de diferentes parámetros que van desde la distinción entre mayúsculas y minúsculas, el orden de la oración, hasta el uso de caracteres especiales. En [23] se muestra un ejemplo de la cantidad de tokens que usa una palabra dependiendo su posición en una oración:

- Minúscula en medio de una oración: ' rojo' - (token: "2266")

- Mayúscula en medio de una oración: ' Rojo' - (token: "2297")
- Mayúscula al comienzo de una oración: 'Rojo' - (token: "7738")

La API de OpenAI ofrece diferentes modelos de consulta por medio de *prompts*, cada uno con diferentes precios por token. De acuerdo con su página [24], los precios actuales expresados en dólares, por cada 1.000 tokens son los siguientes:

Tabla 1. Descripción de costos por el uso de GPT-4

GPT-4		
Modelo	Prompt	Completado
8K contexto	\$0.03 / 1K tokens	\$0.06 / 1K tokens
32K contexto	\$0.06 / 1K tokens	\$0.12 / 1K tokens

Tabla 2. Descripción de costos por el uso de GPT-3.5-turbo

GPT-3.5-Turbo	
Modelo	Uso
gpt-3.5-turbo	\$0.002 / 1K tokens

Previo a la creación de un modelo que genere y haga *prompts* a la API de OpenAI, es posible calcular la cantidad aproximada de tokens que se usarán por medio de herramientas que se pueden usar desde un lenguaje de programación [25] o desde un navegador web [26]. Siendo preferible el uso de la primera dada su velocidad de respuesta y actualización.

1.3. AGRUPAMIENTO/CLUSTERING

Un agrupamiento o *cluster* se describe como un grupo o comunidad con homogeneidad de características internas y una separación explícita de los otros. El análisis de información por agrupamiento entra en el terreno del aprendizaje no supervisado y consiste en la búsqueda a través de una gran variedad de grupos. La diversidad de la información obliga a desarrollar técnicas y herramientas que permitan procesar una gran cantidad de datos en el menor tiempo posible.

Los algoritmos de clustering, también conocidos como de agrupamiento, son técnicas que segmentan un conjunto de datos en grupos, de tal manera que aquellos integrantes de un mismo grupo presenten un mayor grado de similitud entre sí en comparación con los

elementos de grupos diferentes [14]. La cantidad de agrupaciones a generar puede estar definido de antemano o bien, ser determinado por técnicas específicas o el propio algoritmo. Desde un punto de vista formal, un algoritmo de agrupación produce una función de mapeo que asocia cada objeto del conjunto de datos a un grupo específico. El proceso de análisis por agrupamiento se caracteriza por un enfoque bidireccional, compuesto por las siguientes etapas [27]:

- Selección o extracción de características representativas de un conjunto de candidatos, también se utilizan transformaciones para generar características novedosas derivadas de las originales.
- Diseño y elección de un algoritmo agrupador la selección de una medida de proximidad y la construcción de una función de criterio, establece la métrica de similitud entre un conjunto de datos.
- Validación de *clusters* empleando unos estándares de evaluación neutrales, sin preferencia por algún algoritmo.
- Interpretación de resultados que se obtiene a juicio de expertos que analizan e interpretan la partición de los datos.

En [12] se hace mención de modelos avanzados para el reconocimiento de patrones como: basados en error cuadrático, tomando como base técnicas de búsqueda combinatoria, teoría de grafos, lógica difusa, agrupamiento de k -medias, agrupamiento de datos secuenciales, entre otros; que en su mayoría no son implementados, pero pueden brindar una pauta para combinarlos y así construir un modelo más robusto. Por ejemplo, recientemente se ha propuesto una técnica de búsqueda combinatoria GRASP para resolver instancias de coloración de gráficas suaves usando en su fase de construcción un algoritmo de agrupamiento por k -medias.

1.3.1. Algoritmo de k -medias

Este algoritmo tiene como objetivo establecer en k grupos un conjunto de datos dado, donde k representa el número total de agrupaciones previamente especificados por el analista o por algún método de selección del número idóneo de clases. Cuando k -medias clasifica los objetos, los objetos que pertenecen a un mismo grupo son lo más similares, mientras que los objetos de diferentes grupos son tan diferentes como sea posible; cada grupo está representado por el centro o medio de los puntos de datos que pertenecen al grupo [28]. El pseudocódigo básico es:

Empezar

1. Elegir al azar k centros de cluster
2. Mientras los puntos dejen de cambiar la asignación a centroides
Asignar cada punto de datos al centro de cluster más cercano
Establecer los centroides del grupo en función de la posición promedio (media) de cada punto centroide
3. Terminar mientras

Fin

Formalmente, consideremos k el número de grupos y n el número total de observaciones. Donde $i = 1, \dots, n$ mientras que $c = 1, \dots, k$. Sea μ_c la media del grupo c . El objetivo de k -medias es minimizar la suma del error al cuadrado sobre todos los grupos denotado por $J(C)$; así, la función objetivo se plantea como:

$$J(C) = \sum_{c=1}^k \sum_{x_i \in c} \|x_i - \mu_c\|^2.$$

Minimizar esta función objetivo es un problema *NP-Duro*, incluso para $k = 2$ [29]. Es por eso que se puede esperar que k -medias, que es un algoritmo glotón en cada iteración, converja a un mínimo local. K -medias comienza con una partición inicial con k grupos y asigna patrones a grupos para reducir el error al cuadrado. Dado que el error al cuadrado tiende a decrecer con un incremento en el número de grupos k (con $J(C) = 0$ cuando $k = n$), se minimiza para un número fijo de grupos [30].

El algoritmo k -medias requiere algunos parámetros especificados por el usuario, tales como [31]:

- El número de grupos k . Si bien no existe un criterio matemático, hay varias heurísticas disponibles para elegir k . Normalmente, k -medias se ejecuta de forma independiente para diferentes valores de k y se selecciona la partición que parece más significativa para el experto en el dominio. Las diferentes inicializaciones pueden conducir a una agrupación final diferente porque k -medias solo converge a mínimos locales. Una manera de superar los mínimos locales es ejecutar el algoritmo k -medias, para una determinada instancia, con varias particiones iniciales diferentes y elegir la partición con el valor más pequeño del error al cuadrado.
- Una métrica. A pesar de que la métrica más utilizada para calcular la distancia entre los puntos y los centros del clúster es la distancia euclidiana, esta métrica se limita a los límites de agrupación lineal lo que significa que el algoritmo será ineficaz si las agrupaciones tienen geometrías complicadas; es por ello que se han usado otras métricas como la distancia de Mahalanobis, la cual es útil para detectar *clusters* hiperelipsoidales [32].

Cuando se usa k -medias es posible que no se logre el óptimo global. Y debido a que no garantiza la mejor solución, el algoritmo se ejecuta para múltiples intentos iniciales y mejora el resultado en cada paso. Sin embargo, su implementación es sencilla y las soluciones son encontradas rápidamente incluso para grandes conjuntos de datos.

Esto se obtiene por medio de una concepción simplificada de cómo se ve la agrupación óptima:

- El "centro del grupo" es la media aritmética de todos los puntos que pertenecen al grupo.
- Cada punto está más cerca de su propio centro de clúster que de otros centros de clúster.
- Se supone que los centros óptimos son aquellos que minimizan la distancia de cada punto desde su centro asignado.

1.3.2. Modelo de Mezclas Gaussianas

Un modelo de mezclas gaussianas (GMM) es un enfoque probabilístico que parte del supuesto de que todos los datos en un conjunto provienen de una mezcla finita de distribuciones gaussianas, cada una con parámetros desconocidos. Se puede considerar que GMM es una generalización del algoritmo k -medias que añade información sobre la estructura de covarianza de los datos y los centros de las distribuciones latentes. El objetivo de un GMM es identificar la mezcla de distribuciones gaussianas multidimensionales que mejor se ajuste a un determinado conjunto de datos. El pseudocódigo básico es:

Empezar

1. Elegir las conjeturas de partida para la ubicación y forma

2. Si no se alcanza la convergencia:

 Para cada punto, encontrar los pesos que codifican la probabilidad de membresía en cada grupo

 Para cada grupo, actualizar su ubicación, normalización y forma con base en todos los puntos de datos, haciendo uso de los pesos

3. Terminar mientras

Fin

Formalmente, sea k el número de clusters y n el número total de observaciones. Donde $i = 1, \dots, n$ mientras que $c = 1, \dots, k$. Se inicializan aleatoriamente los parámetros μ, Σ, π .

Donde μ es la media que representa su centro, Σ es la covarianza que representa su ancho y π es la probabilidad de mezcla que establece cuán pequeña o grande será la función gaussiana. Después se calcula para cada punto x_i la probabilidad γ_i^c de que el punto x_i pertenezca a la clase c con:

$$J\gamma_i^c = \frac{\pi_c \mathcal{N}(x_i | \mu_c, \Sigma_c)}{\sum_{c=1}^k \pi_c \mathcal{N}(x_i | \mu_c, \Sigma_c)}$$

donde $\mathcal{N}(x | \mu, \Sigma)$ describe la Gaussiana multivariable con:

$$\mathcal{N}(x_i | \mu_c, \Sigma_c) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_i - \mu_c)^T \Sigma_c^{-1} (x_i - \mu_c)\right)$$

γ_i^c da para cada punto x_i la medida de la probabilidad que x_i pertenezca a la clase c entre la probabilidad de x_i sobre todas las clases, por eso, si x_i es muy cercana a una Gaussiana c , esta tendrá valores altos de γ_i^c y valores relativamente bajos para cualquier otro caso.

Como segundo paso, para cada cluster c : se calcula el peso total m_c (en términos generales, la fracción de puntos asignados al grupo c) y se actualiza π_c , μ_c y Σ_c usando γ_i^c con:

$$m_c = \sum_{i=1}^n \gamma_i^c$$

$$\pi_c = \frac{m_c}{m}$$

$$\mu_c = \frac{1}{m_c} \sum_{i=1}^n \gamma_i^c x_i$$

$$\Sigma_c = \frac{1}{m_c} \sum_{i=1}^n \gamma_i^c (x_i - \mu_c)^T (x_i - \mu_c)$$

Finalmente, el primer y segundo paso se repiten hasta que se alcanza la convergencia [33].

Como resultado del algoritmo, se tiene que cada grupo se asocia no con una esfera de bordes duros, sino con un modelo gaussiano suave. Al igual que en el enfoque de k -medias, el número de grupos debe conocerse de antemano y este algoritmo a veces puede pasar por alto el óptimo global, por lo que en la práctica se utilizan múltiples inicializaciones aleatorias para elegir las estimaciones iniciales [31].

1.3.3. Agrupamiento espectral

El agrupamiento espectral (Spectral clustering) es una técnica cuyo objetivo es agrupar los datos que están conectados, pero no necesariamente agrupados dentro de los límites convexos. La idea básica es construir un gráfico ponderado a partir del conjunto de datos inicial donde cada nodo representa un patrón y cada borde ponderado simplemente toma en cuenta la similitud entre dos patrones [34]. Es por ello que este algoritmo de agrupamiento puede verse como un problema de corte de gráfico, que puede abordarse mediante la teoría del gráfico espectral. El núcleo de esta teoría es la descomposición del valor propio de la matriz Laplaciana del gráfico ponderado obtenido de los datos. El pseudocódigo es:

Empezar

1. Calcular A , la matriz de afinidad $n \times n$

2. Obtener el sistema electrónico de A :

Calcular los primeros v *eigen*-vectores de su matriz Laplaciana para definir un vector de características para cada objeto:

Establecer $U = n \times v$ matriz que contiene los *eigen*-vectores normalizados de los v *eigen*-valores más grandes de A en sus columnas

3. Aplicar k -medias en el espacio de fila de U para encontrar k clusters

Fin

Formalmente, sea n el número de puntos de datos que se agruparán y $W = [w_{i,j}]_{n \times n}$ la matriz de peso donde cada $w_{i,j}$ es la similitud entre dos puntos de datos. Entonces, un problema de agrupamiento se puede formular en el problema de corte mínimo, es decir:

$$q^* = \arg \min_{q \in \{-1,1\}^n} \sum_{i,j=1}^n w_{i,j} (q_i - q_j)^2 = q^T L q$$

Donde $q = (q_1, q_2, \dots, q_n)$ es un vector para membresías binarias y cada q_i puede ser -1 o 1 . L es la matriz Laplaciana, definida como $L = D - W$, donde $D = [d_{i,i}]_{n \times n}$ es una matriz diagonal con cada elemento $d_{i,i} = \sum_{j=1}^n w_{i,j}$.

Para la agrupación en varias clases, la función objetivo se puede definir como:

$$J_{norm_mc}(q) = \sum_{z=1}^k \sum_{z' \neq z} \frac{C_{z,z'}(q)}{D_z(q)}$$

Donde k es el número de clusters, $q \in \{1, 2, \dots, k\}^n$, $C_{z,z'} = \sum_{i,j=1}^n \delta(q_i, z)\delta(q_j, z')w_{i,j}$ y $D_z = \sum_{i=1}^n \sum_{j=1}^n \delta(q_i, z)w_{i,j}$. No obstante, encontrar eficientemente la solución que minimice la ecuación anterior es bastante difícil en términos computacionales. Por lo cual, una estrategia común es conseguir primero los v *eigen*-vectores más pequeños de la matriz Laplaciana L (excluyendo el que tiene un *eigen*-valor cero), y proyectar los puntos de datos en el espacio de baja dimensión que abarcan los v *eigen*-vectores. Luego, se aplica un algoritmo de agrupamiento estándar, como k -medias a los puntos de datos del cluster en este espacio de baja dimensión [35].

Por lo anterior, el agrupamiento espectral se basa en dos pasos principales: primero incrustar los puntos de datos en un espacio en el que los agrupamientos son más “obvios”, y luego aplicar un algoritmo de agrupamiento clásico como k -medias. La matriz de afinidad A se forma usando un *kernel* como, por ejemplo, el Gaussiano. Para obtener k grupos, se calculan los primeros v vectores principales de A y se aplican k -medias en las coordenadas de la unidad de la norma.

El agrupamiento espectral funciona bien para una pequeña cantidad de elementos y al igual que k -medias y GMM, requiere que se especifique el número de clusters.

1.3.4. Análisis de componentes principales

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad lineal empleada para extraer información de un espacio de alta dimensionalidad proyectándolo en un subespacio de menor dimensión. Su objetivo es preservar las características esenciales de mayor variabilidad en los datos, eliminando aquellas no esenciales con menor variación. Esto se logra mediante un procedimiento estadístico que utiliza una transformación ortogonal para convertir un conjunto de observaciones de variables potencialmente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas, denominadas componentes principales.

A manera de síntesis, las operaciones que realiza el algoritmo son las siguientes [31]:

- **Estandarización.** Se estandarizan los datos de entrada. Este proceso transforma los datos a una escala común eliminando las diferencias de unidades o magnitudes entre las variables.
- **Obtención de eigen-vectores y eigen-valores.** Se calcula la matriz de covarianza de los datos estandarizados. La matriz de covarianza refleja la relación de dependencia entre las variables, cuantificando la variabilidad conjunta de cada par de variables. A partir de la matriz de covarianza, se procede a calcular los *eigen*-vectores y *eigen*-valores asociados. Los *eigen*-vectores representan direcciones en el espacio multidimensional que maximizan la varianza de los datos, mientras que los *eigen*-valores representan la magnitud de la varianza en cada dirección.
- **Selección de los eigen-vectores más relevantes.** Los *eigen*-valores se ordenan de mayor a menor, identificando aquellos que capturan la mayor varianza en los datos, es así como se seleccionan los d *eigen*-vectores que corresponden a los d *eigen*-valores más grandes, donde d representa el número de dimensiones del nuevo subespacio de características que se desea obtener. Es recomendable que, para fines de presentación gráfica, $d \leq 3$.
- **Selección de los eigen-vectores más relevantes.** Se ordenan los *eigen*-valores de mayor a menor y elegir los d *eigen*-vectores que se correspondan con los *eigen*-vectores d más grandes (donde d es el número de dimensiones del nuevo subespacio de características). Se recomienda que si la suma de la varianza de los primeros tres d *eigen*-vectores es menor a 70%, se omita continuar con su uso para fines de representación visual o clasificación.
- **Construir la matriz de proyección W .** Tras seleccionar los d *eigen*-vectores, se construye una matriz de proyección W . Esta matriz contiene la información necesaria para proyectar los datos originales a un nuevo subespacio de características de dimensión d .
- **Transformar las características originales.** Como último paso, se transforma la base de datos original X , mediante la matriz de proyección W . Esta transformación proyecta los datos originales hacia el nuevo subespacio de características de dimensión d . El nuevo espacio de dimensión representa un porcentaje de los datos originales igual a la suma de las varianzas de las dimensiones d seleccionadas.

La reducción de dimensionalidad mediante PCA elimina la información de los atributos menos importantes, dejando solo los componentes de datos con la mayor varianza, es decir, los datos resultantes conservan la máxima varianza de los datos. Por esta razón, aunque PCA se utiliza para reducir la dimensionalidad de los datos, también puede ser útil como una herramienta de visualización, filtrar ruido y extraer características.

1.3.5. Métodos para determinar el número de clusters

1.3.5.1. Método del codo

El método del codo consiste en ejecutar el algoritmo de k -medias para una instancia dada en un rango de valores definidos de k , por ejemplo, desde $k = 1$ hasta n , y para cada resultado obtenido de k , se calcula la suma de errores al cuadrado (SSE, por siglas en inglés). La SSE representa la suma de las diferencias al cuadrado entre cada objeto y el promedio de esta distancia de su grupo. Es posible usar este parámetro como una medida de la variación intrínseca del grupo. De darse la situación de que todos los atributos de las observaciones dentro de un conjunto sean idénticos, entonces la SSE sería exactamente igual a 0. La SSE se calcula mediante la siguiente ecuación:

$$SSE = \sum_{i=1}^n (x_i - \bar{x})^2$$

Finalmente, los resultados obtenidos se representan gráficamente; si el gráfico de líneas se asemeja a un brazo, entonces el "codo" en el brazo es el valor de k que resulta ser el mejor.

1.3.5.2. Resiliencia generalizada

Dada una gráfica completa con ponderaciones en las aristas, el objetivo es minimizar la suma de las distancias entre vértices con colores iguales; es decir, la suma de las penalizaciones denominada dureza.

Sea $G = (V, E)$ un grafo completo no dirigido; es decir, que los vértices $V = \{1, 2, \dots, n\}$ en G y cada una de sus aristas posibles (i, j) , donde:

$$|V| = n; |E| = n(n-1)/2.$$

Se establece la penalización p_{ij} , tal que:

$$p_{ij} \geq 0, \quad \forall (i, j) \in E.$$

La función de coloración sobre los vértices en $G = (V, E)$ es definida como:

$$C^k : \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, k\},$$

El objetivo del problema de coloración de gráficas suaves generalizado es hallar la coloración que minimice la dureza sujeta a:

$$H(C^k) = \sum_{(i,j) \in E, c^k(i) = c^k(j)}^n p_{ij}.$$

donde k es el total de colores $1 \leq k \leq n$ que identifica a $C(i)$ como el color sobre el vértice i . Para una coloración C^k en un grafo, la función generalizada de dureza de la coloración C^k está determinada por una medida de agrupación de qué tan cercanos se encuentran los datos, esta función se llamará dureza generalizada $H(C)$. Dicha dureza tiene varias definiciones; por ejemplo, cuando se define con base en centroides, $H(C)$ es la distancia entre los datos y el centroide más cercano. Cuando la dureza se define con base en clustering, se tiene la definición *clásica* de coloración de gráficas suaves en donde se minimiza la suma de los subgrafos, donde cada subgrafo representa un color. Si se define por árboles en gráficas $H(C)$ se define a partir del árbol de expansión mínima que, al eliminar las aristas más grandes, genera k árboles mínimos. Si se define como una función de densidad, $H(C)$ se define a partir del diámetro de la vecindad definida y el número de elementos que incluye.

