

**INSTITUTO TECNOLÓGICO DE TOLUCA**

**ANTEPROYECTO DE RESIDENCIA PROFESIONAL**

**“ANÁLISIS DE DATOS CON EL ALGORITMO DE AGRUPAMIENTO  
K-PROTOTYPES”**

**Ingeniería en Sistemas Computacionales**

**PRESENTA:**

**Fabiola Ramírez Guerrero**

**No. CONTROL:**

**C18280980**

**ASESOR INTERNO:**

**Andrea Olimpia Nájera Ávila**

**ASESOR EXTERNO:**

**Dra. Eréndira Rendón Lara**

**METEPEC, ESTADO DE MÉXICO, JUNIO DE 2024**

Agradecimientos

Deben ser concretos, indicando las razones por las cuales se menciona a cada persona o institución.

Resumen

Es una síntesis del proyecto final. Una representación abreviada del proyecto. Resaltando las ideas generales del proyecto y los logros alcanzados, destacando puntos de vista del autor que motiven al lector a leer el documento. No debe exceder de una cuartilla.

Índice

[1 Introducción 1](#_Toc174036587)

[2 Descripción de la empresa u organización y del puesto o área del trabajo del estudiante. 2](#_Toc174036588)

[3 Problema a resolver 2](#_Toc174036589)

[4 Objetivos 2](#_Toc174036590)

[4.1 General 2](#_Toc174036591)

[4.2 Específicos 2](#_Toc174036592)

[5 Justificación 2](#_Toc174036593)

[6 Marco Teórico (fundamentos teóricos) 2](#_Toc174036594)

[6.1 Minería de datos 2](#_Toc174036595)

[6.1.1 Disciplinas relacionadas con minería de datos 2](#_Toc174036596)

[6.2 Clustering 2](#_Toc174036597)

[6.2.1 Objetivos del Clustering 2](#_Toc174036598)

[6.2.2 Importancia del Clustering en el Análisis de Datos 2](#_Toc174036599)

[6.2.3 Aplicaciones del Clustering 2](#_Toc174036600)

[6.3 Algoritmos de agrupamiento (clustering) 2](#_Toc174036601)

[6.3.1 Clustering jerárquico 2](#_Toc174036602)

[6.3.2 Clustering particional 2](#_Toc174036603)

[6.3.3 Algoritmos de agrupamiento de recorridos simples. 2](#_Toc174036604)

[6.4 Similaridad Y Disimilaridad 2](#_Toc174036605)

[6.5 Algoritmos de Clustering 2](#_Toc174036606)

[6.5.1 El algoritmo DBSCAN 2](#_Toc174036607)

[6.5.2 El algoritmo BIRCH 2](#_Toc174036608)

[6.5.3 El Algoritmo K-Means 2](#_Toc174036609)

[6.5.4 El algoritmo k-modes 2](#_Toc174036610)

[6.5.5 Algoritmo k-Prototypes 2](#_Toc174036611)

[6.6 Diagrama de flujo K-Prototype 7](#_Toc174036612)

[6.7 Prueba de escritorio 8](#_Toc174036613)

[6.8 Investigaciones y proyectos usados con el algoritmo K-Prototypes 15](#_Toc174036614)

[6.8.1 Mejora de K-Prototypes mediante optimización evolutiva 15](#_Toc174036615)

[6.8.2 Aplicaciones en bioinformática 15](#_Toc174036616)

[6.8.3 Optimización del Transporte Público en Singapur 16](#_Toc174036617)

[6.8.4 Segmentación de Clientes en Comercio Electrónico en China 16](#_Toc174036618)

[6.8.5 Análisis de Datos Clínicos en Estados Unidos 16](#_Toc174036619)

[6.9 Herramientas utilizadas 17](#_Toc174036620)

[7 Procedimiento y descripción de las actividades realizadas. 19](#_Toc174036621)

[7.1 Requisitos funcionales 19](#_Toc174036622)

[7.2 Requisitos no funcionales 19](#_Toc174036623)

[7.3 Análisis de la base de datos 19](#_Toc174036624)

[8 Resultados 19](#_Toc174036625)

[9 Conclusiones de Proyecto 19](#_Toc174036626)

[10 Recomendaciones 20](#_Toc174036627)

[11 Competencias desarrolladas y/o aplicadas 20](#_Toc174036628)

[11.1 Competencias Específicas 20](#_Toc174036629)

[11.2 Competencias genéricas 21](#_Toc174036630)

[12 Fuentes de información 22](#_Toc174036631)

[13 Anexos 24](#_Toc174036632)

Índice de figuras

[Figura 1 Representación de un algoritmo de agrupamiento o clustering 2](#_Toc174027434)

[Figura 2 Agrupamiento jerárquico 2](#_Toc174027435)

[Figura 3 Clustering particional 2](#_Toc174027436)

Índice de tablas

[Tabla 1 Categoría de productos 8](#_Toc174892194)

[Tabla 2 Centros iniciales 9](#_Toc174892195)

[Tabla 3 Iteración 1 distancia objeto 1 9](#_Toc174892196)

[Tabla 4 Iteración 1 distancia objeto 2 9](#_Toc174892197)

[Tabla 5 Iteración 1 distancia objeto 3 10](#_Toc174892198)

[Tabla 6 Iteración 1 distancia objeto 4 10](#_Toc174892199)

[Tabla 7 Iteración 1 distancia objeto 5 11](#_Toc174892200)

[Tabla 8 Asignación de objetos 11](#_Toc174892201)

[Tabla 9 Reasignación de objetos, grupo 1 12](#_Toc174892202)

[Tabla 10 Reasignación de objetos, grupo 2 12](#_Toc174892203)

[Tabla 11 Iteración 2 distancia objeto 1 12](#_Toc174892204)

[Tabla 12 Iteración 2 distancia objeto 2 13](#_Toc174892205)

[Tabla 13 Iteración 2 distancia objeto 3 13](#_Toc174892206)

[Tabla 14 Iteración 2 distancia objeto 4 14](#_Toc174892207)

[Tabla 15 Iteración 2 distancia objeto 5 14](#_Toc174892208)

[Tabla 16 Asignación de objetos segunda iteración 15](#_Toc174892209)

[Tabla 17 Reasignación de objetos, grupo 1 15](#_Toc174892210)

[Tabla 18 Reasignación de objetos, grupo 2 15](#_Toc174892211)

[Tabla 19 Iteración 3 distancia objeto 1 16](#_Toc174892212)

[Tabla 20 Iteración 3 distancia objeto 2 16](#_Toc174892213)

[Tabla 21 Iteración 3 distancia objeto 3 17](#_Toc174892214)

[Tabla 22 Iteración 3 distancia objeto 4 17](#_Toc174892215)

[Tabla 23 Iteración 3 distancia objeto 5 17](#_Toc174892216)

[Tabla 24 Asignación de objetos tercera iteración 18](#_Toc174892217)

# Introducción

Es el contenido global de lo que va a encontrarse en el documento. Incluye los aspectos relevantes de los antecedentes, del problema, los objetivos, la justificación y la hipótesis (si aplica).

# Descripción de la empresa u organización y del puesto o área del trabajo del estudiante.

El Instituto Tecnológico de Toluca pertenece al sistema de la Secretaría de Educación Pública conocido como Tecnológico Nacional de México. Esta red se encarga de formar profesionales en la Ciencia y la Tecnológica, así como impulsarlos a través de la creación de proyectos y su divulgación.

El ITT se ha consolidado como una de las universidades de calidad en el Estado de México con respecto a esta área. Han tratado de cubrir las áreas industriales y de servicios en varias regiones del país, vinculándose a empresas, organizaciones e instituciones. Promueven entre sus catedráticos y estudiantes valores como la justicia, independencia, democracia y solidaridad. Por eso procuran el deporte y la cultura como estrategia para mentes y cuerpos sanos.

Realizan actividades donde vinculan a sus estudiantes a proyectos con comunidades, así como la movilidad a otras ciudades o países. Y también imparten capacitación continua para sus egresados y personas en general que quieran seguirse preparando.

Misión

Ofrecer servicios de Educación Superior Tecnológica con calidad, equidad y pertinencia, orientados a la formación integral del ser humano mediante la oferta de programas educativos acreditados y que promuevan el desarrollo sustentable para la conformación de una sociedad justa y humana.

Visión

Ser una Institución de Educación Superior Tecnológica basada en valores y reconocida por ofrecer un servicio de calidad, mediante el desarrollo sostenido, sustentable y equitativo de su entorno y a través del alto desempeño de su personal.

# Problema a resolver

Actualmente, las organizaciones se enfrentan a desafíos significativos al intentar analizar grandes conjuntos de datos que contienen tanto variables numéricas como categóricas. Estos datos mixtos son comunes en muchos campos, desde el comercio minorista hasta la investigación científica, y la falta de herramientas adecuadas para procesarlos y agruparlos eficazmente representa un obstáculo importante.

Los métodos tradicionales de análisis de datos, como los algoritmos de agrupamiento estándar, suelen ser efectivos solo con un tipo de variable, ya sea numérica o categórica, pero no ambos. Esta limitación impide la identificación de patrones y tendencias clave en los datos mixtos, lo que puede llevar a interpretaciones incompletas y decisiones menos efectivas. La incapacidad de manejar estos datos de manera eficiente resulta en una visión parcial de las relaciones importantes entre las variables numéricas y categóricas, afectando negativamente la toma de decisiones estratégicas.

Además, la integración y comprensión de datos mixtos es crucial para diversas aplicaciones, desde el análisis de comportamiento del cliente hasta la investigación científica y la gestión operativa. Sin herramientas adecuadas, las organizaciones tienen dificultades para integrar datos de diferentes fuentes y obtener una comprensión holística. Esta deficiencia limita la capacidad para desarrollar estrategias efectivas y mejorar procesos operativos.

Existe una necesidad urgente de herramientas que puedan manejar eficientemente datos mixtos, permitiendo la identificación de patrones y tendencias que no son evidentes cuando se analizan tipos de datos por separado. Un sistema que utilice algoritmos avanzados, como el K-Prototypes, puede agrupar datos mixtos en clústeres significativos, mejorando la precisión del análisis y facilitando la toma de decisiones basada en datos.

Sin una solución adecuada, las organizaciones continuarán enfrentando dificultades para extraer valor significativo de sus datos, lo que puede resultar en decisiones subóptimas y limitar la capacidad de innovación. En un entorno empresarial competitivo, la capacidad de analizar y actuar sobre datos de manera efectiva es un diferenciador clave. La falta de herramientas adecuadas puede poner a las organizaciones en desventaja frente a competidores que sí pueden aprovechar completamente sus datos.

# Objetivos

## General

Implementación del algoritmo de agrupamiento K-Prototypes para encontrar la descripción de patrones en un conjunto de datos con datos numéricos y categóricos.

## Específicos

1. Estudio y análisis de algoritmos de agrupamiento (K-Prototypes)
2. Análisis y comprensión de la base de datos que será utilizada
3. Análisis estadístico de la base de datos que se utilizará
4. Diseño del prototipo del Software
5. Implementación en el lenguaje Phyton
6. Realización de pruebas
7. Presentación de resultados

# Justificación

Existen pocos algoritmos de agrupamiento que trabajen con datos de tipo mezclado (Numérico y categóricos) por esta razón, contar con esta implementación del algoritmo de agrupamiento permitirá, realizar estudios de análisis de datos, para cualquier conjunto de datos de este tipo. Muchos conjuntos de datos contienen una combinación de variables numéricas y categóricas, lo que presenta un desafío considerable para los métodos de agrupamiento tradicionales. La falta de herramientas adecuadas para analizar estos datos mixtos puede llevar a interpretaciones incompletas y decisiones menos efectivas.

La implementación del algoritmo de agrupamiento K-Prototypes permitirá analizar conjuntos de datos mixtos de manera más precisa y eficiente. Este proyecto se enfoca en desarrollar un software que utilice dicho algoritmo para identificar patrones y tendencias en los datos, lo que facilitará una comprensión más profunda de la información disponible.

Al permitir un análisis más completo de estos datos, se podrá mejorar la gestión de recursos, el rendimiento del sistema, y adaptar sus estrategias a las necesidades y comportamientos de los usuarios. La integración del algoritmo K-Prototypes es una herramienta accesible que permita a los usuarios realizar análisis complejos sin necesidad de una formación avanzada en análisis de datos.

El desarrollo de este software no solo solucionará un problema técnico significativo, sino que también abrirá nuevas oportunidades para el análisis de datos. Al ofrecer una solución robusta para el análisis de datos mixtos, el proyecto contribuirá a la toma de decisiones más informadas y estratégicas, mejorando así la eficiencia y efectividad operativa.

# Marco Teórico (fundamentos teóricos)

## Minería de datos

Hoy en día, las empresas que gestionan grandes volúmenes de datos buscan formas de analizarlos para identificar oportunidades que les permitan monitorear las ventas y dirigir sus estrategias de marketing. También buscan detectar tendencias que puedan predecir situaciones futuras. En esencia, el objetivo es identificar modelos o patrones que den sentido a la vasta cantidad de datos disponibles. Esta tarea recae en la minería de datos (MD), que se enfoca en extraer información valiosa o conocimiento a través del análisis detallado de grandes conjuntos de datos, utilizando técnicas y metodologías (algoritmos) diseñados para la clasificación, regresión, agrupamiento y detección de anomalías, entre otras.

El proceso de minería de datos implica ajustar modelos o identificar patrones a partir de datos observados. Este ajuste es generalmente de naturaleza estadística, permitiendo cierto nivel de ruido o error dentro del modelo. Los algoritmos de minería de datos se encargan principalmente de tareas de predicción (para datos desconocidos) y descripción (de patrones). Los objetivos de la predicción y la descripción se logran mediante las tareas fundamentales de la minería de datos: clasificación, regresión, agrupamiento, resumir información, dependencia del modelo, y detección de cambios y desviaciones. La mayoría de los métodos de minería de datos están basados en conceptos de aprendizaje automático, reconocimiento de patrones y estadística. El amplio conjunto de algoritmos disponibles para resolver estos problemas puede resultar abrumador tanto para los analistas de datos experimentados como para los principiantes. Los desarrollos actuales en minería de datos se centran en encontrar métodos que clasifiquen eficientemente grandes bases de datos, con datos de diversas naturalezas o combinaciones de estos.

.

### Disciplinas relacionadas con minería de datos

Por definición, el KDD (Descubrimiento de Conocimiento en Bases de Datos) se considera un campo interdisciplinario que reúne a investigadores de diversas áreas. El término "minería de datos" ha sido comúnmente utilizado por estadísticos, analistas de datos y la comunidad de Sistemas de Información para la Gestión (MIS), mientras que "KDD" ha sido más empleado en inteligencia artificial y en investigaciones de aprendizaje automático.