Se define a la *solidez* de la coloración como la dureza de un grafo dividida entre m , que representa el número promedio de aristas que contribuyen a la dureza. Dicha cantidad m de aristas depende si son centroides, subgrafos, árboles, etc. Matemáticamente se expresa como:

$$S(C_{op}^K) = \frac{H(C_{op}^K)}{m}.$$

La *resiliencia* de la coloración C^k está definida como el porcentaje en el cual decrece la solidez de una coloración con $k - 1$ colores con respecto a la efectuada con k de ellos y se expresa como:

$$R(C_{op}^K) = \frac{S(C_{op}^{K-1}) - S(C_{op}^K)}{S(C_{op}^K)}.$$

Si se encuentra que, al agregar un color nuevo a la coloración del grafo, existe un incremento significativo en su resiliencia, eso se interpreta como que se ha encontrado un número adecuado para clasificar.

Finalmente, al generarse las resiliencias para todas las coloraciones consecutivas que van desde 1 hasta n , y si interpretamos que cada color es un grupo para clasificar a los vértices (objetos), los valores más altos de n indicarán el número de clases que son más adecuadas para clasificar el conjunto. Con base en esta propiedad, es posible determinar un número adecuado de subconjuntos, resolviendo así uno de los grandes problemas del reconocimiento de patrones no supervisado.

2. IMPLEMENTACIÓN DE LOS ALGORITMOS

2.1. DESCRIPCIÓN DE LAS HERRAMIENTAS UTILIZADAS

Las implementaciones de los algoritmos fueron realizadas con el lenguaje de programación Python. Python es un lenguaje bajo licencia de código abierto y multiplataforma que soporta múltiples paradigmas de programación. Entre las características que se tomaron en cuenta al usarlo para desarrollar el modelo propuesto se encuentran:

- *Interpretado*. Al ser un lenguaje interpretado, permite hacer pruebas y desarrollos más rápidos.
- *Legibilidad*. La sintaxis que usa Python favorece a que se tenga una mejor legibilidad y comprensión del código.
- *Librerías*. Posee una amplia gama de librerías para computación científica que facilitan tareas como el manejo y representación de datos.

2.2. PREPROCESAMIENTO DE LAS BASES DE DATOS

2.2.1. Transformación de variables categóricas

En su mayoría, las bases de datos contienen al menos una variable categórica, es decir, con valores representados usando cualidades, descripciones o categorías, ejemplos de ello son el país de origen, estado civil, sexo, entre otras. Sin embargo, estas variables no pueden ser interpretadas por los algoritmos, siendo necesario su transformación a valores numéricos. La transformación de variables categóricas puede realizarse mediante diferentes técnicas, según el tipo y el número de categorías como lo son:

- Codificación *one-hot*: Crea variables binarias para cada categoría, indicando su presencia (1) o ausencia (0). Por ejemplo, si un paciente se encuentra sano (1) o enfermo (0).
- Codificación ordinal: Asigna valores teniendo en cuenta el orden jerárquico. Por ejemplo, Excelente, Bueno, Regular y Malo se transforma en una variable ordinal con valores 4, 3, 2 y 1, respectivamente.
- Codificación de enteros: Asigna valores numéricos a cada categoría, sin tener en cuenta el orden jerárquico. Por ejemplo, una variable categórica con valores Smartphone, Tablet y Computadora se transforma en una variable numérica con valores 1, 2 y 3, respectivamente.
- Combinación de método: Como su nombre lo indica es combinar dos o más técnicas para la transformación de las variables categóricas. Esta aproximación fue la implementada para el desarrollo del modelo.

2.2.2. Limpieza de las bases de datos

Este paso consiste en identificar y corregir datos incompletos o faltantes. Las acciones que se pueden tomar son las siguientes:

- Sustituir el dato faltante por medidas de tendencia central, comúnmente por la media de los datos.
- Eliminar el registro. Consiste en eliminar la fila completa de datos en caso de tener uno o más datos ausentes. Esta fue la técnica usada para el desarrollo de este modelo.

2.2.3. Normalización de las bases de datos

La normalización de la base de datos a utilizar es un requisito, debido a que algunos datos presentan escalas distintas. Este procedimiento implica obtener los valores máximos y mínimos de cada una de las columnas, que fungirán como límites, y reemplazar los demás, asegurando que todos los valores del atributo oscilen entre 1 y 0 mediante la fórmula:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Con el propósito de prevenir la división entre cero, en caso de que los valores mínimos y máximos coincidan, se asignará un valor igual a 1 para cada uno de los valores de una misma columna.

2.3. EVALUACIÓN DEL MODELO

2.3.1. Coeficiente de Correlación de Pearson

Es una medida que determina qué tan relacionadas linealmente están dos variables. Este coeficiente, representado por la letra r , calculado por medio de la siguiente fórmula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Donde:

- n es el tamaño de la muestra.

- x_i, y_i son puntos individuales muestrales
- \bar{x} establece la media muestral

El valor obtenido puede variar entre -1 y +1, y estos pueden interpretarse como se muestra en la Figura 1.

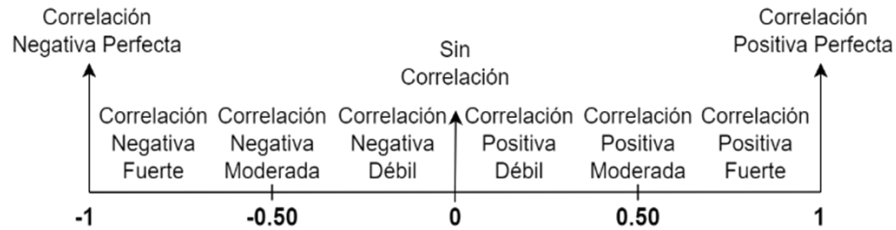


Figura 1. Interpretación del coeficiente de correlación de Pearson

Como es posible notar, a medida que se aleja de 0 la correlación entre dos variables es más fuerte, entre más negativo sea, significa que mientras una variable aumenta, la otra disminuye. Por otro lado, mientras sea más positivo, significa que cuando una variable aumenta, la otra también lo hará.

2.3.2. Coeficiente de silueta

El coeficiente de silueta o puntaje de silueta es una medida de validez de clúster para evaluar el rendimiento de la agrupación. Para calcular el puntaje de silueta para cada observación/punto de datos, se deben encontrar las siguientes distancias para cada observación perteneciente a todos los clusters:

1. Distancia promedio entre la observación y todos los demás puntos de datos en el mismo cluster. Esta distancia también puede llamarse distancia *intracluster* promedio. Esta distancia promedio se denota por $a(i)$.
2. Distancia promedio entre la observación y todos los demás puntos de datos del cluster más cercano siguiente. Esta distancia también puede llamarse distancia media al cluster más cercano. La distancia promedio se denota por $b(i)$.
3. El puntaje de silueta, $s(i)$, para cada muestra se calcula utilizando la siguiente fórmula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i) - b(i)\}}$$

Los valores del coeficiente de silueta van desde -1 hasta 1. Los coeficientes de silueta cercanos a +1 indican que los clusters están bien separados entre sí y claramente distinguibles. Un valor de 0 indica que los clusters son indiferentes, o podemos decir que la distancia entre los clusters no es significativa; los valores negativos indican que la configuración de agrupación puede tener muchos o pocos clusters. Dado que los coeficientes de silueta se utilizan para estudiar la distancia de separación entre los clústeres resultantes, también es posible utilizarlos para seleccionar el número de clusters en técnicas de agrupación.

2.3.3. Matriz de confusión

Para evaluar un modelo, se utiliza con frecuencia la exactitud, definida como la proporción de predicciones correctas realizadas por el modelo y las predicciones generales. Dado por la fórmula:

$$Exactitud = \frac{Total\ Predicciones\ Correctas}{Total\ Predicciones} * 100$$

Aunque esta fórmula es práctica, no proporciona toda la información importante, como el total de predicciones correctas e incorrectas realizadas por el modelo.

La matriz de confusión se conoció por primera vez bajo el término Tabla de contingencia, fue inventada en 1904 por Karl Pearson. Sin embargo, la matriz de confusión a término se hizo popular en la comunidad de aprendizaje automático gracias a Kohavi y Provost en 1998.

La matriz de confusión generada a partir de una instancia con n clases, es una matriz cuadrada $n \times n$ donde las filas se nombran tomando las clases y columnas reales, utilizando las clases proporcionadas por el modelo. De esta manera, se puede identificar claramente cuando el modelo clasifica una clase correcta o incorrectamente. Por lo tanto, la matriz de confusión nos permite evaluar el rendimiento del modelo con respecto a una instancia. Por ejemplo, para una instancia con dos clases, la matriz se vería como la que se muestra en la Tabla 3:

Tabla 3. Matriz de confusión para 2 clases

	Negativo (Modelo)	Positive (Modelo)
Negativo (Real)	Verdadero Negativo (TN)	Falso Positivo (FP)
Positivo (Real)	Falso Negativo (FN)	Verdadero Positivo (TP)

Así:

$$Exactitud = \frac{TN + TP}{TN + FP + FN + TP}$$

y

$$Error\ Tipo\ I = \frac{FP}{TN + FP}$$

$$Error\ Tipo\ II = \frac{FN}{FN + TP}$$

Donde:

- TN = Verdadero Negativo
- TP = Verdadero Positivo
- FN = Falso Negativo
- FP = Falso Positivo

Cada campo en la matriz proporciona un tipo diferente de éxito o error; sin embargo, es importante aclarar que, dependiendo del problema en cuestión que se está resolviendo, los errores pueden tener diferentes significados. Por ejemplo, si deseas utilizar un modelo para diagnosticar enfermedades, existen dos posibles errores que se pueden cometer: diagnosticar a una persona sana como enferma (Error de Tipo I) y diagnosticar a alguien enfermo como sano (Error de Tipo II). Aunque ambos siguen siendo errores, las consecuencias de cometer un Error de Tipo I implicarán más diagnósticos que eventualmente descartarán una posible enfermedad o condición, mientras que las consecuencias de cometer un Error de Tipo II pueden eludir la detección temprana de una enfermedad, lo que puede provocar la progresión de la enfermedad e incluso causar la muerte. En un clasificador, perseguir errores de un tipo implica aumentar los errores del otro tipo, y el tipo de error que se debe reducir dependerá del problema que se esté tratando de resolver.

2.3.4. Validación cruzada

La validación cruzada es un método estadístico para la evaluación y comparación de algoritmos de aprendizaje automático. Este método se basa en la división de los datos disponibles en dos conjuntos distintos: uno destinado al entrenamiento del modelo y otro para fines de validación. El enfoque clásico de esta técnica consiste en cruzar los conjuntos de entrenamiento y validación en rondas sucesivas, garantizando que cada punto de datos tenga la oportunidad de ser validado.

Ejecutar un algoritmo y evaluar su rendimiento estadístico en los mismos datos produce un resultado excesivamente optimista. La validación cruzada se planteó para solucionar este problema, partiendo de la idea de que probar la salida del algoritmo en nuevos datos proporcionaría una buena estimación de su rendimiento [36]

Una única división de datos produce una estimación de validación del riesgo, y promediar varias divisiones produce una estimación de validación cruzada. Diferentes estrategias de división conducen a diferentes riesgos en las estimaciones de validación cruzada [37].

La validación cruzada se destaca por su amplia aplicabilidad en el ámbito del aprendizaje automático, gracias a la naturaleza universal de las heurísticas de división de datos que emplea. Su principal fortaleza reside en la suposición mínima que realiza sobre los datos: únicamente asume una distribución idéntica entre los mismos y la independencia entre las muestras de entrenamiento y validación, permitiendo obtener una estimación precisa del rendimiento del modelo en un conjunto de datos independiente, evitando así el sesgo que podría surgir de la evaluación directa con los datos utilizados para su entrenamiento.

Es por lo anterior que la validación cruzada se puede aplicar a casi cualquier algoritmo en casi cualquier marco de trabajo, en una amplia gama de tareas que incluye, no sin limitarse a ellas, estimación de densidad, regresión y clasificación.

2.3.4.1. Validación cruzada k -fold

La validación cruzada k -fold, se caracteriza por dividir los datos en k segmentos de tamaño similar (o aproximadamente igual). Seguidamente, se llevan a cabo k iteraciones, donde en cada una de ellas se reserva un pliegue diferente de los datos para la validación. Antes de dividirlos en los k pliegues, y para optimizar la representatividad de cada pliegue, se recomienda estratificar los datos. La estratificación consiste en reorganizar los datos de manera que cada pliegue sea una buena representación del conjunto completo.

En el caso de un problema de clasificación binaria donde ambas clases representan el 50 por ciento de los datos, la estratificación adecuada garantizaría que cada pliegue contenga aproximadamente la mitad de las instancias de cada clase. De esta manera, se evita el sesgo que podría surgir si un pliegue contuviera una proporción significativamente mayor de una clase que la otra.

2.3.4.2. Validación cruzada repetida de k -fold

Para obtener una estimación o comparación confiable del rendimiento, siempre se prefieren un gran número de estimaciones. En la validación cruzada de k -fold, solo se obtienen k estimaciones. Un método comúnmente utilizado para aumentar el número de estimaciones es ejecutar la validación cruzada de k -fold varias veces. Los datos se vuelven a mezclar y estratificar antes de cada ronda.

2.3.4.3. Selección del número k de pliegues

En el trabajo [38] se comparan diferentes enfoques para estimar la exactitud: validación cruzada (incluyendo la validación cruzada regular, la validación cruzada de dejar uno fuera y la validación cruzada estratificada) y bootstrap (muestra con reemplazo), recomendando así la validación cruzada estratificada de 10 pliegues (10-fold cross-validation) como el mejor método de selección de modelos, ya que tiende a proporcionar una estimación menos sesgada de la exactitud.

2.3.5. Modelo matricial propuesto para 60 Universidades

Este modelo matricial se propone para representar de una forma más entendible los hallazgos derivados de la implementación del modelo propuesto en la base de datos de 60 Universidades.

Entre las muchas actividades que se desarrollan dentro de una universidad (gestión, difusión de la cultura, actividades deportivas, entre otras), se consideraron que los rubros más importantes son la formación de alumnos a nivel licenciatura y posgrado, así como la investigación. Estos dos rubros son los que se tomaron en cuenta para el modelo propuesto. Se usaron los datos disponibles referentes a docencia como: número de profesores a tiempo completo o parcial, grado máximo de estudios, número de alumnos matriculados, número de alumnos egresados, etcétera. Mientras que en la parte de investigación se pueden tomar en cuenta el número de artículos de investigación que se tienen en diferentes índices internacionales (JCR, ISI, Scopus, Latindex, Zentralblat Math, entre otros), así como el número de patentes generadas o citas en revistas internacionales.

Tomando en cuenta este criterio y con los datos históricos se clasifican las universidades; esto es debido a que a partir de los datos es posible asignar un orden de mayor a menor; por ejemplo, considerando la distancia de los centroides con respecto al origen, los centroides más cercanos al origen implican un desempeño menor (menos estudiantes graduados o menos artículos de investigación generados).

Considerando las dimensiones ya descritas se puede estructurar una matriz donde en uno de los ejes se muestra la clasificación de acuerdo con docencia y en la otra, con investigación.

En la matriz de resultados se ha decidido que el eje horizontal sea la investigación, mientras que el vertical sea de docencia, esto se ejemplifica en la Figura 2.



Figura 2. Ejemplo del modelo matricial propuesto

Este modelo se divide en cuatro cuadrantes de clasificación: el primer cuadrante contendrá a las instituciones estáticas, es decir, con poca docencia y poca investigación. El segundo cuadrante tendrá a las instituciones consolidadas en docencia; esto es, aquellas instituciones con poca investigación y mucha docencia. El tercer cuadrante albergará a las instituciones consolidadas en investigación, es decir, con mucha investigación y poca docencia. Finalmente, en el cuarto cuadrante estarán las instituciones de excelencia, esto significa que aquellas universidades en este sitio poseen los mejores resultados tanto en docencia como en investigación.

Usando la matriz anterior, se usarán flechas para mostrar transiciones existentes en las instituciones, indicando en la punta de cada una el año en el que ocurrió; por otro lado, las instituciones resaltadas serán aquellas que durante todo el estudio se mantuvieron en un mismo grupo.

3. MODELO PROPUESTO DE CLUSTERING AVANZADO

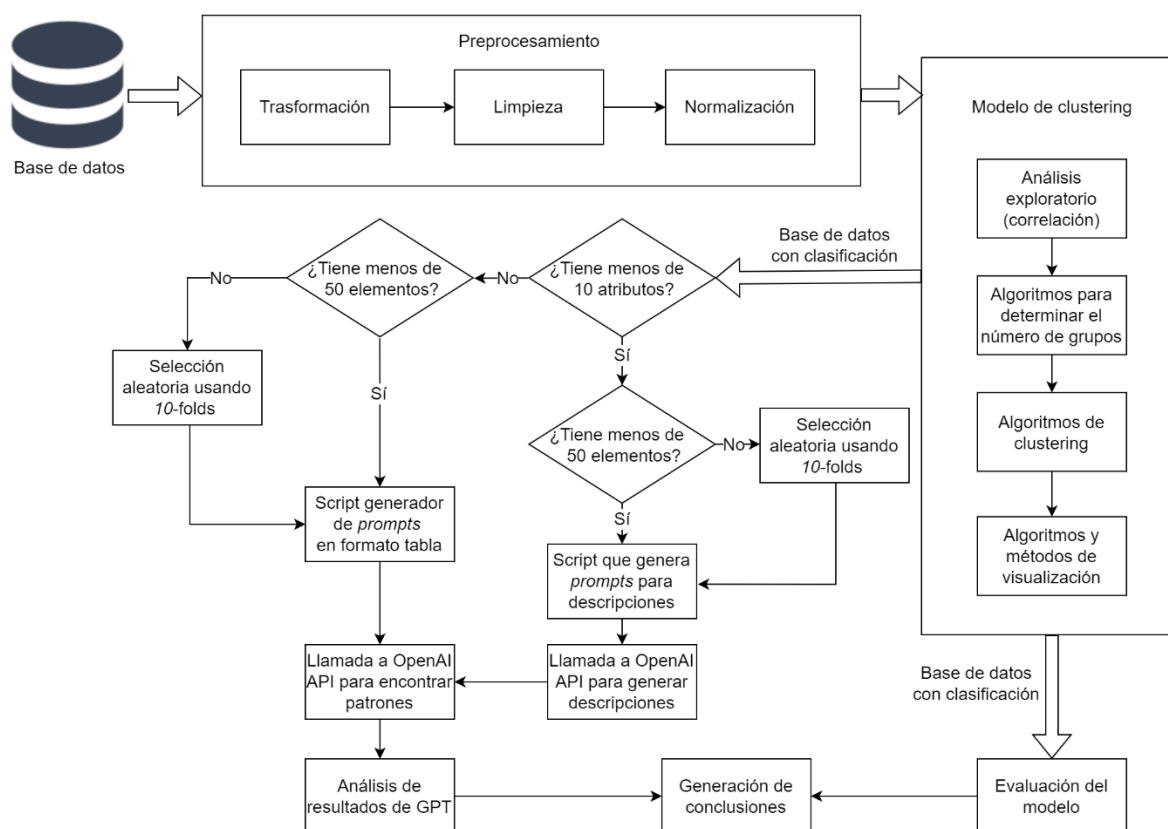


Figura 3. Diagrama de flujo del modelo avanzado de clustering

El proceso que se sigue para formular el modelo avanzado de clustering se describe en la Figura 3.

En esta se observa cómo cualquier base de datos debe de pasar por un preprocesamiento de transformación, limpieza y normalización, tras completar estas tareas es posible aplicar un modelo de clustering que consiste en un análisis exploratorio usando una matriz de

correlación, seguido de determinar el número de grupos en los que se clasificará la base de datos, estos métodos han sido anteriormente descritos en este trabajo; posteriormente la base de datos se clasifica usando tres algoritmos de clustering y finalmente se aplican algoritmos y métodos de visualización, este último paso se realiza solo con fines de una mejor representación y explicación a un público general.

Cuando se tiene la clasificación determinada por cada uno de los modelos, es necesario procesar los resultados usando dos caminos distintos.

El primero consistirá en la evaluación del modelo con los métodos descritos en secciones pasadas para determinar su validez y extraer conclusiones.

El segundo consistirá en procesar la base de datos en función de sus características, teniendo diferentes opciones dependiendo de la naturaleza de la base de datos:

1. Si esta tiene menos de 10 atributos y menos de 50 elementos, entonces un script generará un *prompt* para generar descripciones.
2. Si tiene menos de 10 atributos, pero más de 50 elementos, entonces se hará una selección aleatoria de cada elemento de una clase usando 10-folds y posteriormente pasará al script de generación de *prompts* para descripciones.