Estadística

La estadística ofrece una valiosa asistencia en el análisis de datos, pero a menudo no es suficiente por sí sola y presenta ciertos inconvenientes, como la inadecuación para datos nominales, la dificultad en la interpretación de resultados y la necesidad de que el usuario decida cómo y dónde analizar los datos. No obstante, desempeña un papel crucial en algunos pasos del proceso de KDD, especialmente en la selección y muestreo de datos, en la minería de datos y en la evaluación del conocimiento obtenido.

Inteligencia Artificial y Aprendizaje Automático

Estas disciplinas aportan algoritmos y métodos esenciales para el aprendizaje automático, permitiendo que los sistemas identifiquen patrones y hagan predicciones a partir de los datos. Los enfoques de aprendizaje supervisado y no supervisado son fundamentales para desarrollar modelos predictivos y de agrupamiento.

Bases de Datos y Sistemas de Información

El almacenamiento y recuperación eficiente de datos son vitales en la minería de datos. Las bases de datos relacionales y NoSQL, junto con los sistemas de gestión de bases de datos (DBMS), facilitan la organización y el acceso a grandes conjuntos de datos, apoyando las operaciones de minería.

Matemáticas y Algoritmos

La teoría matemática sustenta muchas técnicas de minería de datos. Los algoritmos de optimización, la teoría de grafos y las matemáticas discretas son esenciales para desarrollar métodos eficientes y efectivos para el análisis de datos.

Visualización de Datos

La capacidad de visualizar datos es crucial para interpretar y comunicar los resultados de la minería de datos. Herramientas y técnicas de visualización ayudan a presentar patrones complejos de manera comprensible y a facilitar la toma de decisiones basadas en datos.

Ciencias de la Computación

Las ciencias de la computación proporcionan los fundamentos técnicos y teóricos necesarios para el desarrollo de software y herramientas de minería de datos. Esto incluye desde el diseño de algoritmos hasta la implementación de sistemas escalables y eficientes.

## Clustering

Clustering es una técnica de aprendizaje no supervisado que se utiliza para agrupar un conjunto de objetos en subgrupos o clusters, de manera que los objetos en un mismo cluster sean más similares entre sí que con los objetos de otros clusters. La similitud entre objetos se define a menudo mediante una medida de distancia, como la distancia Euclidiana, aunque puede variar según el tipo de datos y el algoritmo utilizado.

El Clustering es un enfoque o método popular usado para segmentar grandes conjuntos de datos heterogéneos, dentro de pequeños subconjuntos homogéneos que puedan ser fácilmente manejados, analizados y modelados por separado. Los métodos de agrupamiento reparten un conjunto de objetos dentro de grupos, de tal forma que los objetos en el mismo grupo son más similares mutuamente que objetos en diferentes grupos conforme a algún criterio definido. El Clustering ha sido estudiado en diversas áreas, cada una tiene sus métodos para organizar los datos, en las áreas de análisis de datos y estadística se requiere a los datos en forma de números (cuantitativos). En aprendizaje de máquina, en forma de conceptos y simbolos (cualitativos). En minería de datos frecuentemente se trata con todos los tipos de datos o con la mezcla de estos: cuantitativos (numéricos) y cualitativos (categóricos o simplificación de simbólicos).

### Objetivos del Clustering

Descubrimiento de estructuras: Identificar patrones ocultos en los datos sin etiquetas predefinidas.

Reducción de dimensionalidad: Simplificar grandes conjuntos de datos agrupándolos en clusters representativos.

Detección de anomalías: Identificar puntos de datos que no pertenecen a ningún cluster, lo que puede ser útil para detectar fraudes o fallos.

Segmentación de mercados: Dividir un mercado en segmentos distintos para identificar grupos de clientes con comportamientos similares.

### Importancia del Clustering en el Análisis de Datos

El clustering es una herramienta fundamental en el análisis de datos y tiene aplicaciones en diversas disciplinas. En biología, por ejemplo, se usa para clasificar especies similares basadas en características genéticas. En el comercio electrónico, ayuda a segmentar a los clientes en grupos basados en su comportamiento de compra, permitiendo la personalización de campañas de marketing. En seguridad informática, se emplea para detectar patrones anómalos que podrían indicar ataques cibernéticos.

### Aplicaciones del Clustering

Biología y Medicina: Agrupamiento de genes con funciones similares, clasificación de enfermedades.

Marketing y Ventas: Segmentación de clientes, análisis de preferencias de consumidores.

Redes Sociales: Detección de comunidades en redes sociales, análisis de opiniones y sentimientos.

Seguridad y Fraude: Detección de transacciones fraudulentas, identificación de patrones inusuales en datos de seguridad.

Procesamiento del Lenguaje Natural: Agrupamiento de documentos o términos, análisis de sentimientos.

El clustering no solo organiza los datos en grupos manejables, sino que también facilita el descubrimiento de relaciones complejas y patrones en los datos que no son fácilmente visibles. Esto permite una comprensión más profunda y una toma de decisiones más informada en una variedad de campos.

## Algoritmos de agrupamiento (clustering)

Los algoritmos de agrupamiento reciben como entrada al conjunto de datos o la misma base de datos y sin supervisión de ningún tipo y de forma autónoma, agrupan los datos dentro de k grupos o clusters. Clustering es útil en ciertas aplicaciones de biología, medicina, sociología y otras donde no se encuentran bien definidas las clases. La figura [1] muestra la representación de un algoritmo de agrupación, la entrada del algoritmo se representa por el conjunto de datos u objetos y dependiendo del algoritmo de agrupamiento se solicita o no el parámetro k; la salida, es una etiqueta para cada uno de los elementos del conjunto de datos, la cual indica que dicho dato pertenece a cierta clase, grupo o cluster.

Tabla

Descripción generada automáticamente con confianza baja

Figura 1 Representación de un algoritmo de agrupamiento o clustering

Los algoritmos de **agrupamiento conceptual** desarrollados en aprendizaje de máquina agrupan datos con valores categóricos y también producen descripciones conceptuales de los clusters. La descripción conceptual de los grupos es importante en minería de datos, porque ayuda a la interpretación de los resultados del agrupamiento.

Los algoritmos de **agrupamiento estadístico** cuentan con algoritmos basados en la búsqueda de objetos que tengan conceptos iguales o similares y su eficiencia depende de buenas estrategias de búsqueda. Para los problemas de minería de datos que frecuentemente involucran muchos conceptos y espacios muy grandes de objetos, los métodos basados en conceptos pueden ser de potencial desventaja.

### Clustering jerárquico

Organiza los datos en una secuencia anidada de grupos. Una de sus características es el impacto visual del dendograma que habilita al analista de datos a visualizar los objetos mezclados en grupos o divisiones sucesivas de niveles de proximidad, los métodos de agrupamiento generalmente requieren solo el uso de la matriz de proximidad, estas técnicas son populares en las ciencias biológicas, sociales y de comportamiento por que se necesita construir taxonomías. Los algoritmos de clustering jerárquico tienen como objetivo unir grupos (cada grupo de objetos es manejado como si fuera un sólo objeto) para formar uno nuevo, tendiendo hacia un nivel de generalización mayor (métodos aglomerativos) o bien separando grupos formando nuevos sub-agrupamientos de ellos (métodos divisivos), tendiendo hacia un nivel mayor de particularización. Los algoritmos de clustering jerárquico descomponen una base de datos, en varios niveles de clustering. La descomposición jerárquica es representada por un dendograma o árbol, que divide iterativamente las bases de datos en subconjuntos más pequeños hasta que cada subconjunto consta de un objeto. En tal jerarquía, cada nodo del árbol representa un cluster de las bases de datos.

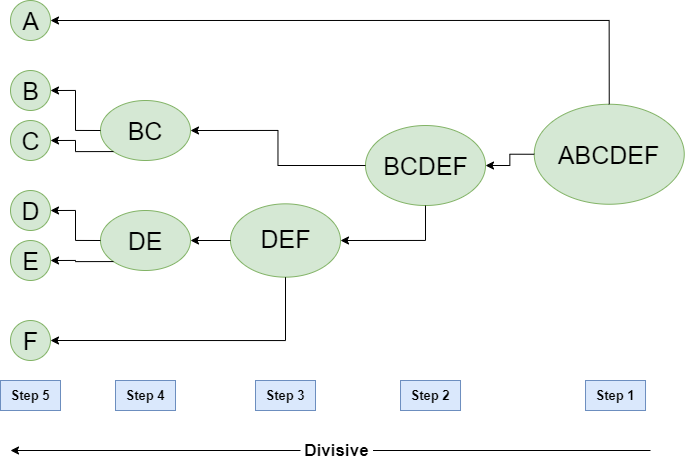


Figura 2 Agrupamiento jerárquico

### Clustering particional

Por su parte el clustering particional genera una simple partición de los datos, en un intento de recuperar los grupos naturales presentes en éstos, para ello cuentan con los datos en forma de matriz de patrones, donde las características han sido medidas en la escala ratio. Las técnicas de partición son usadas para aplicaciones de ingeniería y son apropiadas para una eficiente representación y comprensión de grandes bases de datos. El problema de agrupamiento puede formalizarse de la siguiente manera: Dados patrones en un espacio métrico de d-dimensiones, determinar una partición de los patrones en k cluster o grupos, tal que los patrones en un grupo sean más similares entre si que con otros patrones de diferentes grupos. El valor de k puede o no ser especificado.

Un criterio de agrupamiento como el error cuadrático puede adoptarse, el criterio puede clasificarse como global o local. El criterio global, representa cada grupo por un prototipo y asigna los patrones al grupo de acuerdo con el prototipo más similar. El criterio local, forma grupos utilizando una estructura local en los datos. Por ejemplo, los grupos pueden formarse por la identificación de las regiones de alta densidad en el espacio de patrones o por la asignación de un patrón y los i-ésimos k vecinos más cercanos del mismo grupo. Cada grupo es representado por su centro de gravedad y los objetos restantes son asignados al grupo cuyo objeto representativo sea más cercano.

Estos algoritmos inician con una partición inicial de las bases de datos y entonces usan una estrategia de control iterativa para mover los objetos de un grupo a otro, tratando de optimizar la calidad de los clusters durante el agrupamiento. Pero estos algoritmos presentan las siguientes limitaciones: Pueden aplicarse en problemas en los cuales existen objetos descritos por atributos numéricos exclusivamente, no manejan ausencia de información y muchas veces es necesario especificar el número de grupos que formarán el agrupamiento. Estos algoritmos de partición consideran el agrupamiento como un problema de optimización y pueden sufrir de mínimos locales.

Los algoritmos de partición construyen separan una base de datos en conjuntos de k grupos, donde k puede o no ser un parámetro de entrada. Cada grupo puede ser representado por un centro de gravedad k-means, o por uno de los objetos del grupo localizado cerca de su centro k-medoide, el resto de los objetos son asignados al grupo más cercano. La figura de todos los grupos encontrados por un algoritmo de partición es convexa como en la figura 2.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Figura 3 Clustering particional

### Algoritmos de agrupamiento de recorridos simples.

Estos algoritmos agrupan los objetos cercanos del conjunto de objetos utilizando como base una condición local de agrupamiento, recorriendo una sola vez todo el conjunto de objetos. Dentro de esta familia cabe destacar los algoritmos BIRCH y DBSCAN los cuales agrupan conjuntos de datos utilizando criterios locales.

## Similaridad Y Disimilaridad

Una vez que se identifican los tipos de características, el siguiente desafío consiste en determinar qué tan cerca o lejos se encuentran los objetos entre sí. Para abordar este problema, se emplean medidas de similitud y disimilitud. En términos generales, cuando dos objetos son muy similares entre ellos, el valor de la similitud es alto, mientras que la disimilitud es baja. La disimilitud a menudo se mide a través de una distancia, y existen diversas medidas de distancia que dependen del tipo de característica que se esté analizando. Por ejemplo, si las características son categóricas, no se pueden utilizar medidas de distancia que operen en un espacio métrico.

En la literatura, es común encontrar el término "métrica" para referirse a estas medidas de distancia. En este trabajo, nos referiremos a ellas simplemente como "medidas de distancia". Así, una distancia 𝑑(𝑥,𝑦) entre dos objetos 𝑥 y 𝑦 es una función que cumple con las siguientes propiedades propiedades

* 𝑑(𝑥,𝑦) no negativo
* 𝑑(𝑥,𝑦) identidad Si x=y
* 𝑑(𝑥,𝑦) 𝑑(𝑥,𝑦) simetría
* 𝑑(𝑥,𝑦) 𝑑(𝑥,z) + 𝑑(z,y) desigualdad triangular

## Algoritmos de Clustering

### El algoritmo DBSCAN

Este algoritmo trabaja sobre una noción basada en la densidad de los grupos y está diseñado para descubrir grupos de formas arbitrarias e irregulares en bases de datos espaciales con ruido, para ello utiliza estructuras de indices como árboles R\*-Tree y sus variantes. La idea es que cada rama del árbol no contenga más puntos de los preestablecidos en el parámetro "MinPts". Una de las limitaciones que presenta este algoritmo es el manejo de datos espaciales exclusivamente. El agrupamiento generado por el algoritmo puede variar si se cambia el orden de entrada de los objetos del conjunto de datos.

* Principio: Agrupa puntos que están densamente conectados.
* Ventajas: Detecta clusters de formas arbitrarias y maneja ruido (outliers).
* Limitaciones: Depende del parámetro de densidad y puede variar con el orden de los datos.

### El algoritmo BIRCH

Este algoritmo utiliza una estructura de árbol llamada CF-tree (cluster features tree), el cual almacena en sus nodos descripciones compactas de subgrupos (CF's), donde un CF se define como: número de objetos, suma lineal y suma de los cuadrados de todos los objetos en ese grupo. La metodologia que utiliza este algoritmo es la de generar un árbol de CF's con el conjunto de datos, para obtener las descripciones de los subgrupos y aplicar posteriormente un algoritmo de clasificación no supervisada para agrupar estos subgrupos. Asi, BIRCH transforma el problema de clasificar el conjunto de datos original en clasificar subgrupos de este conjunto de datos. BIRCH fue diseñado para agrupar grandes bases de datos con atributos numéricos, resumiendo la base de datos original hasta ajustarla al tamaño de la memoria principal disponible en el sistema en que se ejecuta, asi minimiza los costos de lectura y escritura de los datos. Las limitaciones que presenta este algoritmo son las siguientes: puede aplicarse en problemas donde se manejan datos numéricos exclusivamente, y no todas las funciones de distancia pueden aplicarse al utilizar este algoritmo, ya que la información manejada por los CF's permite manipular solamente una familia de medidas que puedan expresarse en términos del centroide, radio y diámetro, dependiendo del orden de entrada de los datos.