En ambos casos descritos anteriormente en los puntos 1 y 2, al tener el *prompt* se hará una consulta a la API de OpenAI para intentar extraer más información que con la que se cuenta en ese momento.

3. Si la base de datos tiene más de 10 atributos, pero menos de 50 elementos, se llamará a un script que genera *prompts* en formato de tabla
4. Si tiene más de 10 atributos y más de 10 elementos, se hará una selección aleatoria usando 10-folds con una posterior llamada al script generador de *prompts* en formato tabla.

En cualquiera de los casos descritos en desde el punto 1 al 4, se hará una llamada a la API de OpenAI ya sea tomando en cuenta las descripciones o las tablas, esto con la finalidad de detectar posibles patrones ocultos.

Una vez que se tengan los resultados de GPT, se analizarán, compararán y retroalimentarán con aquellos derivados de la evaluación del modelo, generando así discusiones y conclusiones sobre ellos.

4. INSTANCIAS

Una instancia es una colección de datos de entrada con la información suficiente para obtener una solución; mientras que un problema (base de datos) es una colección de instancias del mismo tipo. Se dice que un algoritmo resuelve un problema, si puede aplicarse a cualquier instancia y siempre garantiza una solución factible. En el presente trabajo se proponen, no sin limitarse a ellas, las siguientes instancias para evaluar y validar el modelo propuesto, separándolas en 5 categorías: salud, nutrición, económicas, investigación, clasificación humana previa y clasificación de resultados de cuestionarios.

4.1. SALUD

4.1.1. Hepatitis dataset

Esta base de datos recopila las mediciones de síntomas y tratamientos de 155 pacientes. Para esta base de datos es necesario considerar y reducir el error de tipo I, donde un error en el diagnóstico puede dejar a un paciente enfermo sin tratamiento. Esta instancia considera 19 características como atributos que son:

- 1) Edad: en años entre 7 y 78 años.
- 2) Sexo: masculino (1), femenino (2).
- 3) Se ha proporcionado un tratamiento con esteroides: sí (2), no (1).
- 4) Se ha proporcionado un tratamiento antiviral: sí (2), no (1).
- 5) El paciente presenta síntomas de fatiga: sí (2), no (1).
- 6) El paciente presenta malestares: sí (2), no (1).
- 7) El paciente presenta anorexia: sí (2), no (1).
- 8) El paciente tiene hígado grande: sí (2), no (1).
- 9) El paciente tiene hígado duro: sí (2), no (1).
- 10) Bazo palpable: sí (2), no (1).
- 11) Arañas vasculares o visualización de manchas rojas sobre la piel: sí (2), no (1).
- 12) Ascitis (almacenamiento de líquido seroso en las inmediaciones del hígado): sí (2), no (1).
- 13) Várices (venas o arterias anormalmente dilatadas): sí (2), no (1).
- 14) Bilirrubina: un líquido producido por el hígado. El nivel normal está entre 0,3 y 1,9 mg / dL. Los datos van desde 0.3 a 8 mg / dL.
- 15) Fosfatasa alcalina (ALP): una enzima relacionada con enfermedades del hígado y los huesos. Los niveles normales están entre 20 y 140UI/L. Los datos están entre 26 y 295UI/L.
- 16) Transaminas oxaloacéticas glutámicas en suero (SGOT): SGOT se libera en la sangre cuando se daña el hígado o el corazón. Rango normal para adultos: 0-42 UI / L. Los datos están entre 14 y 648 UI / L.

- 17) Albúmina. Los niveles bajos de esta proteína presente en la sangre indican que las células hepáticas se encuentran dañadas. Rango normal para adultos: 2.2- 4.2 g / dL. Los datos están entre 2.1 y 6.4 g / dL.
- 18) Tiempo de protrombina: consiste en un monitoreo estandarizado de la coagulación de la sangre. Los datos se distribuyen entre 0 y 100. Cuanto más grande, mejor.
- 19) El paciente se ha realizado histología del hígado. sí (2), no (1).

Existen datos faltantes. De 155 pacientes, solo 43 tienen todos los registros y estos fueron los pacientes utilizados en nuestro análisis, 30 vivos y 13 muertos. Las columnas alfanuméricas se intercambiaron con datos numéricos. Los datos fueron clasificados por los médicos en dos grupos: muertos y vivos.

4.1.2. Mamografías de cáncer de mama

Este conjunto de datos se refiere a la predicción de la gravedad (benigna o maligna) de una lesión mamográfica mamaria de los atributos de BI-RADS y la edad del paciente. Contiene una evaluación BI-RADS, la edad del paciente y tres atributos BI-RADS junto con la verdad básica (el campo de gravedad) para 516 masas benignas y 445 malignas que se han identificado en mamografías digitales de campo completo recolectadas en el Instituto de Radiología de la Universidad Erlangen-Nuremberg entre 2003 y 2006. Cada instancia tiene una evaluación BI-RADS asociada que va de 1 (definitivamente benigna) a 5 (altamente sugestiva de malignidad) asignada en un proceso de revisión doble por parte de los médicos. Suponiendo que todos los casos con evaluaciones de BI-RADS mayores o iguales a un valor dado (que varía de 1 a 5), son malignos y los otros casos benignos, se pueden calcular las sensibilidades y las especificidades asociadas. Estos pueden ser una indicación de qué tan bien funciona un sistema CAD en comparación con los radiólogos. Esta base de datos tiene 6 atributos en total: 1 campo de etiqueta, 1 no predictivo y 4 atributos predictivos. Los atributos se describen a continuación:

- 1) Evaluación BI-RADS: 1 a 5 (ordinal, no predictivo).
- 2) Edad: edad del paciente en años (entero).
- 3) Forma: forma de la masa: redondo = 1 ovalado = 2 lobular = 3 irregular = 4 (nominal).
- 4) Margen: margen de masa: circunscrito = 1 microlobulado = 2 oculto = 3 mal definido = 4 espiculado = 5 (nominal).
- 5) Densidad: densidad de masa alta = 1, iso = 2, bajo = 3, que contienen grasa = 4 (ordinal).
- 6) Gravedad: benigno = 0 o maligno = 1 (binominal, campo de etiqueta).

4.1.3. Biopsias de cáncer de mama

Base de datos de biopsias de 699 casos clínicos reportados por los hospitales de la Universidad de Wisconsin. Cada caso tiene una de las 2 posibles clases: benigno (0) o maligno (1). Hay datos faltantes y contiene un total de 9 atributos, cuya descripción es la siguiente:

- 1) Número de código de muestra: número de identificación
- 2) Grosor del racimo: 1 - 10
- 3) Uniformidad del tamaño de células: 1 - 10
- 4) Uniformidad de la forma de células: 1 - 10
- 5) Adhesión marginal: 1 - 10
- 6) Tamaño de célula epitelial única: 1 - 10
- 7) Núcleo desnudo: 1 - 10
- 8) Cromatina suave: 1 - 10
- 9) Nucleoli normales: 1 - 10
- 10) Mitosis: 1 - 10

La distribución de las clases es de 458 muestras benignas y 241 muestras malignas.

4.2. NUTRICIÓN

4.2.1. Cereales

Esta base de datos fue desarrollada tomando en cuenta la información publicada por el Laboratorio Nacional de Protección al Consumidor en 2011 [39] y en 2019 [40] en la cual se compararon los valores nutrimentales reportados en porciones de 30 g de las marcas de cereales de caja que más se consumen en México, los cuales son:

- Chachitos
- Kellogg's Corn Pops
- Golden Hills Azucaradas
- Kellogg's Zucaritas Nutri Crisp
- Kellogg's Froot Loops
- Maizoro Azucaradas
- Kellogg's Choco Krispis
- Ke Precio Corn Flakes
- Great value Fruti aros
- Nestlé Cookie Crisp
- Nestlé Cheerios
- Nestlé Cini Minis

- Nestlé Lucky Charms
- Nestlé Nesquik
- Nestlé Trix
- Post Honey Bunches
- Quaker Oatmeal Squares

En cuanto a los valores nutrimentales analizados y su unidad de medida que figuran en esta base de datos son:

- Proteína g
- Grasa g
- Azúcares g
- Fibra g
- Carbohidratos complejos g
- Aporte energético en kilocalorías (kcal)

En el trabajo desarrollado por [41] reportan por medio de una tabla de comparación a los cereales que son considerados como más aptos para el consumo y cuáles se recomiendan menos.

4.2.2. Cremas comestibles

Esta base de datos fue desarrollada tomando en cuenta la información publicada por el Laboratorio Nacional de Protección al Consumidor en 2014 [42] y en 2019 [43], donde reportan los estudios realizados en muestras de 100 gramos.

Para el 2014 se analizaron las siguientes marcas:

- Lyncott
- Lala Light
- Alpura Reducida En Grasa
- Cuadritos
- Rancherita
- La Abuelita
- Gina
- La Campesina Suiza
- Flor De Alfalfa
- Santa Clara
- Aguascalientes
- Alpura
- San Marcos

- Selecto Brand
- Nestlé
- Great Value
- Alpura Entera
- Lyncott Entera
- Yoplait
- Lala
- Chilchota
- Lala Media Crema
- Alpura Media Crema

Mientras que del año 2019 se analizaron las siguientes marcas:

- Aguascalientes
- Alpura Entera
- Alpura Ligera
- Chilchota
- Rancherita
- Flor De Alfalfa
- La Abuelita
- Lala
- Lala Light
- Lyncott
- Santa Clara
- Yoplait Crema

Los valores nutrimentales considerados en ambos estudios son:

- Grasa, porcentaje por cada 100 g
- Proteína, porcentaje por cada 100 g
- Sodio mg
- Costo, por cada 100 g

4.3. ECONÓMICAS

4.3.1. Wine dataset

El conjunto de datos que comprenden esta instancia corresponde a los resultados de un análisis químico para determinar la concentración de 13 componentes presentes en 178 vinos de tres variedades distintas cultivados en una misma región italiana [44].

Esta instancia modela un caso en el que vender un producto de buena calidad como una calidad inferior representa una pérdida para el empresario. Por otro lado, si un producto de mala calidad se vende como si fuera mejor, significa una mala reputación para la empresa. Los atributos son:

- 1) Porcentaje alcohólico, este valor varía en un rango que va desde 11.03 a 14.83.
- 2) Cantidad de ácido málico, este es el causante del gusto ácido al fermentar las frutas, tiene un grado de acidez que oscila de 0.74 a 5.8.
- 3) Cantidad de cenizas, derivado de calcinar el vino a 500 ° C, estos valores oscilan entre 1.36 y 3.23.
- 4) La alcalinidad presente en las cenizas, se obtiene al sumar los cationes de amonio presentes dentro de los ácidos del vino, estos valores oscilan de 10.6 a 30.
- 5) Cantidad de magnesio, este valor es importante porque ayuda a precisar la naturaleza de almacenaje del vino, las concentraciones oscilan entre 70 y 162 siempre usando números sin punto decimal.
- 6) Cantidad de fenoles totales, comúnmente conocida como coloración del vino, con valores que oscilan de 0.98 a 3.88.
- 7) Cantidad de flavonoides o, también llamados pigmentos amarillos, estos valores oscilan entre 0.34 y 5.08.
- 8) Cantidad de fenoles no flavanoides, es otro compuesto que está relacionado con el color del vino, posee valores que van como mínimo desde 0.13 hasta 0.66, como máximo.
- 9) Cantidad de proantocianinas, la sustancia de uva más importante en este caso, brinda las propiedades benéficas a la salud del vino, estos valores oscilan de 0.41 a 58.
- 10) Magnitud del color, independientemente de si el vino es rojo o blanco, con valores de 1.28 a 13.
- 11) Tinte del color del vino; no se hace diferencia si el vino es rojo o blanco, su valor oscila entre 0.48 y 1.71.
- 12) OD280 / OD315, mide la densidad de estas proteínas en vinos, las medidas fluctúan entre 1.27 y 4.
- 13) Prolina, este es un aminoácido de gran importancia para el metabolismo del nitrógeno presente en las levaduras, el valor oscila entre 78 y 1680 siempre usando números sin punto decimal.

La instancia se distribuye de la siguiente manera: Clase 1 (Vinos malos): 59. Clase 2 (Vinos Regulares): 71. Clase 3 (Buenos Vinos): 48.

4.4. INVESTIGACIÓN

4.4.1. 60 universidades mexicanas

El Estudio Comparativo de Universidades Mexicanas (ECUM) [45] es un proyecto de investigación desarrollado por la Dirección General de Evaluación Institucional (DGEI) de la Universidad Nacional Autónoma de México (UNAM) que realiza la sistematización, el análisis y la difusión de conjuntos de datos estadísticos obtenidos de fuentes autorizadas y bases de datos acreditadas. Estos datos permiten evaluar el desempeño de las universidades mexicanas en sus funciones primordiales: enseñanza, investigación y promoción de la cultura.

El ECUM no se configura como una clasificación categórica (ranking) de los organismos de educación superior mexicanos, sino que se propone como una opción alternativa a los rankings existentes. Esto se debe a que su propósito no radica en evaluar a las instituciones o establecer jerarquías basadas en ciertos supuestos preestablecidos sobre la calidad o el prestigio de las instituciones y sus programas. En cambio, su propósito fundamental es proporcionar datos objetivos generados de fuentes de acceso público.

Derivado de lo anterior, el ECUM favorece la exposición de datos en su estado original, evitando el uso de agrupaciones o ponderaciones, ya que este tipo de praxis generalmente derivan en resultados que son objetos de debate. Las características distintivas del ECUM posibilitan que sean los propios usuarios quienes establezcan las comparaciones y relaciones que consideren pertinentes entre los diversos rubros de información disponibles, o construyan indicadores personalizados en función de sus necesidades y enfoques de análisis. De igual forma, recae sobre los usuarios la responsabilidad de adaptar sus interpretaciones a las distintas características que presentan las universidades mexicanas entre sí.

Actualmente, el ECUM en su versión 3, incorpora datos desde el año 2009 hasta el 2022 de 60 instituciones de educación superior (45 de carácter público y 15 privadas), las cuales representan más del 50% de la matrícula de educación superior en México. La información recopilada se organiza en las siguientes categorías:

- Profesorado, matrícula estudiantil y programas de estudio ofertados.
- Información sobre producción de patentes requeridas y concedidas, de acuerdo con los registros del Instituto Mexicano de la Protección Industrial (IMPI).
- Participación institucional en artículos, documentos y citas indexadas en bases de datos bibliográficas internacionales de prestigio como ISI, Web of Knowledge, Scopus, SciVerse, entre otras.

- Participación institucional en artículos, documentos y citas indexadas en bases de datos regionales como Periódica (índice de revistas latinoamericanas en ciencias) y Clase (Citas latinoamericanas en ciencias sociales y humanidades).
- Académicos adscritos a las instituciones y al Sistema Nacional de Investigadores (SNI) del Consejo Nacional de Humanidades, Ciencia y Tecnología (Conahcyt), anteriormente conocido como Conacyt.
- Cantidad de revistas de investigación de cada institución indexadas por Latindex (Índice latinoamericano de publicaciones científicas seriadas) y el Índice Conahcyt.
- Cantidad de cuerpos académicos reconocidos en el Programa para el Desarrollo Profesional Docente (PRODEP), anteriormente conocido como Programa Nacional de Mejoramiento de Profesorado (PROMEP).
- Cantidad de programas de posgrado reconocidos en el Sistema Nacional de Posgrados (SNP) del Conahcyt, anteriormente conocido como Padrón Nacional de Posgrados de Calidad (PNPC).
- Cantidad de programas de educación superior de cada institución avalados por agencias reconocidas por el Consejo para la Acreditación de la Educación Superior (COPAES) y los Comités Interinstitucionales para la Evaluación de la Educación Superior (CIEES).

Los datos obtenidos de estos estudios para los indicadores anteriormente descritos, se publican en una plataforma que brinda un acceso fácil y navegación intuitiva desde el Explorador de Datos del Estudio Comparativo de Universidades Mexicanas, que actualmente se encuentra en su tercera versión (ExECUM3).

Tomando en cuenta la base de datos ExECUM, para la primera dimensión, la cual corresponde a los datos de docencia, se usó la siguiente información:

Profesores que imparten cursos

- Contrato: Tiempo completo, 3/4 de tiempo, 1/2 tiempo, por horas.
- Escolaridad: Técnico superior universitario, licenciatura, especialidad, maestría, doctorado.

Número de alumnos graduados

- Nivel: Licenciatura, especialidad, maestría, doctorado.

Programas académicos ofertados

- Nivel: Licenciatura, especialidad, maestría, doctorado.

Por otro lado, para la segunda dimensión, la cual corresponde a investigación, se usó la información descrita a continuación:

Investigadores SNI

- Investigadores: Candidato, nivel I, nivel II, nivel III

Cuerpos académicos PROMEP

- Consolidado, en consolidación, en formación.

ISI

- Artículos: Análisis por autor, producción institucional, citas, colaboradores.
- Documentos: Análisis por autor, producción institucional, citas, colaboradores.

SCOPUS

- Artículos: Producción institucional, análisis por autor, colaboradores, citas
- Documentos: Producción institucional, análisis por autor, colaboradores, citas

Patentes

- Solicitadas u otorgadas

Revistas

- Latindex o índice Conahcyt

Posgrados pertenecientes al SNP

- Doctorado públicos y privados de reciente creación, en desarrollo, consolidados y competencia internacional.
- Maestría públicos y privados de reciente creación, en desarrollo, consolidados y competencia internacional.
- Especialidad públicos y privados de reciente creación, en desarrollo, consolidados y competencia internacional.

4.5. CLASIFICACIÓN HUMANA PREVIA

4.5.1. Cuadernos

Es una base de datos generada a partir de los estudios publicados por PROFECO sobre cuadernos escolares en México [46], esta consiste en el estudio de 34 cuadernos profesionales de 105, 100 y 90 hojas; cosidos, con espiral de plástico y metálico, las características que tomaron en cuenta y se analizaron son:

- Costo
- Acabados
- Gramos por hoja indicados (g/m²)
- Gramos por hoja reales (g/m²)
- Resistencia al borrado
- Resistencia de hoja
- Resistencia de tapa

Una de las ventajas que posee esta base de datos es que cuenta con una clasificación asignada por PROFECO tras tomar en cuenta los resultados de sus análisis, estos son:

- Scribe 7510, Muy bueno
- Estrella Único, Muy bueno
- Manufacturas 8-A, Muy bueno
- Estrella/ Star Kid, Bueno
- Sazz, Suficiente
- U-Pak, Suficiente
- Norma Cosido, Excelente
- Estrella Cosido, Excelente
- Office Max Espiral, Muy bueno
- Scribe Excellence, Muy bueno
- Scribe In Black, Muy bueno
- Manufacturas 8-A, Muy bueno
- First Class 5038, Muy bueno
- First Class 5040, Muy bueno
- First Class 4517, Muy bueno
- Norma Jean Book, Muy bueno
- Scribe Clásico Cosido, Muy bueno
- Scribe Kids Cosido, Muy bueno
- First Class 4507, Bueno
- Estrella Drive Linea A, Bueno
- Scribe In Colors 1070, Bueno
- First Class Cuaderno Cosido 9408, Bueno
- First Class Profesional 5039, Bueno
- Scribe Poly Cover 2050, Bueno
- Norma, Bueno
- Estrella 0520, Bueno
- First Class Cosido Profesional 9214, Bueno
- Norma Click Urbana, Bueno
- U-Pak 1157525405, Bueno
- Norma Kiut, Bueno

- Norma Girly Notes, Suficiente
- Norma Uno, Suficiente
- Nine To Five, Suficiente

Sin embargo, no muestran mayor detalle de si existe una relación entre aquellos cuadernos con la misma calificación.

4.6. CLASIFICACIÓN DE RESULTADOS DE CUESTIONARIOS

4.6.1. Implementación de nuevos modelos educativos

Esta base de datos se generó a partir del trabajo presentado por [47], donde se propone una estrategia educativa basada en la escuadra invertida, un modelo pedagógico que combina el estudio en casa con el aprendizaje interactivo en el aula, para adaptarse al cambio de escenarios presenciales a híbridos en la carrera de Comercio Internacional y Aduanas de la UPEGro. Su estrategia fue validada por medio de cuestionarios aplicados a 48 participantes.