* Principio: Utiliza un árbol CF (Cluster Features) para agrupar datos.
* Ventajas: Eficiente en memoria y adecuado para grandes bases de datos numéricas.
* Limitaciones: Solo funciona con datos numéricos y tiene restricciones en las medidas de distancia.

### El Algoritmo K-Means

El algoritmo k-Means fue propuesto hace poco más de tres décadas y es uno de los algoritmos de agrupamiento más usados en una amplia variedad de áreas. k-Means recibe como parámetros el número de agrupamientos a formar y se encuentra definido sobre datos continuos, es decir, únicamente permite trabajar con objetos descritos por medio de un conjunto de atributos numéricos. Este algoritmo calcula iterativamente los centros de los agrupamientos mientras que al mismo tiempo minimiza una función objetivo. k-Means usa la distancia Euclidiana para comparar objetos y promedios para calcular los centros de los agrupamientos, lo que no le permite trabajar con atributos no numéricos.

1. Inicialización: Selecciona k centros de clúster iniciales aleatoriamente del conjunto de datos.
2. Asignación de Clústeres: Asigna cada punto de datos al centro de clúster más cercano usando la distancia euclidiana.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

1. Recalcular Centros: Calcula el nuevo centroide de cada clúster como el promedio de todos los puntos asignados al clúster.

Un dibujo de una persona

Descripción generada automáticamente con confianza media

1. Iterar: Repite los pasos de asignación y recalculación hasta que los centros no cambien significativamente entre iteraciones.
2. Convergencia: El algoritmo termina cuando los centros de clúster se estabilizan.

### El algoritmo k-modes

Permite extender k-means para ser utilizado con variables categóricas, a partir del cálculo de una medida de disimilitud que permita comparar observaciones categóricas, y la utilización de modas en lugar de medias para calcular los clusters. Por motivos de eficiencia computacional, el algoritmo k-modes sigue estos pasos para un número prestablecido de k clusters:

1. Selecciona k modas, una para cada cluster.
2. Asigna cada observación al cluster con la moda más cercana según la medida de disimilitud.
3. Actualiza la moda de cada cluster después de cada asignación.
4. Recalcula la medida de disimilitud después de que todas las observaciones han sido asignadas, y procede a reubicar aquellas más cercanas a otro cluster que al propio.
5. Recalcula las modas para los clusters afectados.
6. Repite los dos últimos pasos hasta que ninguna observación cambie de cluster después de un ciclo completo para todo el conjunto de datos.

Las principales características del algoritmo K-Modes son:

* Utiliza una medida de disimilitud para comparar objetos categóricos.
* Reemplaza el cálculo de promedios por el de modas.
* Usa un método basado en frecuencias para actualizar las modas de los grupos.

El algoritmo K-Modes sigue una estructura similar al K-Means, pero está adaptado para manejar datos categóricos, lo que lo hace adecuado para aplicaciones donde los datos no son numéricos. La integración de K-Means y K-Modes en el algoritmo K-Prototypes permite trabajar con conjuntos de datos mixtos, combinando la efectividad de ambos métodos.

* Principio: Extiende K-Means para datos categóricos usando una medida de disimilitud basada en frecuencias.
* Ventajas: Agrupa datos categóricos de manera eficiente.
* Limitaciones: Solo trabaja con datos categóricos.

### Algoritmo k-Prototypes

En 1997, Joshua Zhexue Huang presentó el algoritmo K-Prototypes, diseñado para el agrupamiento de datos mixtos.

Uno de los principales desafíos en la minería de datos es cómo particionar grandes conjuntos de datos de manera eficiente en grupos homogéneos, lo que constituye el objetivo central de la tarea descriptiva conocida como clustering o agrupamiento.

Existen métodos de agrupamiento que destacan por su capacidad para manejar grandes volúmenes de datos, pero su aplicación se limita a conjuntos de datos que contienen únicamente atributos numéricos. Sin embargo, muchos conjuntos de datos también incluyen atributos tanto numéricos como categóricos, lo que deja a los métodos tradicionales fuera de consideración para este tipo de datos.

El algoritmo K-Prototypes se basa en el modelo del K-Means, pero supera la limitación de solo trabajar con datos numéricos, ya que está diseñado específicamente para datos mixtos (numéricos y categóricos). Se introduce el concepto de centroide para representar el prototipo de un grupo con atributos mixtos; para los atributos categóricos se utiliza una medida de similitud, mientras que para los atributos numéricos se aplica una ecuación específica para calcular la distancia entre los objetos y los centroides de los grupos. Ambos componentes se combinan para determinar el centroide que representa el prototipo del grupo con atributos mixtos.

Este algoritmo está optimizado para trabajar con grandes conjuntos de datos mixtos, y su principal objetivo es reducir el costo computacional asociado con el agrupamiento de este tipo de datos.

El algoritmo k-Prototypes es un algoritmo de agrupamiento restringido que permite agrupar grandes conjuntos de datos mixtos. Constituye una integración de los algoritmos k-Modes y k-Means. Este algoritmo se basa en el paradigma del algoritmo k-means, preservando su eficiencia, pero eliminando la limitación de trabajar con únicamente datos de tipo numérico.

El algoritmo k-Modes fue la primera extensión del algoritmo k-Means orientada al agrupamiento de datos categóricos. Sigue la misma idea que el k-Means, pero con la principal diferencia en la medida de similitud utilizada para comparar objetos. Sus características principales incluyen el uso de una medida de disimilaridad para comparar objetos, el reemplazo del uso de promedios por el de modas, y un método basado en frecuencias para actualizar las modas. Este algoritmo fue diseñado exclusivamente para agrupar grandes conjuntos de datos categóricos.

El algoritmo k-Prototypes integra tanto al k-Means como al k-Modes para eliminar la limitación de trabajar únicamente con un solo tipo de datos. Asume que la medida de disimilaridad entre atributos numéricos se define por el cuadrado de la distancia Euclidiana y la medida de disimilaridad entre atributos categóricos se define por el número de incoincidencias de categorías entre objetos.

**Funciones de distancia para K-Prototype**

* **Distancia para atributos numéricos o distancia Euclidiana ():**

Sean x, y objetos con atributos numéricos (x1, x2 ,...,xn ) y (y1, y2,...,yn) respectivamente, la distancia Euclidiana se calcula con la la siguiente ecuación:

* **Distancia para atributos categóricos ():**

Para datos categóricos, sean b y c, objetos con atributos categóricos (b1, b2, ,...,bn ) y (c1 ,c2 ,..,cn ) respectivamente, la distancia se calcula con la ecuación:

Donde y es un peso para los atributos categóricos en el prototipo j, en este caso es igual a uno, y...

* **Distancia para atributos mezclados ():**

Sean b y c. dos objetos (b1, b2,...,bn) y (c1 ,c2 ,..,cn ), donde los atributos tienen valores tanto numéricos como categóricos. Para el algoritmo K-Prototypes la distancia que existe entre los objetos b y c, se calcula como la suma de la , más se y denota con la ecuación:

=

**Funcionamiento**

Integración de K-Means y K-Modes:

* K-Means: Utiliza la distancia euclidiana para medir la similitud entre objetos numéricos.
* K-Modes: Utiliza la disimilitud de coincidencia para datos categóricos, basándose en la frecuencia de las categorías.

**Parámetro 𝛾:**

* Controla el balance entre atributos numéricos y categóricos.
* Un valor bajo de 𝛾 da más peso a los atributos numéricos, mientras que un valor alto prioriza los categóricos.

​**Proceso de Clustering:**

Paso 1: Elegir k centros iniciales.

Paso 2: Se eligen k-prototipos aleatoriamente.

Paso 3: Para cada objeto de la base de datos se calcula la distancia (D...) existente entre el objeto y los k-prototipos.

Paso 4: El objeto es asignado al prototipo que presente la menor distancia. En caso de empate seleccione un prototipo al azar, en este caso se toma el primero. (El paso 3 y 4 se efectúa para cada uno de los objetos).

Paso 5: Cuando todos los objetos son asignados a un prototipo los k-prototipos son recalculados.

Paso 6: Los nuevos prototipos son asignados calculando un promedio entre los objetos asignados a ese prototipo

Paso 7: Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia entonces el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Paso 8: Se repiten los pasos del 5 al 8 si algún objeto cambió de prototipo en caso contrario terminar. Si después de un número determinado de interacciones el algoritmo no converge termine el algoritmo con las asignaciones actuales. (El paso 7 y 8 se efectúa para cada uno de los objetos)

Paso 9. Fin del algoritmo

Para la prueba de escritorio de este algoritmo el fin del algoritmo será cuando se logre la estabilización de grupos, es decir, cuando no existan movimientos de objetos entre grupos (movimientas=0

**Aplicaciones**

El K-Prototypes es útil en áreas donde los conjuntos de datos contienen tanto atributos numéricos como categóricos, como en estudios de mercado, análisis de clientes, y ciencias sociales. Permite identificar patrones y agrupamientos en datos heterogéneos, ofreciendo una visión más completa que los algoritmos que solo manejan un tipo de dato.

## Diagrama de flujo K-Prototype

Diagrama

Descripción generada automáticamente

Figura 4 Diagrama de flujo k-prototype

## Prueba de escritorio

Tenemos un conjunto de datos con información sobre clientes de una tienda en línea. Cada cliente tiene 4 atributos:

1. Edad (numérico)
2. Ingresos anuales (numérico)
3. Categoría de producto favorito (categórico: Electrónica, Ropa, Comida, Libros)
4. Estado del cliente: (nuevo, regular, fiel)

Tabla 1 Categoría de productos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente |
| 1 | 25 | 40,000 | Electrónica | nuevo |
| 2 | 34 | 50,000 | Ropa | regular |
| 3 | 22 | 30,000 | Comida | fiel |
| 4 | 45 | 70,000 | Libros | nuevo |
| 5 | 23 | 35,000 | Electrónica | fiel |

Aplicación del Algoritmo K-Prototypes

Para datos mixtos (numéricos y categóricos), la disimilitud total entre un cliente 𝑖 y un prototipo 𝑦 se calcula con:

=

1. Inicialización: Seleccionamos aleatoriamente k centros iniciales, 𝑘=2
2. Se eligen k-prototipos aleatoriamente, objetos 1 y 3

k=2 prototipos iniciales.

Tabla 2 Centros iniciales

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 1 | 25 | 40,000 | Electrónica | nuevo | K1 |
| 3 | 22 | 30,000 | Comida | fiel | K2 |

1. Para cada objeto de la bdd se calcula la distancia () existente entre el objeto y los k-prototipos

Tabla 3 Iteración 1 distancia objeto 1

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 1** con el prototipo **K1** | | Distancia del **objeto 1** con el prototipo **K2** | |
| Datos numéricos: =((25-25)^2-(40000-40000)^2)^(1/2)=0 | Datos categóricos: (Electrónica, Electrónica)=0 + (nuevo, nuevo)=0 Total=0 | Datos numéricos: =((22-25)^2-(40000-30000)^2)^(1/2)=10000.00045 | Datos categóricos: (Electrónica, comida)=1 + (nuevo, fiel)=1 Total=2 |
| Distancia total: 0 | | Distancia Total: 10002.00045 | |

1. Al objeto 1 se le asigna el valor de **K1**
2. Para cada objeto de la bdd se calcula la distancia () existente entre el objeto y los k-prototipos

Tabla 4 Iteración 1 distancia objeto 2

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 2** con el prototipo **K1** | | Distancia del **objeto 2** con el prototipo **K2** | |
| Datos numéricos: =((34-25)^2-(40000-40000)^2)^(1/2)= 10000.00405 | Datos categóricos: (ropa, electrónica)= 1 + (regular, nuevo)= 1  Total= 2 | Datos numéricos: =((34-22)^2-(50000-30000)^2)^(1/2)=20000.0036 | Datos categóricos: (ropa, comida)= 1 + (regular, fiel)=1 Total=2 |
| Distancia total: 10002.00405 | | Distancia Total: 20002.0036 | |

1. Al objeto 2 se le asigna el valor de **K1**
2. Para cada objeto de la bdd se calcula la distancia () existente entre el objeto y los k-prototipos

Tabla 5 Iteración 1 distancia objeto 3

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 3** con el prototipo **K1** | | Distancia del **objeto 3** con el prototipo **K2** | |
| Datos numéricos: =((22-25)^2-(30000-40000)^2)^(1/2)= 10000.00045 | Datos categóricos: (comida, electrónica)= 1+ (fiel, nuevo)= 1  Total=2 | Datos numéricos: =((22-22)^2-(30000-30000)^2)^(1/2)=0 | Datos categóricos: (comida, comida)=0 + (fiel, fiel)=0  Total=0 |
| Distancia total: 10002.00045 | | Distancia Total: 0 | |

1. Al objeto 3 se le asigna el valor de **K2**
2. Para cada objeto de la bdd se calcula la distancia () existente entre el objeto y los k-prototipos

Tabla 6 Iteración 1 distancia objeto 4

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 4** con el prototipo **K1** | | Distancia del **objeto 4** con el prototipo **K2** | |
| Datos numéricos: =((45-25)^2-(70000-40000)^2)^(1/2)= 30000.00667 | Datos categóricos: (libros, electrónica)= 1+ (nuevo, nuevo)= 0  Total=1 | Datos numéricos: =((45-22)^2-(70000-30000)^2)^(1/2)=40000.00661 | Datos categóricos: (libros, comida)=1 + (nuevo, fiel)=1  Total=2 |
| Distancia total: 30001.00667 | | Distancia Total: 40002.00661 | |

1. Al objeto 4 se le asigna el valor de **K1**
2. Para cada objeto de la bdd se calcula la distancia () existente entre el objeto y los k-prototipos

Tabla 7 Iteración 1 distancia objeto 5

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 5** con el prototipo **K1** | | Distancia del **objeto 5** con el prototipo **K2** | |
| Datos numéricos: =((23-25)^2-(35000-40000)^2)^(1/2)= 5000.0004 | Datos categóricos: (electrónica, electrónica)=0 + (fiel, nuevo)=1  Total=1 | Datos numéricos: =((23-22)^2-(35000-30000)^2)^(1/2)=5000.0001 | Datos categóricos: (electrónica, comida)=1 + (fiel, fiel)=0  Total=1 |
| Distancia total: 5001.0004 | | Distancia Total: 5001.0001 | |