Estos cuestionarios consistieron en responder las siguientes preguntas, con sus posibles valores:

1. ¿Consideras que el profesor apoyó en tus dudas de clase proporcionando orientación?
 - a. Sí
 - b. No
2. ¿Consideras que la estrategia del docente fue adecuada en el desarrollo de tus actividades en la clase de Comercio Electrónico?
 - a. Sí
 - b. No
 - c. No lo sé
3. De los siguientes aparatos electrónicos, ¿Cuál utilizaste durante las clases de Comercio Electrónico?
 - a. Computadora portátil
 - b. Smartphone/Tablet
4. ¿Fueron funcionales las clases a distancia?
 - a. Sí
 - b. No
5. ¿La aplicación de Classroom para el desarrollo de la clase de Comercio electrónico fue adecuado para tu aprendizaje?
 - a. Totalmente de acuerdo
 - b. Parcialmente de acuerdo
 - c. Mayormente en desacuerdo
 - d. Totalmente en desacuerdo

6. ¿La aplicación Google Meet para el desarrollo de la clase de Comercio electrónico fue adecuado para tu aprendizaje?
 - a. Excelente
 - b. Bueno
 - c. Regular

No obstante, el análisis de los resultados se limita a concluir si el modelo fue aceptado o no.

5. RESULTADOS

Para un mejor entendimiento, las bases de datos han sido separadas como se presentaron en la sección anterior, es decir, en 5 categorías las cuales son: salud, nutrición, económicas, investigación y con clasificación humana. A su vez, los resultados de cada base de datos han sido divididos en dos secciones, en una primera sección se muestran los resultados obtenidos al implementar las técnicas de *clustering*, mientras que en la segunda sección se muestra el análisis posterior usando GPT para intentar encontrar posible información no considerada con anterioridad.

5.1. SALUD

Por lo general, en casos relacionados con la salud, es preferible asegurarse de que todos los pacientes en riesgo sean tratados, aunque esto signifique tener falsos positivos, es decir, diagnosticar a algunos pacientes sanos como enfermos. Por esta razón, los modelos que resuelven este tipo de bases de datos tratan de disminuir el Error Tipo II, donde un error en el diagnóstico puede dejar a un paciente enfermo sin tratamiento.

5.1.1. Hepatitis

Para la base de datos de Hepatitis, las tablas 4-6 muestran los resultados de las matrices de confusión que pertenecen a tres métodos de agrupamiento, y la tabla 7 muestra los resultados obtenidos para exactitud y errores de Tipo I y II.

Tabla 4. Matriz de confusión para la instancia Hepatitis solucionada con k -medias

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	16	3
Saludable (Real)	29	112

Tabla 5. Matriz de confusión para la instancia Hepatitis solucionada con GMM

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	14	5
Saludable (Real)	19	122

Tabla 6. Matriz de confusión para la instancia Hepatitis solucionada con Spectral

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	17	2
Saludable (Real)	68	73

Tabla 7. Resultados para la instancia Hepatitis solucionada con tres métodos de clustering

	Hepatitis		
	<i>k</i> -medias	Gauss	Spectral
Exactitud	80.00%	85.00%	56.25%
Error Tipo I	15.78%	26.31%	10.52%
Error Tipo II	20.56%	13.47%	48.22%
Tiempo (s)	0.3642	0.3614	0.7978

En este caso, la mayor exactitud la proporciona el modelo de mezclas Gaussianas con un 85%, y en este caso, después de analizar la matriz de confusión, coincide que es la mejor técnica, ya que considerando la minimización del Error de Tipo II, este método solo deja fuera al 13.5% de los pacientes en comparación con las otras dos técnicas, que tienen un 20.56% y 48.22% para *k*-medias y agrupamiento espectral, respectivamente.

Se presenta una comparación de los resultados obtenidos de nuestro modelo en relación con otras técnicas de clasificación [48] en la tabla 8. Nuestro mejor resultado se muestra en negrita, mientras que nuestros otros resultados se presentan en cursiva.

Tabla 8. Comparación de clasificadores para la instancia Hepatitis

Técnica	Exactitud (%)	Tipo de prueba
Nuestro (Gauss)	85.00	Subconjuntos sin entrenamiento
<i>Nuestro (k-medias)</i>	80.00	<i>Subconjuntos sin entrenamiento</i>
IBK	66.45	Subconjuntos con entrenamiento
Bagging	64.52	Subconjuntos con entrenamiento
JRip	63.23	Subconjuntos con entrenamiento
MP	62.58	Subconjuntos con entrenamiento
J48	58.06	Subconjuntos con entrenamiento
<i>Nuestro (Spectral)</i>	56.25	<i>Subconjuntos sin entrenamiento</i>

En este caso, nuestro modelo no solo obtiene mejores resultados en términos de exactitud, sino que va un paso más allá de lo propuesto en [48] y evalúa los Tipos de Errores para tomar la mejor decisión sobre qué técnica es la más adecuada para esta instancia.

5.1.1.1. Análisis GPT de resultados de Hepatitis de la mejor técnica (GMM)

En este caso particular, es necesario determinar si el paciente tiene una probabilidad de mortalidad y si este es el caso, determinar las razones, es por esto que el análisis se centra en determinar las causas de errores Tipo II.

- Determinar causas de errores Tipo II

- Consulta

Tomando en cuenta los siguientes datos, ¿podrías decir que tienen en común y si vivieron o murieron?

- Respuesta

Los datos que has proporcionado corresponden a dos pacientes con hepatitis que tienen algunas características en común. Algunas de ellas son:

- Son hombres (SEX = 1).
- Tomaron esteroides (STEROID = 1).
- Tenían fatiga y malestar (FATIGUE = 1, MALAISE = 1).
- Tenían el hígado grande y firme (LIVER BIG = 2, LIVER FIRM = 1).
- Tenían arañas vasculares, ascitis y varices (SPIDERS = 2, ASCITES = 2, VARICES = 2).
- Tenían el mismo nivel de albumina (ALBUMIN = 3.8).

No puedo decir con certeza si los pacientes vivieron o murieron, sin embargo, puedo intentar hacer una predicción basada en algunos factores que podrían influir en el resultado. Por ejemplo, de acuerdo con diversos estudios, los pacientes con hepatitis que tienen un nivel de bilirrubina mayor de 3 mg/dL, un nivel de albumina menor de 3 g/dL, o un tiempo de protrombina mayor de 16 segundos tienen un mayor riesgo de mortalidad. Además, según otros estudios, los pacientes con hepatitis que tienen histología positiva, ascitis, varices o arañas vasculares tienen un peor pronóstico que los que no las tienen. Teniendo en cuenta esto, en conjunto con su ascitis, varices y arañas vasculares, lo que también empeora su situación y podríamos suponer son susceptibles a morir.

- Hallazgo

Es importante tomar en cuenta estudios más recientes en la materia para poder determinar la importancia de ciertos factores en la mortalidad de esta enfermedad, pudiendo así generar clasificadores más confiables para este tipo de aplicación.

5.1.2. Mamografías de cáncer de mama

Para la instancia de mamografías de cáncer de mama, los resultados obtenidos de las matrices de confusión se describen en las tablas 9-11, y la exactitud y los errores de Tipo I y II proporcionados para cada método de agrupación se muestran en la tabla 12.

Tabla 9. Matriz de confusión para la instancia mamografías solucionada con *k*-medias

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	694	164
Saludable (Real)	94	828

Tabla 10. Matriz de confusión para la instancia mamografías solucionada con GMM

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	721	137
Saludable (Real)	85	837

Tabla 11. Matriz de confusión para la instancia mamografías solucionada con Spectral

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	661	197
Saludable (Real)	79	843

Tabla 12. Resultados para la instancia mamografías solucionada con tres métodos de clustering

	Mamografías de cáncer de mama		
	<i>k</i> -medias	Gauss	Spectral
Exactitud	85.50%	87.52%	84.49%
Error Tipo I	19.11%	15.96%	22.96%
Error Tipo II	10.19%	9.21%	8.56%
Tiempo (s)	1.1848	1.1459	0.9876

Analizando los resultados obtenidos de clasificar la instancia de mamografías de cáncer de mama, la técnica con el menor Error Tipo I es Gauss, pero su Error Tipo II es relativamente alto (9.21%), aunque el Error Tipo I no se aleja mucho de la anterior cifra (15.96%). La mejor opción para el cáncer de mama es Spectral, que solo diagnostica un 8.59% de los pacientes con Error Tipo II, aunque comente un 22.96% de Error Tipo I.

A continuación, se presenta una comparativa de los resultados obtenidos de nuestro modelo frente a otras técnicas de clasificación [49] resolviendo la misma instancia de mamografías de cáncer de mama en la Tabla 13. Nuestro mejor resultado está en negrita, mientras que nuestros otros resultados están en cursiva.

Tabla 13. Comparación de clasificadores para la instancia mamografías de cáncer de mama

Técnica	Exactitud (%)	Tipo de prueba
NB	89.8	Subconjuntos con entrenamiento
NBNB	89.6	Subconjuntos con entrenamiento
IBk	89.6	Subconjuntos con entrenamiento
MLPMLP	89.5	Subconjuntos con entrenamiento
(Bos)(Bag)	89	Subconjuntos con entrenamiento
<i>Nuestra (Gauss)</i>	<i>87.52</i>	<i>Subconjuntos sin entrenamiento</i>
<i>Nuestra (k-medias)</i>	<i>85.5</i>	<i>Subconjuntos sin entrenamiento</i>
Nuestra (Spectral)	84.49	Subconjuntos sin entrenamiento

Una vez más se demuestra que, dependiendo de la instancia, es necesario tener en cuenta otras medidas de rendimiento, ya que, aunque el resultado seleccionado por nuestro modelo es el peor en exactitud, el análisis realizado anteriormente (ver Tabla 12) muestra que este es el más indicado para resolver esta instancia debido a la reducción del Error de Tipo II.

5.1.2.1. Análisis GPT de resultados de Mamografías de cáncer de mama de la mejor técnica (Spectral)

Las preguntas para esta instancia se orientaron para determinar las causas que podrían conducir al modelo a diagnosticar a una persona sana como enferma (Error de Tipo I) y a una enferma como sana (Error Tipo II).

- Determinar causas de errores Tipo I

- Consulta

Se tomó una muestra del 10% de los elementos y se generó la siguiente consulta cargando los datos previamente:

¿Qué tienen en común los siguientes datos, los cuales corresponden a datos relacionados con mamografías?

- Respuesta

Los datos que mencionas corresponden a características de las imágenes de mamografía que se usan para evaluar el riesgo de cáncer de mama. Estos datos se basan en el Sistema de Base de Datos e Informes de Imágenes de la Mama (BI-RADS®), que es un sistema estandarizado para clasificar las mamografías según su apariencia y la probabilidad de cáncer. Los valores posibles son: 0 (incompleto), 1 (negativo), 2 (hallazgos benignos), 3 (probablemente benigno), 4 (sospechoso de malignidad), 5 (altamente sospechoso de malignidad) y 6 (conocido por ser maligno).

Según los datos, hay cuatro personas que tienen una evaluación BI-RADS de 5, que significa altamente sospechosa de cáncer, que también tienen una forma y un margen del tumor o la lesión que pueden indicar malignidad. Las demás personas tienen una evaluación BI-RADS de 4 o 3, que significa sospechosa o probablemente benigna de cáncer, pero con una forma y un margen del tumor o la lesión que suelen ser benignos. La densidad del tumor o la lesión varía entre 2 y 3, lo que significa que tienen una densidad media o alta, que puede dificultar la detección del cáncer.

- Hallazgo

El modelo toma en cuenta las densidades del tumor y valores de BI-RADS para determinar haberlos clasificado como enfermos a pesar de que no lo estén, lo cual es una buena decisión, ya que sirve para descartar posibles errores de medición o humanos o anomalías.

- Determinar causas de errores Tipo II

- Consulta

Se tomó una muestra del 10% de los elementos y se generó la siguiente consulta cargando los datos previamente:

¿Qué tienen en común los siguientes datos, los cuales corresponden a datos relacionados con mamografías?

- Respuesta

Estos datos muestran que todas las mamografías tienen una evaluación BI-RADS de 4 y 5, lo que significa que son altamente sospechosas de malignidad y que se necesita una biopsia urgente para confirmar el diagnóstico. La mayoría de las mamografías tienen una forma irregular, un margen especulado o mal definido y una densidad heterogéneamente densa, lo que son características típicas de las masas malignas. La edad de las personas varía desde 37 hasta 81 años, lo que indica que el cáncer de mama puede afectar a mujeres de cualquier edad, aunque es más común en las mayores de 50 años. Solo una mamografía tiene una densidad de 1, lo que significa que el tejido mamario es graso. Esto puede facilitar la detección de anomalías en la mamografía, pero no implica que el riesgo de cáncer de mama sea menor.

- Hallazgo

Gracias al uso de GPT es posible determinar que se puede mejorar el modelo no supervisado si los datos se pasan por un filtro previo de GPT o, en su defecto, se le da una mayor importancia a determinados atributos, como lo son: el resultado de la BI-RADS y su edad.

5.1.3. Biopsias de cáncer de mama

Para el caso de las biopsias de cáncer de mama, los resultados obtenidos de las matrices de confusión se describen en las tablas 14-16, y la exactitud, así como los errores de Tipo I y Tipo II para cada método de agrupamiento, se muestran en la tabla 17.

Tabla 14. Matriz de confusión para la instancia Biopsias de cáncer de mama solucionada con k -medias

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	1762	43
Saludable (Real)	68	867

Tabla 15. Matriz de confusión para la instancia Biopsias de cáncer de mama solucionada con Gauss

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	1694	118
Saludable (Real)	2	926

Tabla 16. Matriz de confusión para la instancia Biopsias de cáncer de mama solucionada con Spectral

	Enfermo (Modelo)	Saludable (Modelo)
Enfermo (Real)	1777	37
Saludable (Real)	35	891

Tabla 17. Resultados para la instancia Biopsias de cáncer de mama solucionada con tres métodos de clustering

	Biopsias de cáncer de mama		
	<i>k</i> -medias	Gauss	Spectral
Exactitud	95.94%	95.62%	97.37%
Error Tipo I	2.38%	6.51%	2.03%
Error Tipo II	7.27%	0.21%	3.77%
Tiempo (s)	1.0552	1.0703	0.9584

En relación con los resultados de las biopsias de cáncer de mama, se obtiene la mejor exactitud con la técnica Spectral de clustering (97.37%). Sin embargo, dado que en casos de salud necesitamos reducir el Error de Tipo II, el mejor método es Gauss (0.21%).

En la Tabla 18, son comparados los resultados obtenidos contra otras técnicas de clasificación [48]. Nuestro mejor resultado se encuentra en negrita, mientras que nuestros otros resultados están en cursiva.

Tabla 18. Comparación de clasificadores para la instancia Biopsias de cáncer de mama

Técnica	Exactitud (%)	Tipo de prueba
<i>Nuestra (Spectral)</i>	<i>97.37</i>	<i>Subconjuntos sin entrenamiento</i>
JRip	96.28	Subconjuntos con entrenamiento
<i>Nuestra (k-medias)</i>	<i>95.94</i>	<i>Subconjuntos sin entrenamiento</i>
Bagging	95.85	Subconjuntos con entrenamiento
MP	95.85	Subconjuntos con entrenamiento
Nuestra (Gauss)	95.62	Subconjuntos sin entrenamiento
IBK	95.14	Subconjuntos con entrenamiento
J48	94.56	Subconjuntos con entrenamiento

A pesar de que los resultados obtenidos aparentemente indican que nuestra propuesta utilizando la clasificación espectral es la mejor técnica, el análisis en la tabla 17 muestra que la mejor técnica es Gauss, no obstante de tener una menor exactitud.

5.1.3.1. Análisis GPT de resultados de Biopsias de cáncer de mama (Gauss)

- Determinar causas de errores Tipo I

- Consulta

Se tomó una muestra del 10% de los elementos y se generó la siguiente consulta cargando los datos previamente:

¿Qué tienen en común los siguientes datos, los cuales corresponden a datos relacionados con biopsias?

- Respuesta

Los casos que tienen valores más altos para los atributos como el grosor del racimo, la uniformidad del tamaño y la forma de las células, la adhesión marginal y el tamaño de célula epitelial única, son más propensos a ser malignos que benignos. Si observas los datos que me mostraste, verás que algunos tienen valores altos para algunos de estos atributos, como el núcleo desnudo, la cromatina suave o los nucleoli normales. Estos datos podrían indicar que se trata de casos malignos, pero para confirmarlo se recomienda realizar una revisión de los resultados de la biopsia por parte de un segundo patólogo experto, especialmente en los casos de mayor complejidad o incertidumbre.

- Hallazgo

La respuesta proporcionada añade un grado de incertidumbre, es decir, aunque el modelo haya predicho que son malignos, la respuesta a la consulta indica que es altamente recomendable tomar en cuenta una segunda opinión.

- Determinar causas de errores Tipo II

- Consulta

Se tomó una muestra del 10% de los elementos y se generó la siguiente consulta cargando los datos previamente:

¿Qué tienen en común los siguientes datos, los cuales corresponden a datos relacionados con biopsias?

- Respuesta

Una posible forma de determinar qué tienen en común estos registros es comparar sus valores con los promedios de cada atributo para las clases benignas y malignas. Por ejemplo, el grosor del racimo (clump thickness) tiene un promedio de 2.91 para los casos benignos y de 7.19 para los casos malignos. Si observas los registros ingresados, verás que todos tienen un valor de 5 u 8 para este atributo, lo que indica que están más cerca de la clase maligna. Lo mismo ocurre con otros atributos como la uniformidad del tamaño de células, la uniformidad de la forma de células, la adhesión marginal y el tamaño de célula epitelial única. Estos atributos tienen valores más altos para los casos malignos que para los benignos. Por lo tanto, se podría decir que estos registros tienen en común que presentan características que los hacen más propensos a ser clasificados como malignos que como benignos.

- Hallazgo

Gracias a los resultados proporcionados se puede generar un posible nuevo y mejor modelo no supervisado para esta aplicación en específico y el cual sea nuevamente comparado con estudios reportados en la literatura. Este modelo consistiría en hacer una preclasificación de los datos tomando en cuenta la característica *clump thickness* o explorar otras formas de modelado del problema en la cual se resalte esta característica; un ejemplo de ello, sería el uso de la coloración de gráficas [50].

5.2. NUTRICIÓN

5.2.1. Cereales

En el trabajo desarrollado en [51] se reportan por medio de una tabla comparativa los resultados obtenidos al resolver la instancia usando técnicas de reconocimiento de patrones, pudiendo determinar cuáles son los cereales considerados como más aptos para el consumo y cuáles se recomiendan menos, concluyendo que las marcas americanas y las de Nestlé pasaron a ser menos saludables, mientras que Maizoro y Kellogg's mejoraron su calidad con el paso del tiempo. La tabla que resume esto es la siguiente.

Tabla 19. Comparación entre los cereales de caja en 2011 y 2019

Resultado 2011	Resultado 2019	Nombre	Contenido
Saludable	Saludable	Chachitos	Se mantiene
Menos saludable	Saludable	Kellogg's Corn Pops	Mejora
Menos saludable	Saludable	Golden Hills Azucaradas	Mejora
Menos saludable	Saludable	Kellogg's Zucaritas Nutri Crisp	Mejora
Menos saludable	Saludable	Kellogg's Froot Loops	Mejora
Menos saludable	Saludable	Maizoro Azucaradas	Mejora
Menos saludable	Saludable	Kellogg's Choco Krispis	Mejora
Saludable	Saludable	Ke Precio Corn Flakes	Se mantiene
Menos saludable	Saludable	Great value Fruti aros	Mejora
Saludable	Menos saludable	Nestlé Cookie Crisp	Empeora
Saludable	Menos saludable	Nestlé Cheerios	Empeora
Saludable	Menos saludable	Nestlé Cini Minis	Empeora
Saludable	Menos saludable	Nestlé Lucky Charms	Empeora
Saludable	Menos saludable	Nestlé Nesquik	Empeora
Saludable	Menos saludable	Nestlé Trix	Empeora
Saludable	Menos saludable	Post Honey Bunches	Empeora
Saludable	Menos saludable	Quaker Oatmeal Squares	Empeora

5.2.1.1. Análisis con GPT

- Análisis de los resultados del año 2019

Tomando en cuenta los resultados que se muestran durante el año 2019, se hizo la siguiente consulta:

- Consulta

¿Qué tienen en común los siguientes cereales de caja?