1. Al objeto 5 se le asigna el valor de **K2**
2. Todos los objetos han sido asignados en un prototipo como se aprecia en la tabla 8

Tabla 8 Asignación de objetos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 1 | 25 | 40,000 | Electrónica | nuevo | K1 |
| 2 | 34 | 50,000 | Ropa | regular | K1 |
| 3 | 22 | 30,000 | Comida | fiel | K2 |
| 4 | 45 | 70,000 | Libros | nuevo | K1 |
| 5 | 23 | 35,000 | Electrónica | fiel | K2 |

1. Se asignan los nuevos prototipos, para cada prototipo se calcula el promedio de los atributos numéricos y a través de la medida de similaridad y la moda para los atributos categóricos como se muestra en las tablas 9 y 10

Tabla 9 Reasignación de objetos, grupo 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 1 | 25 | 40,000 | Electrónica | nuevo | K1 |
| 2 | 34 | 50,000 | Ropa | regular | K1 |
| 4 | 45 | 70,000 | Libros | nuevo | K1 |
| Promedio | 34.67 | 53333.33 | Ropa | nuevo | K1 |

Tabla 10 Reasignación de objetos, grupo 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 3 | 22 | 30,000 | Comida | fiel | K2 |
| 5 | 23 | 35,000 | Electrónica | fiel | K2 |
| Promedio | 22.5 | 32500 | Comida | fiel | K2 |

1. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 11 Iteración 2 distancia objeto 1

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 1** con el prototipo **K1** | | Distancia del **objeto 1** con el prototipo **K2** | |
| Datos numéricos: =((25-34.67)^2-(40000-53333.33)^2)^(1/2)= 13333.33351 | Datos categóricos: (electrónica, ropa)= 1+ (nuevo, nuevo)= 0  Total=1 | Datos numéricos: =((25-22.5)^2-(40000-32500)^2)^(1/2)=7500.000417 | Datos categóricos: (electrónica, comida)=1 + (nuevo, fiel)=1  Total=2 |
| Distancia total: 13334.33351 | | Distancia Total: 7502.000417 | |

1. Al objeto 1 se le asigna el valor de **K1**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 12 Iteración 2 distancia objeto 2

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 2** con el prototipo **K1** | | Distancia del **objeto 2** con el prototipo **K2** | |
| Datos numéricos: =((34-34.67)^2-(50000-53333.33)^2)^(1/2)= 3333.330067 | Datos categóricos: (ropa, ropa)= 0+ (regular, nuevo)= 1  Total=1 | Datos numéricos: =((34-22.5)^2-(50000-32500)^2)^(1/2)=17500.00378 | Datos categóricos: (ropa, comida)=1 + (regular, fiel)=1  Total=2 |
| Distancia total: 3334.330067 | | Distancia Total: 17502.00378 | |

1. Al objeto 2 se le asigna el valor de **K1**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 13 Iteración 2 distancia objeto 3

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 3** con el prototipo **K1** | | Distancia del **objeto 3** con el prototipo **K2** | |
| Datos numéricos: =((22-34.67)^2-(53333.33-30000)^2)^(1/2)= 23333.33344 | Datos categóricos: (comidas, ropa)= 1+ (fiel, nuevo)= 2  Total=2 | Datos numéricos: =((22-22.5)^2-(30000-32500)^2)^(1/2)=2500.00005 | Datos categóricos: (comida, comida)=0 + (nuevo, fiel)=1  Total=1 |
| Distancia total: 23335.33344 | | Distancia Total: 2501.00005 | |

1. Al objeto 3 se le asigna el valor de **K2**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 14 Iteración 2 distancia objeto 4

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 4** con el prototipo **K1** | | Distancia del **objeto 4** con el prototipo **K2** | |
| Datos numéricos: =((45-34.67)^2-(70000-53333.33)^2)^(1/2)= 16666.6732 | Datos categóricos: (libros, ropa)= 1+ (nuevo, nuevo)= 0  Total=1 | Datos numéricos: =((45-22.5)^2-(70000-32500)^2)^(1/2)=37500.00675 | Datos categóricos: (libros, comida)=1 + (nuevo, fiel)=1  Total=2 |
| Distancia total: 16667.6732 | | Distancia Total: 37502.00675 | |

1. Al objeto 4 se le asigna el valor de **K1**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 15 Iteración 2 distancia objeto 5

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 5** con el prototipo **K1** | | Distancia del **objeto 5** con el prototipo **K2** | |
| Datos numéricos: =((23-34.67)^2-(35000-32500)^2)^(1/2)= 18333.33371 | Datos categóricos: (electrónica, ropa)= 1+ (fiel, nuevo)= 1  Total=2 | Datos numéricos: =((23-22.5)^2-(35000-32500)^2)^(1/2)=2500.00005 | Datos categóricos: (electrónica, comida)=1 + (fiel, fiel)=0  Total=1 |
| Distancia total: 18335.33371 | | Distancia Total: 2501.00005 | |

1. Al objeto 5 se le asigna el valor de **K2**
2. Todos los objetos han sido asignados en un prototipo como se aprecia en la tabla 16

Tabla 16 Asignación de objetos segunda iteración

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 1 | 25 | 40,000 | Electrónica | nuevo | K2 |
| 2 | 34 | 50,000 | Ropa | regular | K1 |
| 3 | 22 | 30,000 | Comida | fiel | K2 |
| 4 | 45 | 70,000 | Libros | nuevo | K1 |
| 5 | 23 | 35,000 | Electrónica | fiel | K2 |

1. Se asignan los nuevos prototipos, para cada prototipo se calcula el promedio de los atributos numéricos y a través de la medida de similaridad y la moda para los atributos categóricos como se muestra en las tablas 9 y 10

Tabla 17 Reasignación de objetos, grupo 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 1 | 25 | 40,000 | Electrónica | nuevo | K2 |
| 3 | 22 | 30,000 | Comida | fiel | K2 |
| 5 | 23 | 35,000 | Electrónica | fiel | K2 |
| Promedio | 23.33 | 35000 | Electrónica | Fiel | K2 |

Tabla 18 Reasignación de objetos, grupo 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 2 | 34 | 50,000 | Ropa | regular | K1 |
| 4 | 45 | 70,000 | Libros | nuevo | K1 |
| Promedio | 39.5 | 60000 | Ropa | REGULAR | K1 |

1. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 19 Iteración 3 distancia objeto 1

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 1** con el prototipo **K1** | | Distancia del **objeto 1** con el prototipo **K2** | |
| Datos numéricos: =((25-39.5)^2-(40000-60000)^2)^(1/2)= 20000.00526 | Datos categóricos: (electrónica, ropa)= 1+ (nuevo, regular)= 1  Total=2 | Datos numéricos: =((25-23.33)^2-(40000-35000)^2)^(1/2)=5000.000279 | Datos categóricos: (electrónica, electrónica)=0+ (nuevo, fiel)=1  Total=1 |
| Distancia total: 20002.00526 | | Distancia Total: 5001.000279 | |

1. Al objeto 1 se le asigna el valor de **K2**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 20 Iteración 3 distancia objeto 2