Chachitos

Kellogg's Corn Pops

Golden Hills Azucaradas

Kellogg's Zucaritas Nutri Cr

Kellogg's Froot Loops

Maizoro Azucaradas

Kellogg's Choco Krispis

Ke Precio Corn Flakes

Great value Fruti aros

- Respuesta

Son cereales de desayuno: Todos los cereales mencionados están diseñados y comercializados como opciones de desayuno.

Contienen azúcar: Estos cereales son conocidos por tener un contenido significativo de azúcar añadida para mejorar su sabor y atractivo.

Son producidos por diferentes marcas: Los cereales mencionados provienen de diferentes marcas, como Kellogg's, Great Value y Maizoro.

Tienen formas y sabores distintivos: Cada cereal tiene una forma y sabor únicos. Por ejemplo, Chachitos tiene forma de ruedas, Froot Loops tiene forma de aros y colores brillantes, Choco Krispis tiene un sabor a chocolate, y así sucesivamente.

- Hallazgo

Si bien la respuesta puede resultar genérica, también brinda una característica no reportada anteriormente, la cual es que, que en general producen cereales más saludables son tres: Kellogg's, Great Value y Maizoro.

Esto puede llevar a una búsqueda más extensiva, por ejemplo, haciendo la siguiente consulta:

¿Qué tienen en común Great Value y Maizoro?

La respuesta fue:

Great Value y Maizoro comparten algunas similitudes en cuanto a su naturaleza y oferta de productos:

Marcas privadas: Tanto Great Value como Maizoro son marcas privadas, lo que significa que son propiedad de minoristas y no son marcas específicas de una sola empresa de alimentos. Great Value es una marca exclusiva de Walmart, mientras que Maizoro es una marca propia de tiendas de comestibles en México.

Es así, como se pudo determinar que tanto Kellogg's como Great Value son fabricados por una empresa argentina llamada Gollek, la cual es una empresa establecida en Latinoamérica dedicada a la producción y distribución de productos a base de arroz inflado; por lo anterior, podemos deducir que, en general, aquellas marcas producidas por empresas latinoamericanas son consideradas como una opción más saludable.

5.2.2. Cremas comestibles

En el trabajo desarrollado en [41] se aplicaron técnicas de clasificación no supervisada como lo es PCA y *K*-medias para clasificar las bases de datos [42], [43], mismo donde se muestran las siguientes gráficas:

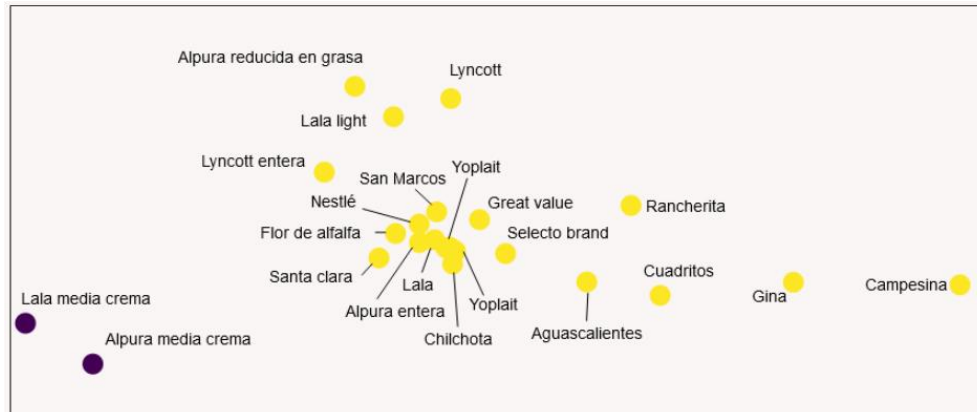


Figura 4. Clasificación de cremas 2014

La anterior muestra el resultado de aplicar PCA y técnicas de agrupación a las cremas en 2014 las cuales se dividen en medias cremas (morado) y el resto (amarillo)

Usando esta misma metodología se generó una clasificación y representación para las cremas en 2019 las cuales se dividen en cremas vegetales ligeras (amarillo) y el resto (morado)

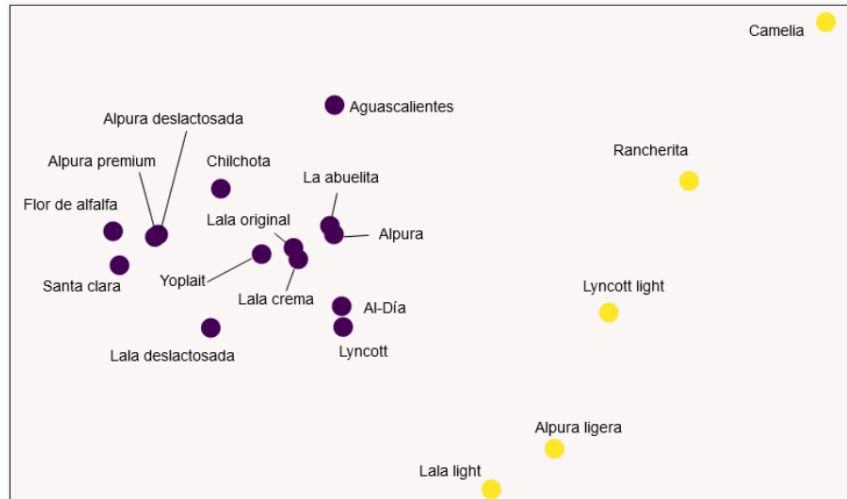


Figura 5. Clasificación de cremas 2019

5.2.2.1. Análisis de los resultados reportados con GPT

- Análisis del año 2014 con GPT

- Consulta

¿Qué tienen en común las siguientes cremas?

Lala media crema

Alpura media crema

- Respuesta

Las cremas Lala media crema y Alpura media crema comparten varias similitudes:

Tipo de crema: Ambas son variedades de crema láctea conocidas como "media crema". La media crema es una crema láctea con un contenido de grasa medio, generalmente alrededor del 15% al 20%.

Uso culinario: Tanto Lala media crema como Alpura media crema son utilizadas en la cocina para agregar suavidad, cremosidad y sabor a diferentes platillos y postres.

- Hallazgo

En este análisis en particular no se tuvo información de mayor relevancia.

- Análisis del año 2019 con GPT

- Consulta

¿Qué tienen en común las siguientes cremas?

Camelia

Rancherita

Lyncott light

Alpura ligera

Lala light

- Respuesta

Las cremas Camelia, Rancherita, Lyncott light, Alpura ligera y Lala light comparten algunas características en común:

Son productos lácteos: Todas estas cremas son derivados lácteos y se elaboran a partir de leche.

Variante "light": Todas estas marcas ofrecen una variante "light" o baja en grasas de sus respectivas cremas. Estas versiones suelen tener un contenido reducido de grasa en comparación con las cremas regulares.

- Hallazgo

En este caso se obtuvo una información importante y es que las cremas vegetales son consideradas como una variante a la "light".

5.3. ECONÓMICAS

5.3.1. Vinos

En la industria del vino, vender un producto de buena calidad como uno de calidad inferior representa una pérdida para el empresario. Por otro lado, si se vende un producto de mala calidad como uno de buena calidad, significa una mala reputación para la empresa derivando en la posible pérdida de clientes y, eventualmente, pérdidas económicas. En este caso, es preferible vender un buen vino como uno regular en lugar de vender un vino regular como uno bueno, ya que esto le da una mala reputación a la empresa. Los resultados de resolver esta instancia se muestran en las matrices de confusión siguientes:

Tabla 20. Matriz de confusión para vinos clasificada con k-medias

	<i>k</i> -means		
	Malo (Modelo)	Regular (Modelo)	Bueno (Modelo)
Malo (Real)	124	0	0
Regular (Real)	1	134	5
Bueno (Real)	0	0	96

Tabla 21. Matriz de confusión para vinos clasificada con GMM

	GMM		
	Malo (Modelo)	Regular (Modelo)	Bueno (Modelo)
Malo (Real)	124	0	0
Regular (Real)	9	122	9
Bueno (Real)	0	0	96

Tabla 22. Matriz de confusión para vinos clasificada con Spectral

	Spectral		
	Malo (Modelo)	Regular (Modelo)	Bueno (Modelo)
Malo (Real)	117	7	0
Regular (Real)	20	113	7
Bueno (Real)	0	1	95

Un análisis superficial sugeriría que las técnicas de *k*-medias y Gauss tienen un buen rendimiento, debido a su exactitud del 98.3% y 95% respectivamente. Sin embargo, al observar los datos, es evidente que *k*-medias es la mejor opción, ya que, de los 140 vinos regulares, solo uno se vendió como vino de mala calidad (menos del 1% de los vinos regulares) y 5 vinos regulares se vendieron como vinos de buena calidad (3.47%). Lo cual es mucho mejor que lo que sucede con el método Gauss, donde se venden 9 vinos regulares como malos (6.43%) y otros 9 vinos regulares se venden como buenos (6.43%). Aunque en principio, la exactitud del método Spectral no es tan mala, es desfavorable en términos de clasificación de vinos, ya que, de considerar 124 vinos malos, clasifica 7 como regulares (5.6%), de los 140 vinos regulares clasifica 20 como malos (14.28%) y 7 regulares se clasifican como buenos (5.6%).

La tabla 23 establece una comparación de los resultados obtenidos de nuestro modelo en contraste con otras técnicas de clasificación [52]. El mejor resultado del modelo está en negrita, mientras que otros resultados están en cursiva.

Tabla 23. Comparativa entre los resultados obtenidos con otros modelos reportados.

Técnica	Exactitud (%)	Tipo de prueba
RDA	100	Subconjuntos con entrenamiento
QDA	99.4	Subconjuntos con entrenamiento
LDA	98.9	Subconjuntos con entrenamiento
kNN, Manhattan k=1	98.7	Subconjuntos con entrenamiento
Nuestro (k-means)	98.33	Subconjuntos sin entrenamiento
kNN, Euclidiana k=1	95.5	Subconjuntos con entrenamiento
<i>Our (Gauss)</i>	<i>95</i>	<i>Subconjuntos sin entrenamiento</i>
<i>Our (Spectral)</i>	<i>90.27</i>	<i>Subconjuntos sin entrenamiento</i>

La Tabla anterior muestra que los resultados del modelo no supervisado propuesto son tan buenos en cuanto a exactitud como los mejores modelos supervisados reportados en la literatura.

5.3.1.1. Análisis GPT de resultados de Vinos con k -medias

Las preguntas para esta instancia se orientaron para determinar las causas que podrían conducir al modelo a clasificar un vino de forma diferente a la clasificación propuesta por los expertos humanos.

- Determinar causas de clasificar un vino originalmente “Regular” en “Malo”
 - Consulta

La siguiente información representa los resultados de un análisis químico de vinos cultivados en la misma región de Italia, pero obtenidos de tres variedades diferentes:

Se carga la información de 7 elementos “Regulares” seleccionados aleatoriamente

¿Qué tienen en común con los siguientes datos?

Se carga la información de los elementos originalmente en la categoría de “Regular” que, de acuerdo con el modelo se deberían de clasificar en la categoría de “Malo”.

- Respuesta

Comparando los datos que me has proporcionado, se podría decir que los dos últimos vinos son considerados de peor calidad porque tienen una mayor acidez volátil, un menor grado alcohólico, un menor contenido en flavonoides y un menor índice OD280/OD315, que son algunos de los criterios que determinan la calidad en el vino. Además, tienen un color más intenso y un pH más alto, lo que puede afectar al equilibrio y la estabilidad del vino.

- Hallazgo

Tomando en cuenta la respuesta proporcionada se puede determinar que, en realidad, al hacer un análisis de las propiedades químicas del vino, sí pueden ser considerados de una calidad inferior.

- Determinar causas de clasificar un vino originalmente “Malo” en “Regular”

- Consulta

La siguiente información representa los resultados de un análisis químico de vinos cultivados en la misma región de Italia, pero obtenidos de tres variedades diferentes:

Se carga la información de 6 elementos clasificados como “Regulares” seleccionados aleatoriamente.

¿Qué tienen en común con el siguiente dato?

Se carga la información del único elemento originalmente “Malo” que, de acuerdo con el modelo se deberían de clasificar en la categoría de “Regular”.

- Respuesta

El último elemento tiene el valor más alto de OD280/OD315 de vinos diluidos, que es una medida de la calidad del vino. Un valor más alto de OD280/OD315 indica un mayor contenido de flavonoides y una mayor capacidad antioxidante del vino. Este elemento también tiene un valor más alto de prolina, que es un aminoácido que se asocia con la estabilidad y la longevidad del vino. La prolina también ayuda a preservar el color y el aroma del vino, lo que puede mejorar su calidad sensorial. Así que el último elemento podría ser el vino de mejor calidad de los que se muestran en la tabla.

- Hallazgo

Tomando en cuenta la respuesta proporcionada se puede determinar que, en realidad, al hacer un análisis de las propiedades químicas del vino, sí pueden ser considerados de una calidad superior.

- Determinar causas de clasificar un vino originalmente “Regular” en “Excelente”

- Consulta

La siguiente información representa los resultados de un análisis químico de vinos cultivados en la misma región de Italia, pero obtenidos de tres variedades diferentes:

Se carga la información del 10% de elementos clasificados como “Regulares” seleccionados aleatoriamente.

¿Qué tienen en común con el siguiente dato?

Se carga la información del único elemento originalmente “Regular” que, de acuerdo con el modelo se deberían de clasificar en la categoría de “Excelente”.

- Respuesta

El último vino podría ser considerado de mejor calidad porque tiene un mayor contenido de alcohol, que le da más cuerpo y estructura al vino, y una mayor intensidad de color, que le da más atractivo visual. También tiene un contenido de proantocianidinas similar al de los otros vinos, lo que le aporta beneficios antioxidantes y antiinflamatorios.

- Hallazgo

Tomando en cuenta la respuesta proporcionada se puede determinar que, en realidad, al hacer un análisis de las propiedades químicas del vino, sí puede ser considerado de una calidad superior, asimismo, la respuesta explica las razones por las cuales las propiedades químicas son importantes para que el vino sea de una calidad superior.

- Determinar causas de clasificar un vino originalmente “Excelente” en “Regular”

- Consulta

La siguiente información representa los resultados de un análisis químico de vinos cultivados en la misma región de Italia, pero obtenidos de tres variedades diferentes:

Se carga la información del 10% de elementos clasificados como “Excelentes” seleccionados aleatoriamente.

¿Qué tienen en común con el siguiente dato?

Se carga la información del único elemento originalmente “Excelente” que, de acuerdo con el modelo se debería de clasificar en la categoría de “Regular”.

- Respuesta

El vino que has mencionado podría considerarse de peor calidad porque tiene un nivel más bajo de dióxido de azufre libre (1.35 mg/L) que los otros vinos (entre 1.39 y 1.65 mg/L). El dióxido de azufre libre es un conservante que protege al vino de la oxidación y el crecimiento microbiano. Un nivel bajo de dióxido de azufre libre podría indicar una mayor susceptibilidad a la alteración o el deterioro del vino.

Además, el vino tiene un nivel más alto de acidez fija (2.32 g/L) que los otros vinos (entre 2.26 y 2.58 g/L). La acidez fija es la suma de los ácidos orgánicos presentes en el vino, principalmente el ácido tartárico y el ácido málico. La acidez fija influye en el sabor, el equilibrio y la estabilidad del vino. Un nivel alto de acidez fija podría hacer que el vino sea demasiado ácido o agrio.

- Hallazgo

Tomando en cuenta la respuesta proporcionada se puede determinar que, en realidad, al hacer un análisis de las propiedades químicas del vino, el único elemento considerado como “Excelente” sí puede ser considerado de una calidad inferior, asimismo, la respuesta explica las razones por las cuales las propiedades químicas son importantes para que el vino pueda ser considerado de una calidad inferior.

5.4. INVESTIGACIÓN

5.4.1. 60 universidades mexicanas

Con respecto a la evaluación de la base de datos ExECUM, se obtuvieron dos distintos tipos de resultados: aquellos que incluyen a la UNAM y aquellos que no la incluyen, esto es debido a que su presencia supone un desbalance para las instancias ya que esta institución se encuentra muy alejada de las demás en términos de tamaño y, por lo tanto, en su capacidad docente y de investigación, con lo cual las distancias entre las demás instituciones se acortan. Por esta razón los resultados que se mostrarán a continuación son obtenidos sin tomar en cuenta a la UNAM. Un claro ejemplo de esto se puede notar en la Figura 6 en comparativa con la Figura 7, ambas figuras se construyeron a partir de PCA. Es así que, de todas las dimensiones que corresponden a investigación, solo se usó la componente principal proyectada en el eje de las abscisas, mientras que para docencia todas sus dimensiones también fueron reducidas a una sola componente principal que se proyecta en el eje de las ordenadas.

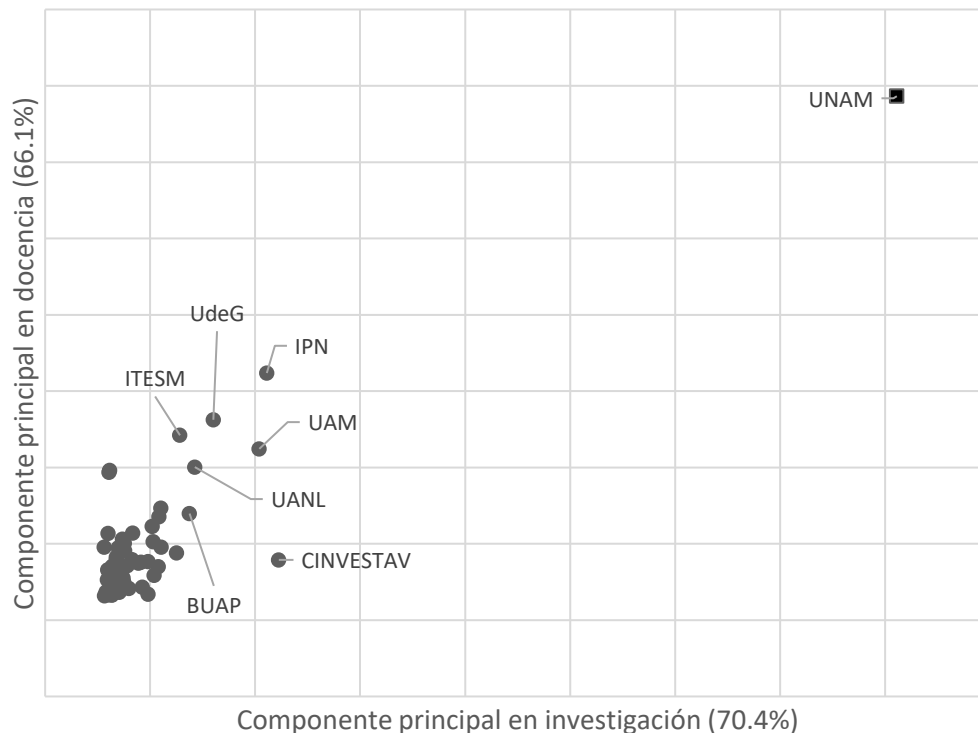


Figura 6. PCA promedio de 2009 a 2017 aplicado a docencia e investigación con UNAM

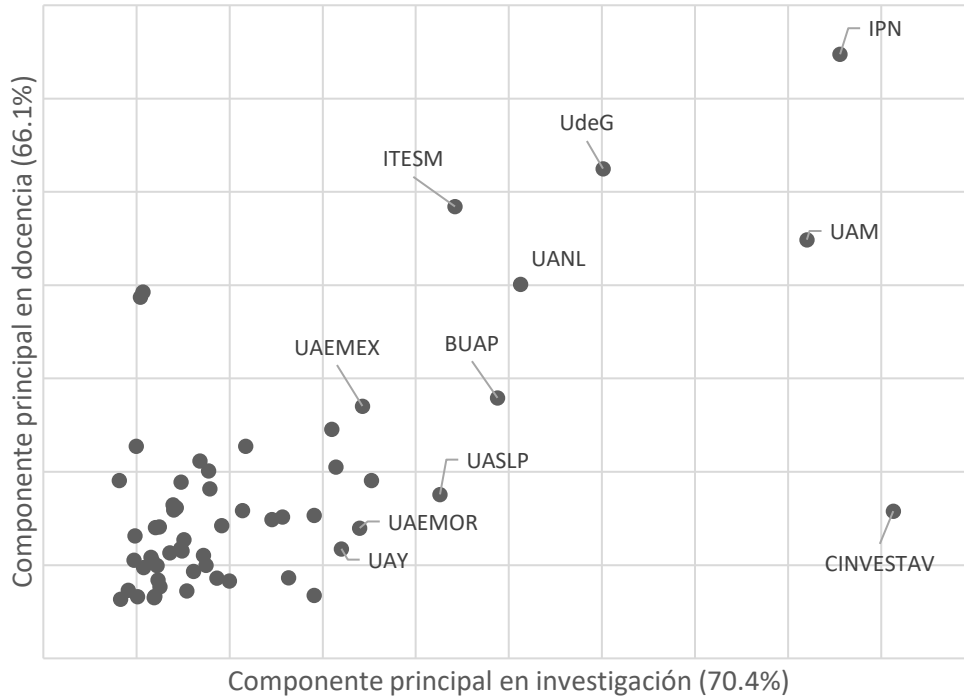


Figura 7. PCA promedio de 2009 a 2017 aplicado a docencia e investigación sin UNAM

Como se aprecia en las gráficas anteriores, la UNAM se encuentra muy alejada de las demás instituciones y ocasiona que todas ellas se vean como un solo grupo; sin embargo, al eliminar a la UNAM, se hace clara una separación entre las instituciones, a simple vista se puede ver que el IPN, UAM y CINVESTAV las mejores instituciones en investigación y en docencia son el IPN y la UdeG. Cabe mencionar que las gráficas anteriores solamente son representativas del total de los datos en un cierto porcentaje del total de la información disponible, debido a que, en el caso de la componente en investigación, PCA mantiene una representatividad total del 70.4%, mientras que en la componente principal de docencia la representatividad es de 66.1%.

5.4.2. Resultados de *k*-medias

Tabla 24. Resultados de 2009 a 2017 aplicando *k*-medias

Sumario de resultados de <i>k</i> -medias									
	2009	2010	2011	2012	2013	2014	2015	2016	2017
Más docencia, más investigación	IPN, UAM, UdeG	IPN, UAM, UdeG	IPN, UAM, UANL, UdeG	IPN, UAM, UANL, UdeG	IPN, UAM, UANL, UdeG	IPN, ITESM, UAM, UANL, UdeG	BUAP, IPN, ITESM, UAM, UANL, UdeG	BUAP, IPN, ITESM, UAM, UANL, UdeG	BUAP, IPN, ITESM, UAM, UANL, UdeG
Más docencia, menos investigación	ITESM, UANL, UPN, UVM	ITESM, UANL, UPN, UVM	ITESM, UPN, UVM	BUAP, ITESM, LASALLE, UABC, UAEMex, UAS, UP, UPN, UV, UVM	ITESM, UABC, UAEMex, UAS, UPN, UV, UVM	UAEMex, UPN, UVM	UABC, UAEMex, UPN, UV, UVM	LASALLE, UABC, UAS, UP, UPN, UTM, UV, UVM	LASALLE, UABC, UAS, UPN, UTM, UV, UVM
Menos docencia, más investigación	CINVESTAV	CINVESTAV	BUAP, CINVESTAV, UASLP	CINVESTAV	CINVESTAV, UASLP	BUAP, CINVESTAV, UASLP	CINVESTAV	CINVESTAV	UASLP, IBERO, CINVESTAV

La tabla anterior muestra la clasificación con *k*-medias. Los elementos no incluidos pertenecen al cuadrante 1, el cual incluye instituciones que no son buenas en ninguno de los dos rubros. Puede observarse a simple vista que hay instituciones que se mantienen durante todo el tiempo

Se aprecian otras instituciones que también varían su posición durante el periodo analizado, estas transiciones se pueden observar en la figura 4, donde se hace evidente el esfuerzo de las instituciones por mantener su estatus o mejorarlo durante todo el estudio.

Para representar las transiciones determinadas por el algoritmo *k*-medias de las Universidades, se usó el modelo matricial propuesto, quedado como se aprecia en la siguiente figura.

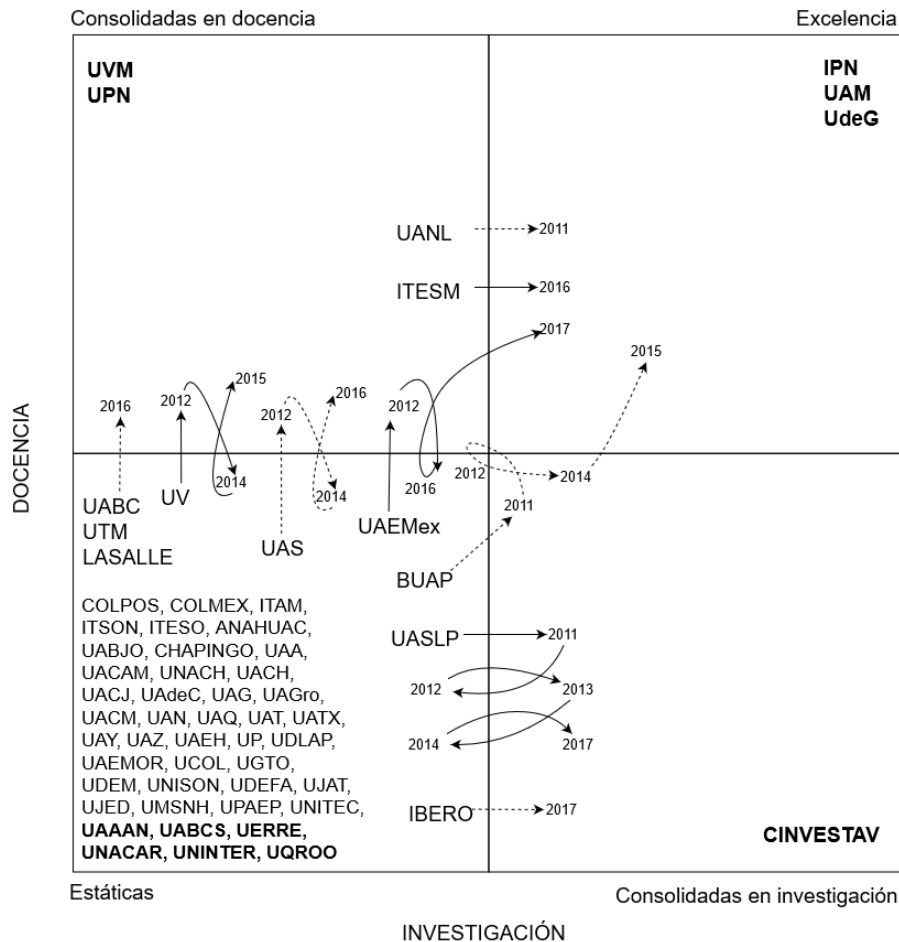


Figura 8. Representación visual de los resultados de 2009 a 2017 aplicando k-medias

Aquellas universidades que se remarcan en negro, son aquellas denominadas: estáticas, es decir, durante todo el periodo del análisis se mantuvieron en el mismo cuadrante.

5.4.3. Resultados de GMM

Tabla 25. Resultados de 2009 a 2017 aplicando GMM

Sumario de resultados de GMM									
	2009	2010	2011	2012	2013	2014	2015	2016	2017
Más docencia, más investigación	IPN, UAM, UdeG	IPN, UAM, UANL, UdeG	IPN, UAM, UANL, UdeG	IPN, UAM, UANL, UdeG	IPN, UAM	IPN, UAM	IPN, UAM	IPN, UAM, BUAP, ITESM, UAEMex, UANL, UdeG	IPN, UAM, BUAP, ITESM, UAEMex, UANL, UdeG
Más docencia, menos investigación	ITESM, UANL, UPN, UVM	ITESM, UPN, UVM	ITESM, UPN, UVM	ITESM, UPN, UVM	BUAP, ITESM, UANL, UPN, UAEMex, UVM, UdeG	BUAP, ITESM, LASALLE, UABC, UAEMex, UANL, UAS, UPN, UV, UVM, UdeG	BUAP, ITESM, LASALLE, UABC, UAEMex, UANL, UAS, UPN, UV, UVM, UdeG	LASALLE, UABC, UAS, UNITEC, UP, UPN, UV, UVM	
Menos docencia, más investigación	CINVESTAV	CINVESTAV, UAEMOR, UGTO, UV	CINVESTAV, UAEMOR, UGTO, UV	CINVESTAV, UAEMOR, UGTO, UV	CINVESTAV	CINVESTAV	CINVESTAV	CINVESTAV	CINVESTAV, IBERO, UASLP

La tabla anterior muestra la clasificación con GMM y los elementos no incluidos pertenecen a las instituciones estáticas, es decir, que no son buenas en ninguno de los dos rubros.

Cabe mencionar que los resultados de este algoritmo clasificador son muy similares a los del algoritmo *k*-medias.

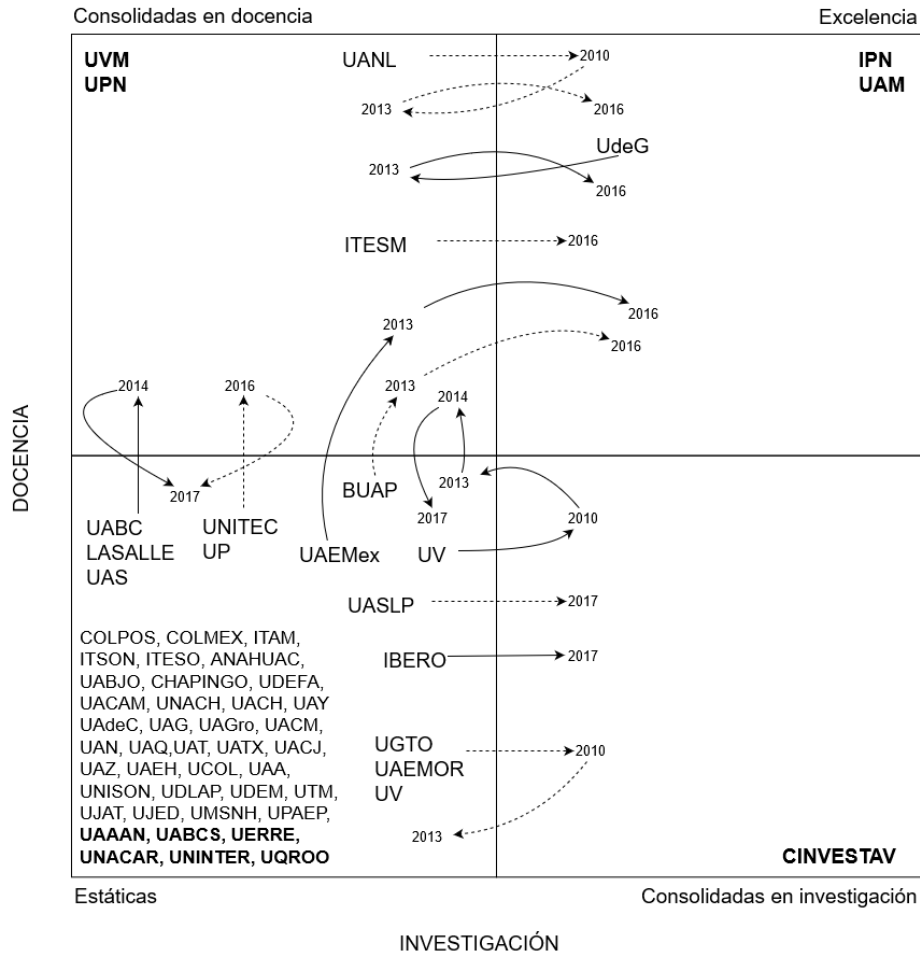


Figura 9. Representación visual de los resultados de 2009 a 2017 aplicando GMM

Para el caso del análisis desarrollado con este algoritmo, se tienen algunas ligeras diferencias como el contar con menos universidades consolidadas en el cuadrante de excelencia.

5.4.4. Resultados de agrupamiento espectral

Sumario de resultados de agrupamiento espectral										
	2009	2010	2011	2012	2013	2014	2015	2016	2017	
Más docencia, más investigación	ANAHUAC, ITSON, LASALLE, UAA, UABIO, UACH, UAG, UAGro, UAdec, UDEFA, UJAT, UJED, UNACH, UNITEC, UPN, UVM	ANAHUAC, ITSON, LASALLE, UABIO, UACH, UAG, UAN, UAGro, UAdec, UDEFA, UJAT, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ANAHUAC, ITSON, LASALLE, UAA, UABIO, UACH, UAGro, UAN, UAS, UAT, UATx, UAdec, UDEFA, UDEM, UJAT, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ANAHUAC, ITSON, LASALLE, UAA, UABIO, UACH, UAGro, UAN, UAT, UATx, UAdec, UDEFA, UDEM, UJAT, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ANAHUAC, IBERO, ITSON, LASALLE, UAA, UABIO, UACH, UAGro, UAN, UAS, UAT, UATx, UAdec, UDEFA, UDEM, UJAT, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ITSON, LASALLE, UAA, UABIO, UAG, UAGro, UAN, UAT, UATx, UAdec, UDEFA, UDEM, UJAT, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ANAHUAC, LASALLE, UABIO, UAG, UAdec, UDEM, UNACH, UNITEC, UPAEP, UPN, UVM	ANAHUAC, LASALLE, UABIO, UACH, UACI, UAG, UAN, UAT, UATx, UAY, UAZ, UAdec, UDEFA, UDEM, UJAT, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ANAHUAC, LASALLE, UAA, UABIO, UACI, UAG, UAN, UATx, UAdec, UDEM, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM	ITESO, LASALLE, UAA, UABIO, UACI, UAG, UAN, UATx, UAdec, UDEM, UJED, UNACH, UNITEC, UP, UPAEP, UPN, UVM
Menos docencia, más investigación	COLMEX, COLPOS, ITAM, UACAM, UACM, UDLAP	CHAPINGO, COLMEX, COLPOS, ITAM, UACAM, UACM, UDLAP	CHAPINGO, COLMEX, COLPOS, ITAM, UACAM, UACM, UDLAP	COLMEX, COLPOS, ITAM, UDLAP	COLPOS, ITAM	CHAPINGO, COLMEX, COLPOS, ITAM, UACM, UDLAP	CHAPINGO, COLMEX, COLPOS, ITAM, UACM, UAY, UCOL		CHAPINGO, COLPOS, UTM, ITAM, UCOL	
Menos docencia, menos investigación	CHAPINGO, ITESO, UAAAN, UABCS, UAN, UATx, UDEM, UERRE, UNACAR, UNINTER, UNINTER, UQROO	ITESO, UAAAN, UABCS, UATx, UTM, UDEM, UERRE, UNACAR, UNINTER, UQROO	ITESO, UAAAN, UABCS, UATx, UTM, UDEM, UERRE, UNACAR, UNINTER, UQROO	CHAPINGO, ITESO, UAAAN, UABCS, UTM, UACAM, UACM, UERRE, UNACAR, UNINTER, UQROO	CHAPINGO, COLMEX, ITESO, UAAAN, UABCS, UTM, UACAM, UERRE, UNACAR, UNINTER, UNINTER, UQROO	ITESO, UAAAN, UABCS, UTM, UACAM, UERRE, UNACAR, UNINTER, UQROO	ITESO, ITSON, UAAAN, UABCS, UTM, UACAM, UAGro, UAN, UATx, UDEFA, UDLAP, UERRE, UJED, UNACAR, UNINTER, UQROO	CHAPINGO, COLMEX, COLPOS, ITAM, ITESO, ITSON, UTM, UAA, UAAAN, UABCS, UACAM, UACM, UAGro, UAGro, UACM, UACM, UAGro, UERRE, UERRE, UCOL, UDLAP, UNINTER, UQROO	COLMEX, ITSON, UAAAN, UABCS, UACAM, UACM, UDEFA, UDLAP, UERRE, UNACAR, UNINTER, UQROO	

Figura 10. Resultados de 2009 a 2017 aplicando agrupamiento espectral

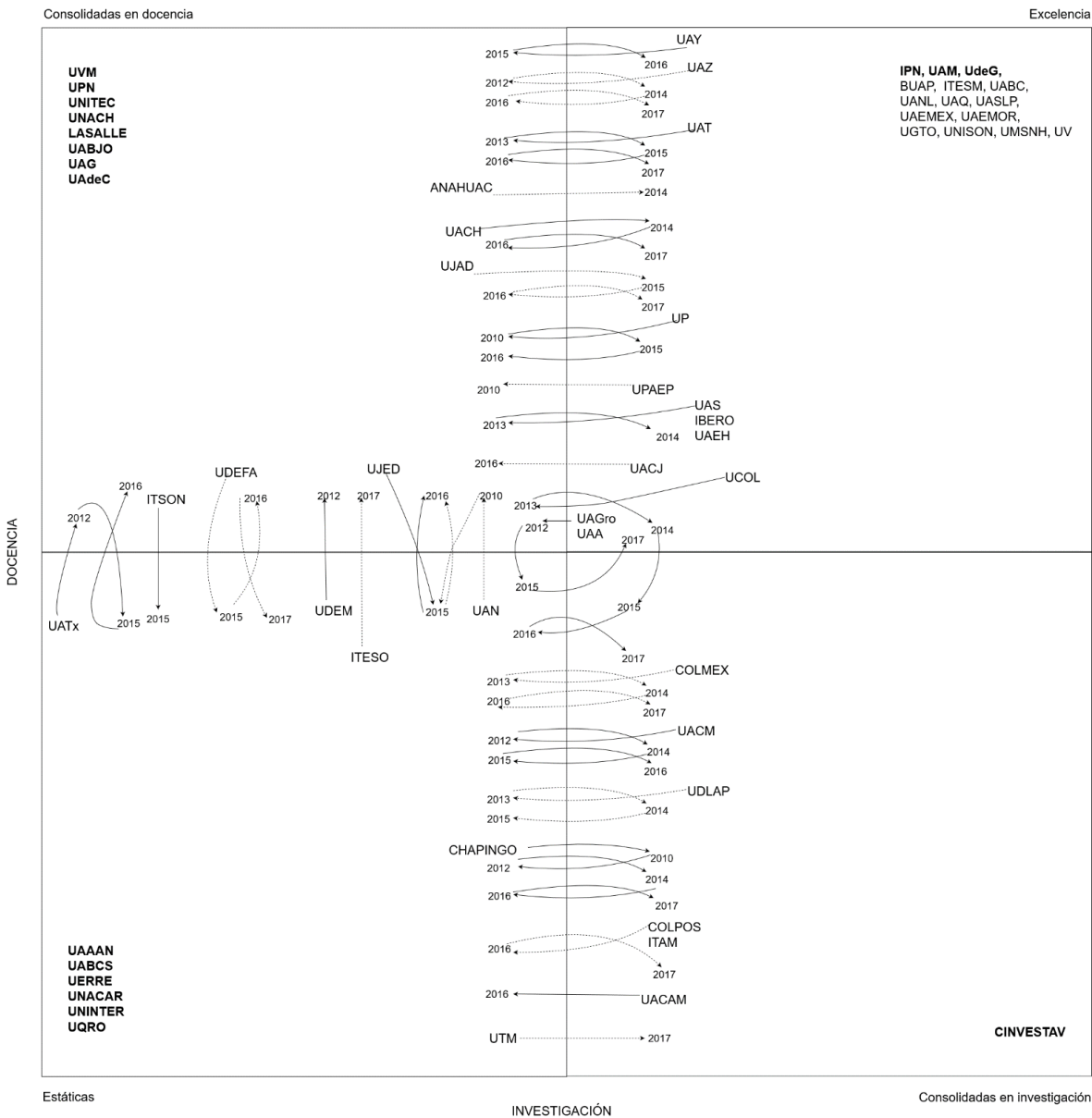


Figura 11. Representación visual de los resultados de 2009 a 2017 aplicando agrupamiento espectral

En lo que respecta a los resultados obtenidos por los algoritmos k -medias y GMM, estos guardan cierta relación, pues resultan que ambos son consistentes el uno con el otro; sin embargo, los resultados del agrupamiento espectral son complementarios a las dos técnicas ya mencionadas y analizadas, esto se debe a que el agrupamiento espectral enfatiza los cambios en las universidades menos favorecidas; donde se muestra que algunas universidades del primer cuadrante quieren mejorar, ya sea en docencia o en investigación, aunque sus esfuerzos son más modestos.

5.4.5. Análisis de resultados con GPT

Los hallazgos anteriormente mencionados en este trabajo y en [53], se limitan a qué tipo de grupo pertenecen las Universidades, así como las transiciones por las que pasan cada una de ellas desde los años 2009 al 2017. No obstante, es posible hacer un análisis más profundo si se usa GPT como herramienta auxiliar para este efecto, así, considerando la información obtenida de los algoritmos se analizaron los siguientes casos:

- Instituciones consolidadas en docencia e investigación

Lejos de obviar la información de que aquellas instituciones que se encuentran agrupadas en el mismo recuadro comparten las características de docencia e investigación, es posible realizar una consulta por medio de GPT sobre qué tienen en común las instituciones IPN, UAM y UdeG, las cuales coinciden en estar consolidadas en docencia e investigación.

- Consulta:

¿Qué tienen en común el Instituto Politécnico Nacional, la Universidad Autónoma Metropolitana y la Universidad de Guadalajara?

- Respuesta

El Instituto Politécnico Nacional (IPN), la Universidad Autónoma Metropolitana (UAM) y la Universidad de Guadalajara (UDG) son instituciones educativas prominentes en México. Aunque cada una de ellas tiene sus características distintivas, comparten algunas similitudes y puntos en común:

Son instituciones públicas: Tanto el IPN, la UAM y la UDG son universidades públicas en México. Esto significa que están financiadas y administradas por el gobierno, lo que les permite ofrecer educación de calidad a un costo más accesible para los estudiantes.

Ofrecen una amplia variedad de programas académicos: Las tres instituciones tienen una oferta académica diversa que abarca diversas áreas de estudio, incluyendo ciencias exactas, ingenierías, ciencias sociales, humanidades, artes y más. Cuentan con múltiples facultades y escuelas que brindan programas de licenciatura, maestría y doctorado.

Son reconocidas a nivel nacional e internacional: El IPN, la UAM y la UDG gozan de prestigio y reconocimiento tanto en México como en el extranjero. Sus programas académicos, investigaciones y contribuciones al desarrollo científico y tecnológico del país las han posicionado como referentes en sus respectivas áreas.

Fomentan la investigación y la innovación: Estas instituciones promueven la investigación y la innovación como parte fundamental de su labor académica. Cuentan con centros de investigación, laboratorios y grupos de investigación en los que se llevan a cabo proyectos científicos y tecnológicos de vanguardia.

Contribuyen al desarrollo regional y nacional: El IPN, la UAM y la UDG desempeñan un papel importante en el desarrollo económico, social y cultural de sus respectivas regiones y del país en general. A través de la formación de profesionales altamente capacitados, la generación de conocimiento y la vinculación con el sector productivo, contribuyen al crecimiento y la competitividad de México.

- Hallazgo

De lo anterior se pueden deducir las siguientes relaciones no consideradas previamente:

- Son instituciones públicas
- Fomentan la investigación y la innovación
- Contribuyen al desarrollo regional y nacional

- Instituciones con más docencia y menos investigación

Debido a la cantidad de transiciones existentes en este cuadrante entre las diferentes técnicas de clasificación no supervisada, se hizo un conteo y se determinaron algunas instituciones con mayor presencia durante el periodo analizado, las cuales son: Instituto Tecnológico de Estudios Superiores de Monterrey, la Universidad Autónoma de Nuevo León, la Universidad Pedagógica Nacional y la Universidad del Valle de México, de esta forma se generó la siguiente información.

- Consulta

¿Qué tienen en común el Instituto Tecnológico de Estudios Superiores de Monterrey, la Universidad Autónoma de Nuevo León, ¿la Universidad Pedagógica Nacional y la Universidad del Valle de México?

○ Respuesta

Ofrecen una amplia variedad de programas académicos: Estas instituciones cuentan con una diversa oferta educativa que abarca diferentes áreas de estudio, incluyendo ciencias exactas, ingenierías, ciencias sociales, humanidades, educación, negocios y más. Tienen múltiples facultades y escuelas que ofrecen programas de licenciatura, maestría y doctorado.

Son instituciones reconocidas: El ITESM, la UANL, la UPN y la UVM son instituciones reconocidas a nivel nacional e internacional. Han establecido su reputación a través de la calidad de sus programas académicos, su excelencia educativa y la formación de profesionales altamente capacitados.

Fomentan la investigación y la innovación: Estas instituciones promueven la investigación científica, la innovación y el desarrollo tecnológico. Cuentan con centros de investigación, laboratorios y grupos de investigación que contribuyen al avance del conocimiento en diversas disciplinas.

Vinculación con el sector empresarial: El ITESM y la UVM, en particular, tienen una fuerte vinculación con el sector empresarial. Estas instituciones fomentan la colaboración con empresas y organizaciones para promover el emprendimiento, la transferencia de tecnología y la generación de proyectos conjuntos que impulsan el desarrollo económico.

Compromiso social y educación: La UPN y la UANL, en especial, tienen un enfoque en la educación y la formación de profesionales en el ámbito pedagógico y educativo. Estas instituciones se dedican a la formación de docentes y profesionales comprometidos con la educación y el desarrollo social.

○ Hallazgo

De lo anterior se pueden deducir las siguientes relaciones no consideradas previamente:

- El ITESM y la UVM, en particular, tienen una fuerte vinculación con el sector empresarial.
- La UPN y la UANL, en especial, tienen un enfoque en la educación y la formación de profesionales en el ámbito pedagógico y educativo.

Esto permite explicar claramente la razón por la cual estas cuatro instituciones se encuentran en el grupo que tienen más docencia y menos investigación.

- Instituciones con menos docencia y más investigación

De manera análoga al análisis anterior, a causa de la cantidad de transiciones en el cuadrante, se hizo una contabilización de las instituciones con mayor aparición del periodo analizado, obteniendo las siguientes: Benemérita Universidad Autónoma de Puebla, Centro de Investigación y de Estudios Avanzados del IPN y la Universidad Autónoma de San Luis Potosí.

- Consulta

¿Qué tienen en común las siguientes instituciones?

Benemérita Universidad Autónoma de Puebla

Centro de Investigación y de Estudios Avanzados del IPN

Universidad Autónoma de San Luis Potosí

- Respuesta

Las instituciones mencionadas comparten las siguientes características comunes:

Son instituciones educativas de nivel superior en México. Tanto la Benemérita Universidad Autónoma de Puebla (BUAP), el Centro de Investigación y de Estudios Avanzados del IPN (CINVESTAV) y la Universidad Autónoma de San Luis Potosí (UASLP) son reconocidas como instituciones de educación superior en el país.

Son instituciones públicas. Las tres instituciones son financiadas y administradas por el gobierno mexicano, lo que implica que ofrecen educación asequible y accesible a los estudiantes.

Tienen una amplia oferta académica. Cada una de estas instituciones cuenta con una variedad de facultades y programas académicos en diferentes áreas del conocimiento, como ciencias, ingeniería, humanidades, artes, ciencias sociales, entre otros.

Son reconocidas por su calidad educativa. Las tres instituciones tienen una reputación establecida y son reconocidas tanto a nivel nacional como internacional por su excelencia académica y la calidad de sus programas de estudio.

Contribuyen a la investigación y desarrollo. Estas instituciones promueven la investigación científica y tecnológica, fomentando la generación de conocimiento y la innovación en diversos campos.

Cuentan con infraestructura y recursos adecuados. Estas instituciones tienen campus extensos, bibliotecas, laboratorios, centros de investigación y otras instalaciones necesarias para el desarrollo de las actividades académicas.

Aunque estas instituciones comparten características comunes, también tienen particularidades y enfoques educativos propios que las distinguen entre sí.

- Hallazgo

Si bien los resultados entre esta respuesta y la anterior pueden resultar muy similares, dentro de las respuestas existe un apartado que marca la diferencia y es que en este caso no se menciona que estas instituciones cuenten con una amplia comunidad estudiantil.

5.5. Clasificación humana previa

5.5.1. Cuadernos

La Procuraduría Federal del Consumidor (PROFECO) tiene como objetivo fundamental defender los derechos de los consumidores mexicanos otorgándoles las herramientas necesarias para ejercer sus derechos de manera efectiva y fomentar la confianza ciudadana en el mercado. La PROFECO promueve un modelo de consumo responsable, basado en la información, la sostenibilidad, la seguridad y la salud. De esta manera, busca corregir las injusticias que puedan surgir en el mercado, fortalecer el mercado interno y contribuir al bienestar general de la población [54]. En este sentido, la PROFECO dispone con el Laboratorio Nacional de Protección al Consumidor (LNPC), cuya función es elaborar y publicar estudios de calidad sobre la calidad de productos de consumo generalizado, evaluando su impacto en la economía, seguridad y nutrición [55]. Uno de los tantos estudios publicados por ellos es el de útiles escolares, en el cual se incluyen comparaciones de los cuadernos disponibles en el mercado mexicano [46].

De esta forma y retomando estos estudios, para obtener esta base de datos se generó una tabla con las calificaciones propuestas por PROFECO en sus estudios, la cual se resume a continuación:

Tabla 26. Clasificación de los cuadernos reportados por PROFECO

Modelo	Calificación asignada por PROFECO
Scribe 7510	Muy bueno
Estrella Único	Muy bueno
Manufacturas 8-A	Muy bueno
Estrella/ Star Kid	Bueno
Sazz	Suficiente
U-Pak	Suficiente
Norma Cosido	Excelente
Estrella Cosido	Excelente
Office Max Espiral	Muy bueno
Scribe Excellence	Muy bueno
Scribe In Black	Muy bueno
Manufacturas 8-A	Muy bueno
First Class 5038	Muy bueno
First Class 5040	Muy bueno
First Class 4517	Muy bueno
Norma Jean Book	Muy bueno
Scribe Clásico Cosido	Muy bueno
Scribe Kids Cosido	Muy bueno
First Class 4507	Bueno
Estrella Drive Linea A	Bueno
Scribe In Colors 1070	Bueno
First Class Cuaderno Cosido 9408	Bueno
First Class Profesional 5039	Bueno
Scribe Poly Cover 2050	Bueno
Norma	Bueno
Estrella 0520	Bueno
First Class Cosido Profesional 9214	Bueno
Norma Click Urbana	Bueno
U-Pak 1157525405	Bueno
Norma Kiut	Bueno
Norma Girly Notes	Suficiente
Norma Uno	Suficiente
Nine To Five	Suficiente

5.5.2. Análisis de clasificación con GPT

- Análisis de la categoría “suficiente”

- Consulta 1

¿Qué tienen en común los siguientes cuadernos?

Norma Girly Notes

Norma Uno

Nine To Five

- Respuesta

Los tres cuadernos mencionados, Norma Girly Notes, Norma Uno y Nine To Five, comparten algunas características comunes:

Marca: Todos son productos de la marca Norma, lo que indica que pertenecen a la misma línea o serie de productos de la compañía.

Función: Los tres cuadernos están diseñados para tomar notas o escribir, lo que implica que son adecuados para uso escolar, de oficina o personal.

- Consulta 2

¿Qué tienen en común los siguientes cuadernos?

Sazz

U-Pak

- Respuesta

Lo que tienen en común los cuadernos Sazz y U-Pak es que son marcas mexicanas de productos escolares que ofrecen diferentes tipos de encuadernación, diseño y calidad. Ambas marcas tienen páginas web donde se pueden consultar sus catálogos y precios.

- Hallazgo

Las dos consultas anteriores derivaron en una investigación adicional en la cual se detectó que, en el caso de las marcas de Norma, los cuadernos calificados como “suficiente” corresponden a la línea “económica” de Norma y a la línea destinada al público femenino, en el cual se prioriza el diseño sobre la funcionalidad. Por otro lado, se encontró que todas las marcas calificadas como “suficiente” son marcas mexicanas producidas en el país.

- Análisis de la categoría “Excelente”

Posterior al análisis de los cuadernos clasificados como “Suficiente”, se analizaron aquellos cuadernos que caen dentro de la categoría “Excelente”, los cuales resultaron ser solo dos.

- Consulta

¿Qué tienen en común las siguientes marcas de cuadernos?:

Norma Cosido

Estrella Cosido

- Respuesta

Las marcas de cuadernos Norma Cosido y Estrella Cosido tienen en común el estilo de encuadernación llamado "cosido". Aquí tienes algunas características comunes:

Encuadernación cosida: Ambas marcas utilizan el método de encuadernación cosida para sus cuadernos. Esto implica que las hojas están unidas mediante un cosido con hilo o hilo metálico en el lomo del cuaderno. Esta técnica proporciona una mayor durabilidad y resistencia al cuaderno, permitiendo que las páginas se mantengan en su lugar incluso con un uso frecuente.

Diseño tradicional: Tanto Norma Cosido como Estrella Cosido suelen ofrecer cuadernos con un diseño clásico y tradicional. Estos cuadernos suelen tener tapas duras o semiduras, con colores sobrios y discretos, adaptados a un estilo más formal y profesional.

- Hallazgo

Aquellos cuadernos cosidos con un diseño tradicional son considerados de mejor calidad.

5.6. Clasificación de resultados de cuestionarios

5.6.1. Implementación de nuevos modelos educativos

Otra de las aplicaciones del modelo propuesto es el análisis y clasificación de grupos de interés derivados de cuestionarios contestados por un público objetivo. Para este caso en particular es de la viabilidad sobre la implementación de nuevos modelos educativos en una Universidad. Las preguntas de este cuestionario fueron:

1. ¿Consideras que el profesor apoyó en tus dudas de clase proporcionando orientación?
2. ¿Consideras que la estrategia del docente fue adecuada en el desarrollo de tus actividades en la clase de Comercio Electrónico?
3. De los siguientes aparatos electrónicos, ¿cuál utilizaste durante las clases de Comercio Electrónico?
4. ¿Fueron funcionales las clases a distancia?
5. ¿La aplicación de Classroom para el desarrollo de la clase de Comercio electrónico fue adecuado para tu aprendizaje?
6. ¿La aplicación Google Meet para el desarrollo de la clase de Comercio electrónico fue adecuado para tu aprendizaje?

Haciendo un análisis de Pearson en los previo se puede determinar si existe alguna correlación entre las respuestas obtenidas:

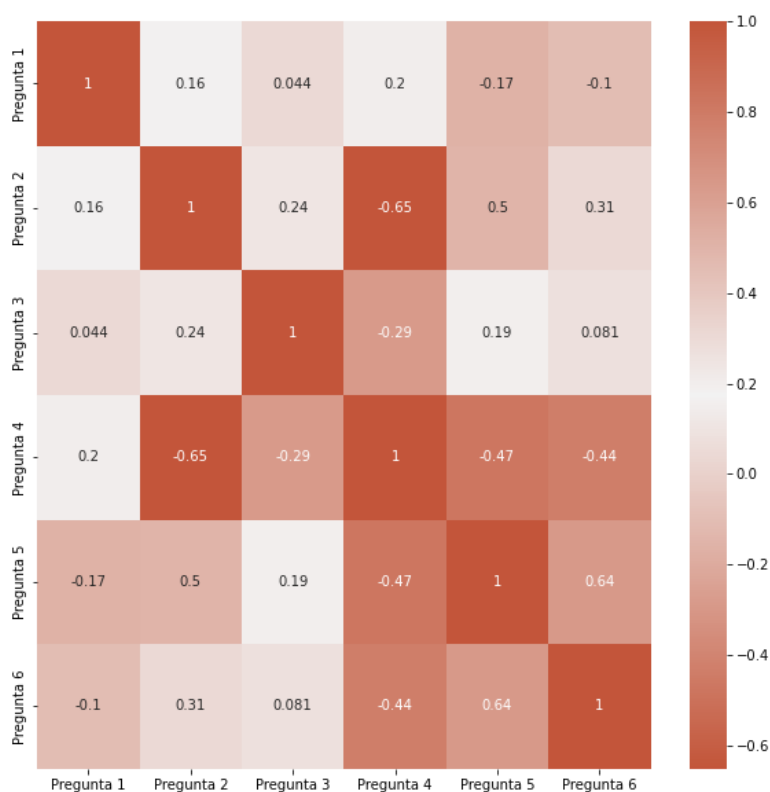


Figura 12. Matriz de correlación de los cuestionarios

Debido a este análisis previo, es posible determinar que existen las siguientes correlaciones:

- Moderada negativa entre la pregunta 2 y 4.
Lo cual quiere decir que, aunque consideran que la estrategia docente fue adecuada, las clases a distancia no. Esto abre un panorama a determinar las posibles razones, como la ausencia, falla o dificultad de acceso a internet y posibles estrategias para resolver este problema.
- Moderada positiva entre la pregunta 5 y 6.
Esto indica que la aplicación de las herramientas digitales de Google Classroom y Meet fueron apropiadas.

Posterior a ello, se determinó el número adecuado de clases usando el método del codo, resultando ser 4 (ver línea verde).

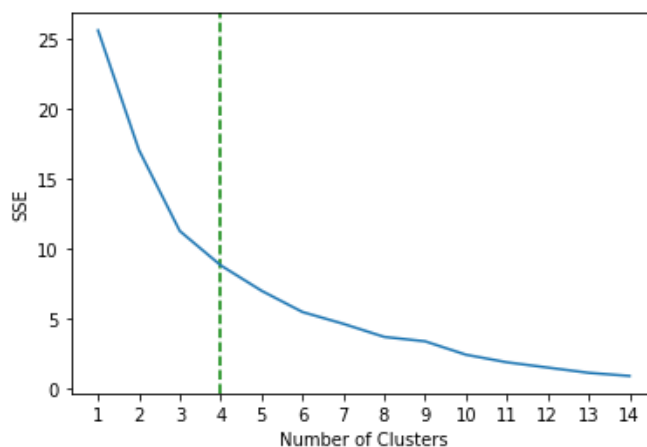


Figura 13. Número adecuado de clases para clasificar los resultados.

El número de clases se ingresó al modelo teniendo resultados iguales en cuanto a la clasificación de la base de datos y finalmente se usaron métodos de reducción de dimensiones, conservando un 93% de varianza de los datos usando tres dimensiones, esto solo fue con fines una mejor visualización para el público en general usando un diagrama de dispersión.

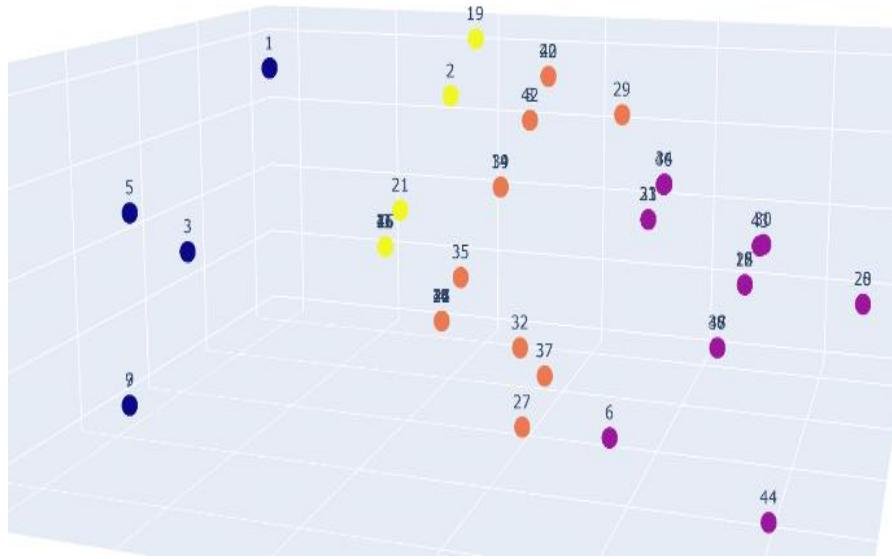


Figura 14. Clasificación de los encuestados en 3 dimensiones.

5.6.2. Determinación de etiquetado con GPT

- Consulta

La siguiente información representa los resultados de una clasificación de estudiantes que contestaron una encuesta, ¿Cómo se pueden etiquetar a los grupos generados?:

Se carga la información de la encuesta y la clasificación generada

- Respuesta

Un posible etiquetado para los datos ingresados sería:

Categoría 1: Muy satisfecho con profesor, estrategia y herramientas digitales.

Categoría 2: Satisfecho con profesor y estrategia, con variaciones en medios y herramientas.

Categoría 3: Satisfecho con profesor, pero con dudas sobre la estrategia.

Categoría 4: Insatisfecho con profesor y clases a distancia.

- Hallazgo

Gracias a esta recomendación de etiquetado es posible separar a las categorías con nombre: Categoría 1 (Morado): Muy satisfecho con profesor, estrategia y

herramientas digitales. Categoría 2 (Naranja): Satisfecho con profesor y estrategia, con variaciones en medios y herramientas. Categoría 3 (Amarillo): Satisfecho con profesor, pero con dudas sobre la estrategia. Categoría 4 (Azul): Insatisfecho con profesor y clases a distancia. Y también da la apertura para ofrecer una solución personalizada que se adapte a las necesidades y expectativas de cada alumno, principalmente a aquellos dentro de la categoría 4, para abordar diversas necesidades y niveles de satisfacción de los estudiantes, generando así un aprendizaje adaptativo y significativo.

Gracias a lo anterior se puede favorecer la participación activa y autonomía del alumno en el proceso de aprendizaje, obteniendo así, conclusiones de mayor relevancia.

Es importante mencionar que esta metodología puede ser aplicada en contextos similares y cuestionarios en general.

6. CONCLUSIONES Y APORTACIONES

6.1. Conclusiones

En este proyecto se realizó un estudio exhaustivo de las características generales de los modelos de clasificación no supervisada y se ha propuesto una metodología que se adapta a diferentes tipos de instancias, permitiendo así el desarrollo e implementación de un modelo avanzado de reconocimiento de patrones no supervisado que permite resolver aplicaciones en diferentes campos de estudio.

El modelo propuesto se fundamenta en el uso de diversos algoritmos de clustering para clasificar los datos, y en otras técnicas de aprendizaje no supervisado que permiten evaluar la calidad de los resultados y, a su vez, reportar los hallazgos de una manera más efectiva y comprensible.

Para el análisis de los resultados del modelo, se utilizó la tecnología de GPT, que permitió generar textos descriptivos y explicativos de los patrones encontrados.

La calidad del modelo se evaluó usando instancias de clasificación bien conocidas como: vinos, cáncer de mama y hepatitis; instancias de investigación reportadas recientemente en la literatura, como lo son aquellas derivadas de estudios de PROFECO y ranking de universidades; así como aquellas generadas por estudios propios. Se compararon los resultados obtenidos por la propuesta con los resultados reportados en la literatura y se encontró que el modelo supera a los métodos existentes en términos de exactitud, robustez, información proporcionada y capacidad de generalización.

Gracias al modelo propuesto, se pudieron encontrar patrones y nueva información no reportada anteriormente en la literatura y se da la apertura, partiendo desde este proyecto, a nuevas líneas de investigación, así como el desarrollo de modelos más sofisticados o específicos de acuerdo con la instancia que se busca resolver.

Actualmente se está desarrollando un modelo comparativo tomando en cuenta diferentes modelos de inteligencia artificial basados en el aprendizaje profundo validando la calidad de sus resultados e implicaciones.

Como trabajo a futuro, se recomienda ampliar el conjunto de datos, las aplicaciones a considerar, así como explorar otras técnicas de clustering y medidas de similitud que puedan mejorar el rendimiento del modelo.

6.2. Aportaciones

Con base en esta investigación:

- Se aceptó la publicación de un resumen llamado: *An Unbiased Analysis of Soft Drinks in Mexico Using Artificial Intelligence*. En la revista *Annals of Nutrition and Metabolism*, 2024. IF: 3.2.
- Se aceptó la publicación de un resumen llamado: *Nutritional Classification of Maria Cookies Using Artificial Intelligence*. En la revista *Annals of Nutrition and Metabolism*, 2024. IF: 3.2.
- Se presentó el artículo llamado: *Nutrition Analysis and Classification of Dry Cat Foods Using Pattern Recognition Techniques*. En la revista *XIKUA Boletín Científico de la Escuela Superior de Tlahuelilpan*, 2024.
- Se publicó un resumen llamado: *Comparative Study with Pattern Recognition to Determine the Quality of the Most Consumed Flavored Milks in Mexico*. En la revista *Annals of Nutrition and Metabolism*, 2023. IF: 3.2.
- Se publicó el artículo llamado: *Unsupervised learning model to determine a classification based on the cost-benefit ratio of school notebooks in Mexico*. En el *ECORFAN Journal-Taiwan*, 2023.
- Se publicó el artículo llamado: *A Comparative Ranking Model among Mexican Universities Using Pattern Recognition. Mathematics*. En la revista *Mathematics*, 2021. IF: 2.3.
- Se publicó el artículo llamado *A Model Based on Clustering Techniques for Type I and Type II Errors in Some Applications in Business and Medical Analysis*. En la revista *Fuzzy Economic Review*, 2020.
- Se publicó un resumen llamado: *Comparative Study to Determine the Classification Among the Most Consumed Creams in Mexico Using Pattern Recognition*. En la revista *Annals of Nutrition and Metabolism*, 2020. IF: 3.2.
- Se publicó un resumen llamado: *Pattern Recognition to Determine the Quality of the Most Consumed Breakfast Cereals in Mexico*. En la revista *Annals of Nutrition and Metabolism*, 2020. IF: 3.2.
- Se presentó el artículo llamado: *A Model Based on Clustering Techniques for Type I and Type II Errors in a Business Application*. En el *XX SIGEF Congress* 2019.
- Se publicó el artículo llamado: *Proposal of a simple recommendation system for small and medium enterprises for decision making based on unsupervised*. En el *Journal of Bussines Development Strategies*, 2019.

REFERENCIAS

- [1] B. D. Ripley, *Pattern recognition and neural networks*, 1. paperback ed. 1997, Reprinted 2009. Cambridge: Cambridge Univ. Press, 2009.
- [2] A. Carpio, A. Simón, A. Torres, and L. F. Villa, ‘Pattern recognition in data as a diagnosis tool’, *J.Math.Industry*, vol. 12, no. 1, p. 3, Dec. 2022, doi: 10.1186/s13362-022-00119-w.
- [3] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, ‘Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda’, *J Ambient Intell Human Comput*, vol. 14, no. 7, pp. 8459–8486, Jul. 2023, doi: 10.1007/s12652-021-03612-z.
- [4] J. Wang, H. Li, C. Dong, J. Wang, B. Zheng, and T. Xing, ‘An Underwater Side-Scan Sonar Transfer Recognition Method Based on Crossed Point-to-Point Second-Order Self-Attention Mechanism’, *Remote Sensing*, vol. 15, no. 18, p. 4517, Sep. 2023, doi: 10.3390/rs15184517.
- [5] M. Ala’raj, M. F. Abbod, and M. Majdalawieh, ‘Modelling customers credit card behaviour using bidirectional LSTM neural networks’, *J Big Data*, vol. 8, no. 1, p. 69, Dec. 2021, doi: 10.1186/s40537-021-00461-7.
- [6] J. J. Bird, M. Pritchard, A. Fratini, A. Ekart, and D. R. Faria, ‘Synthetic Biological Signals Machine-Generated by GPT-2 Improve the Classification of EEG and EMG Through Data Augmentation’, *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 3498–3504, Apr. 2021, doi: 10.1109/LRA.2021.3056355.
- [7] J. Goodhue and Y. Wei, ‘Classification of Trademark Distinctiveness using OpenAI GPT 3.5 model’, *Available at SSRN 4351998*, 2023.
- [8] M. Soni and V. Wade, ‘Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms’, *arXiv preprint arXiv:2303.17650*, 2023.
- [9] D. Castillo, ‘Clustering Documents with OpenAI, LangChain, and HDBSCAN’, Clustering Documents with OpenAI, LangChain, and HDBSCAN. Accessed: Jan. 06, 2024. [Online]. Available: <https://dylancastillo.co/clustering-documents-with-openai-langchain-hdbscan/>
- [10] S. Liu and C. G. Healey, ‘Abstractive Summarization of Large Document Collections Using GPT’, 2023, doi: 10.48550/ARXIV.2310.05690.
- [11] X. Zhu, Z. Ghahramani, and J. D. Lafferty, ‘Semi-supervised learning using gaussian fields and harmonic functions’, in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 2003, pp. 912–919.

- [12] J. F. Cruz, P. L. Velázquez, M. A. G. ANDRADE, S. G. D. L. C. SILVA, and E. A. R. GARCÍA, ‘Un sistema clasificador utilizando coloración de gráficas suaves’, *Revista de Matemática: Teoría y Aplicaciones*, vol. 24, no. 1, pp. 129–156, 2017.
- [13] X. Zhu, Z. Ghahramani, and J. D. Lafferty, ‘Semi-supervised learning using gaussian fields and harmonic functions’, in *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 2003, pp. 912–919.
- [14] B. D. Ripley, *Pattern recognition and neural networks*, 1. paperback ed. 1997, Reprinted 2009. Cambridge: Cambridge Univ. Press, 2009.
- [15] OpenAI, ‘GPT-4’. Accessed: May 24, 2023. [Online]. Available: <https://openai.com/product/gpt-4>
- [16] Y. Wardat, M. A. Tashtoush, R. AlAli, and A. M. Jarrah, ‘ChatGPT: A revolutionary tool for teaching and learning mathematics’, *EURASIA J Math Sci Tech Ed*, vol. 19, no. 7, p. em2286, Jul. 2023, doi: 10.29333/ejmste/13272.
- [17] D. Gašević, G. Siemens, and S. Sadiq, ‘Empowering learners for the age of artificial intelligence’, *Computers and Education: Artificial Intelligence*, p. 100130, Feb. 2023, doi: 10.1016/j.caeai.2023.100130.
- [18] M. Dowling and B. Lucey, ‘ChatGPT for (Finance) research: The Bananarama Conjecture’, *Finance Research Letters*, vol. 53, p. 103662, May 2023, doi: 10.1016/j.frl.2023.103662.
- [19] S. S. Biswas, ‘Role of Chat GPT in Public Health’, *Ann Biomed Eng*, vol. 51, no. 5, pp. 868–869, May 2023, doi: 10.1007/s10439-023-03172-7.
- [20] B. Guo *et al.*, ‘How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection’. arXiv, Jan. 18, 2023. Accessed: May 24, 2023. [Online]. Available: <http://arxiv.org/abs/2301.07597>
- [21] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, ‘A SWOT analysis of ChatGPT: Implications for educational practice and research’, *Innovations in Education and Teaching International*, pp. 1–15, Mar. 2023, doi: 10.1080/14703297.2023.2195846.
- [22] K. Roose, ‘The brilliance and weirdness of ChatGPT’, *The New York Times*, 2022, Accessed: May 23, 2023. [Online]. Available: <https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html>
- [23] Raf, ‘What are tokens and how to count them? | OpenAI Help Center’. Accessed: Jun. 12, 2023. [Online]. Available: <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them>
- [24] OpenAI, ‘Pricing’. Accessed: Jun. 12, 2023. [Online]. Available: <https://openai.com/pricing>

- [25] J. Shantanu, ‘Tiktoken’. OpenAI, Jun. 12, 2023. Accessed: Jun. 12, 2023. [Online]. Available: <https://github.com/openai/tiktoken>
- [26] OpenAI, ‘Tokenizer’. Accessed: Jun. 12, 2023. [Online]. Available: <https://platform.openai.com/tokenizer>
- [27] R. Xu and D. Wunsch, ‘Survey of clustering algorithms’, *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.
- [28] J. MacQueen, ‘Some methods for classification and analysis of multivariate observations’, 1967.
- [29] P. Drineas, A. M. Frieze, R. Kannan, S. Vempala, and V. Vinay, ‘Clustering in Large Graphs and Matrices.’, in *SODA*, Citeseer, 1999, pp. 291–299.
- [30] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. in Prentice Hall advanced reference series. Englewood Cliffs, N.J: Prentice Hall, 1988.
- [31] J. T. Vanderplas, *Python data science handbook: essential tools for working with data*, First edition. Sebastopol, CA: O’Reilly Media, Inc, 2016.
- [32] J. Mao and A. K. Jain, ‘A self-organizing network for hyperellipsoidal clustering (HEC)’, *Ieee transactions on neural networks*, vol. 7, no. 1, pp. 16–29, 1996.
- [33] C. M. Bishop, *Pattern recognition and machine learning*. in Information science and statistics. New York: Springer, 2006.
- [34] U. Von Luxburg, ‘A tutorial on spectral clustering’, *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
- [35] R. Jin, F. Kang, and C. H. Ding, ‘A probabilistic approach for optimizing spectral clustering’, in *Advances in neural information processing systems*, 2006, pp. 571–578.
- [36] F. Mosteller and J. W. Tukey, *Data analysis and regression. A second course in statistics*. 1977.
- [37] S. Arlot and A. Celisse, ‘A survey of cross-validation procedures for model selection’, *Statist. Surv.*, vol. 4, no. none, Jan. 2010, doi: 10.1214/09-SS054.
- [38] R. Kohavi, ‘A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection’, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, in IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143.
- [39] PROFECO, ‘Estudio de calidad: Cereales para niños’, 2011. Accessed: May 23, 2023. [Online]. Available: https://www.gob.mx/cms/uploads/attachment/file/119166/Estudio_Cereales_para_nin__os_30-44_Abril_2011.pdf

- [40] PROFECO, ‘Estudios de calidad de cereales’, 2019. Accessed: May 23, 2023. [Online]. Available: https://www.gob.mx/cms/uploads/attachment/file/426685/ESTUDIOS_DE_CALIDAD_CEREALES_.pdf
- [41] D. Urueta-Hinojosa, P. Lara-Velazquez, M. Gutierrez-Andrade, S. De-los-Cobos-Silva, E. Rincon-Garcia, and A. Mora-Gutierrez, ‘COMPARATIVE STUDY TO DETERMINE THE CLASSIFICATION AMONG THE MOST CONSUMED CREAMS IN MEXICO USING PATTERN RECOGNITION’, in *ANNALS OF NUTRITION AND METABOLISM*, KARGER ALLSCHWILERSTRASSE 10, CH-4009 BASEL, SWITZERLAND, 2020, pp. 134–134.
- [42] PROFECO, ‘Estudio de calidad: Cremas comestibles’, 2014. Accessed: May 23, 2023. [Online]. Available: https://www.gob.mx/cms/uploads/attachment/file/112275/Estudio_Cremas_Comestibles_58-68.pdf
- [43] PROFECO, ‘Cremas comestibles’, 2019. Accessed: May 23, 2023. [Online]. Available: https://issuu.com/profeco/docs/revista_del_consumidor_diciembre_2019
- [44] A. L. Waterhouse and S. E. Ebeler, Eds., *Chemistry of wine flavor*. in ACS symposium series, no. 714. Washington, DC : [New York]: American Chemical Society ; Distributed by Oxford University Press, 1998.
- [45] UNAM, ‘ExECUM3’, Estudio Comparativo de Universidades Mexicanas. Accessed: May 23, 2023. [Online]. Available: <https://www.execum.unam.mx/>
- [46] PROFECO, ‘Útiles escolares’. Accessed: Jan. 07, 2023. [Online]. Available: https://www.gob.mx/cms/uploads/attachment/file/752527/ESTUDIO_DE_CALIDAD_UTILIDADES_ESCOLARES.pdf
- [47] J. A. Suárez, D. E. Urueta, M. Esquivel, J. A. Salgado, and M. D. Román, ‘Estrategia de aprendizaje escuadra invertida: propuesta en la materia de comercio electrónico en la Universidad Politécnica del Estado de Guerrero.’, Guerrero, México, Dec. 08, 2023. Accessed: Jan. 06, 2024. [Online]. Available: <https://acortar.link/GqouGX>
- [48] B. V. Ramana and R. S. Kumar Boddu, ‘Performance Comparison of Classification Algorithms on Medical Datasets’, in *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, Jan. 2019, pp. 0140–0145. doi: 10.1109/CCWC.2019.8666497.
- [49] A. K. Tanwani, J. Afridi, M. Z. Shafiq, and M. Farooq, ‘Guidelines to select machine learning scheme for classification of biomedical datasets’, presented at the European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Springer, 2009, pp. 128–139.
- [50] D. E. U. Hinojosa *et al.*, ‘Classic colouring problems as special cases of the soft graph colouring model’, *IJTPM*, vol. 19, no. 2, p. 131, 2019, doi: 10.1504/IJTPM.2019.100604.

- [51] D. Urueta-Hinojosa, P. Lara-Velazquez, M. Gutierrez-Andrade, S. De-los-Cobos-Silva, E. Rincon-Garcia, and R. Mora-Gutierrez, 'PATTERN RECOGNITION TO DETERMINE THE QUALITY OF THE MOST CONSUMED BREAKFAST CEREALS IN MEXICO', in *ANNALS OF NUTRITION AND METABOLISM*, KARGER ALLSCHWILERSTRASSE 10, CH-4009 BASEL, SWITZERLAND, 2020, pp. 159–159.
- [52] W. Duch, 'Benchmark datasets used for classification: comparison of results'. Accessed: Oct. 05, 2020. [Online]. Available: <http://www.is.umk.pl/~ Duch/projects/projects/datasets.html#Ljubljana>
- [53] D. E. Urueta, P. Lara, M. Á. Gutiérrez, S. G. de-los-Cobos, E. A. Rincón, and R. A. Mora, 'A Comparative Ranking Model among Mexican Universities Using Pattern Recognition', *Mathematics*, vol. 9, no. 14, p. 1615, Jul. 2021, doi: 10.3390/math9141615.
- [54] e.Economía, 'Procuraduría Federal del Consumidor - Desconcentrados', Secretaría de Economía. Accessed: Mar. 22, 2023. [Online]. Available: <https://e.economia.gob.mx/procuraduria-federal-del-consumidor/>
- [55] AMEXCID, 'Catálogo de Capacidades Mexicanas de Cooperación Internacional para el Desarrollo'. Accessed: Mar. 22, 2023. [Online]. Available: <https://de.sre.gob.mx/capacidades/objetivos-de-desarrollo-sostenible/producci%C3%B3n-y-consumo-responsables/profeco/estudios-de-calidad>



Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

ACTA DE DISERTACIÓN PÚBLICA

No. 00024

Matrícula: 2181800351

Modelo avanzado de reconocimiento de patrones no supervisado y aplicaciones.

En la Ciudad de México, se presentaron a las 17:00 horas del día 31 del mes de octubre del año 2024 en la Unidad Iztapalapa de la Universidad Autónoma Metropolitana, los suscritos miembros del jurado:

DR. JOSE FRANCISCO GOMEZ AGUILAR
DRA. BIBIANA OBREGON QUINTANA
DRA. HERICA SANCHEZ LARIOS
DR. JORGE ENRIQUE LAVIN DELGADO
DR. PEDRO LARA VELAZQUEZ

Bajo la Presidencia del primero y con carácter de Secretario el último, se reunieron a la presentación de la Disertación Pública cuya denominación aparece al margen, para la obtención del grado de:

DOCTOR EN CIENCIAS (CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN)

DE: DANIEL EDAHI URUETA HINOJOSA

y de acuerdo con el artículo 78 fracción IV del Reglamento de Estudios Superiores de la Universidad Autónoma Metropolitana, los miembros del jurado resolvieron:

Aprobar

Acto continuo, el presidente del jurado comunicó al interesado el resultado de la evaluación y, en caso aprobatorio, le fue tomada la protesta.



DANIEL EDAHI URUETA HINOJOSA
ALUMNO

REVISÓ

MTRA. ROSALVA SERRANO DE LA PAZ
DIRECTORA DE SISTEMAS ESCOLARES

DIRECTOR DE LA DIVISIÓN DE CBI

DR. ROMAN LINARES ROMERO

PRESIDENTE

DR. JOSE FRANCISCO GOMEZ AGUILAR

VOCAL

DRA. BIBIANA OBREGON QUINTANA

VOCAL

DRA. HERICA SANCHEZ LARIOS

VOCAL

DR. JORGE ENRIQUE LAVIN DELGADO

SECRETARIO

DR. PEDRO LARA VELAZQUEZ