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 2** con el prototipo **K1** | | Distancia del **objeto 2** con el prototipo **K2** | |
| Datos numéricos: =((34-39.5)^2-(50000-60000)^2)^(1/2)= 10000.00151 | Datos categóricos: (ropa, ropa)= 0+ (regular, regular)= 0  Total=0 | Datos numéricos: =((34-23.33)^2-(50000-32500)^2)^(1/2)= 15000.00379 | Datos categóricos: (ropa, electrónica)=1 + (regular, fiel)=1  Total=2 |
| Distancia total: 10000.00151 | | Distancia Total: 15002.00379 | |

1. Al objeto 2 se le asigna el valor de **K1**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 21 Iteración 3 distancia objeto 3

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 3** con el prototipo **K1** | | Distancia del **objeto 3** con el prototipo **K2** | |
| Datos numéricos: =((22-39.5)^2-(30000-60000)^2)^(1/2)= 30000.0051 | Datos categóricos: (comidas, ropa)= 1+ (fiel, regular)= 1  Total=2 | Datos numéricos: =((22-23.33)^2-(30000-32500)^2)^(1/2)= 5000.000177 | Datos categóricos: (comida, electrónica)=1 + (fiel, fiel)=0  Total=1 |
| Distancia total: 30002.0051 | | Distancia Total: 5001.000177 | |

1. Al objeto 3 se le asigna el valor de **K2**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 22 Iteración 3 distancia objeto 4

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 4** con el prototipo **K1** | | Distancia del **objeto 4** con el prototipo **K2** | |
| Datos numéricos: =((45-39,5)^2-(70000-60000)^2)^(1/2)= 10000.00151 | Datos categóricos: (libros, ropa)= 1+ (nuevo, regular)= 1  Total=2 | Datos numéricos: =((45-23.3)^2-(70000-35000)^2)^(1/2)= 35000.00671 | Datos categóricos: (libros, electrónica)=1 + (nuevo, fiel)=1  Total=2 |
| Distancia total: 10002.00151 | | Distancia Total: 35002.00671 | |

1. Al objeto 4 se le asigna el valor de **K1**
2. Se recalcula la distancia entre los objetos y los nuevos prototipos, si el objeto presenta una menor distancia el objeto se mueve del prototipo asignado originalmente al prototipo que presente la menor distancia

Tabla 23 Iteración 3 distancia objeto 5

|  |  |  |  |
| --- | --- | --- | --- |
| Distancia del **objeto 5** con el prototipo **K1** | | Distancia del **objeto 5** con el prototipo **K2** | |
| Datos numéricos: =((23-39.5)^2-(35000-60000)^2)^(1/2)= 25000.00544 | Datos categóricos: (electrónica, ropa)= 1+ (fiel, regular)= 1  Total=2 | Datos numéricos: =((23-23.3)^2-(35000-32500)^2)^(1/2)= 0.33 | Datos categóricos: (electrónica, electrónica) =0 + (fiel, fiel)=0  Total=0 |
| Distancia total: 25002.00544 | | Distancia Total: 0.33 | |

1. Al objeto 5 se le asigna el valor de **K2**
2. Todos los objetos han sido asignados en un prototipo como se aprecia en la tabla 16

Tabla 24 Asignación de objetos tercera iteración

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cliente | Edad | Ingresos Anuales ($) | Categoría de Producto Favorito | Estado del cliente | Prototipo |
| 1 | 25 | 40,000 | Electrónica | nuevo | K2 |
| 2 | 34 | 50,000 | Ropa | regular | K1 |
| 3 | 22 | 30,000 | Comida | fiel | K2 |
| 4 | 45 | 70,000 | Libros | nuevo | K1 |
| 5 | 23 | 35,000 | Electrónica | fiel | K2 |

El algoritmo realiza el mismo procedimiento hasta que converja, en este caso converge debido a que ya no hubo movimientos, es decir, que movimiento=0.

## Investigaciones y proyectos usados con el algoritmo K-Prototypes

### Mejora de K-Prototypes mediante optimización evolutiva

Investigadores han propuesto mejoras al algoritmo K-Prototypes usando técnicas de optimización evolutiva como los algoritmos genéticos. Estos métodos buscan optimizar la selección de prototipos iniciales y el parámetro 𝛾 γ, mejorando la precisión y estabilidad del clustering.

### Aplicaciones en bioinformática

En el campo de la bioinformática, el K-Prototypes se ha utilizado para agrupar datos genéticos y clínicos, donde los atributos incluyen información tanto numérica (como la expresión génica) como categórica (como el tipo de enfermedad).

### Optimización del Transporte Público en Singapur

Contexto: En un estudio realizado por la Universidad Nacional de Singapur, K-Prototypes se utilizó para analizar datos de transporte público, incluyendo tiempos de viaje (numéricos) y tipos de usuarios (categóricos).

Resultados: La investigación ayudó a mejorar la planificación de rutas y la distribución de recursos en el transporte público.

### Segmentación de Clientes en Comercio Electrónico en China

Contexto: Una gran empresa de comercio electrónico en China utilizó K-Prototypes para segmentar a sus clientes basándose en el historial de compras y preferencias de productos.

Resultados: La segmentación permitió personalizar las ofertas y mejorar la experiencia del cliente, resultando en un aumento significativo en la satisfacción y las tasas de conversión de ventas.

### Análisis de Datos Clínicos en Estados Unidos

Contexto: Un hospital en Estados Unidos aplicó K-Prototypes para analizar datos clínicos de pacientes, incluyendo datos numéricos (edad, resultados de pruebas) y categóricos (diagnósticos, tratamientos).

Resultados: Los clusters formados ayudaron a identificar patrones en los datos clínicos, lo cual mejoró la personalización de tratamientos y estrategias de atención.

## Herramientas

**Visual Studio Code (VS Code)**

Visual Studio Code es un editor de código desarrollado por Microsoft que, a pesar de su ligereza, ofrece una gran potencia. Incluye soporte para la depuración, control de versiones Git, resaltado de sintaxis, IntelliSense (autocompletado inteligente), snippets, y refactorización de código. Su diseño extensible permite añadir funcionalidades a través de extensiones, haciéndolo altamente adaptable a diferentes lenguajes y necesidades de desarrollo. Una de sus características más útiles es la terminal integrada, que facilita la ejecución de comandos durante el desarrollo de software. Visual Studio Code no está restringido en cuanto al tipo de aplicaciones que se pueden desarrollar con él, ya que se adapta de manera flexible a las necesidades de cada proyecto. Estas características lo convierten en una opción preferida para entornos que requieren la integración de lenguajes y herramientas para el análisis de datos multivariantes.

**Python**

Python es un lenguaje de programación interpretado, de alto nivel y con un propósito general, que destaca por su diseño enfocado en la claridad y legibilidad del código, utilizando una sintaxis sencilla y significativa. Su popularidad se debe en gran parte a su facilidad de aprendizaje y a la vasta comunidad de desarrolladores que ofrece numerosos recursos y bibliotecas para diversas áreas, como el análisis de datos, el desarrollo web, la inteligencia artificial, entre otros.

Python fue creado por Guido Van Rossum, un programador originario de los Países Bajos. El desarrollo de Python comenzó en 1989 en el Centrum Wiskunde & Informatica (CWI) como un proyecto personal que Van Rossum emprendió durante las vacaciones navideñas. El nombre "Python" proviene del programa de televisión "Monty Python's Flying Circus", un show de la BBC del que Van Rossum era un gran admirador.

**Evolución de Python**

La primera versión de Python, la 0.9.0, fue lanzada por Guido Van Rossum en 1991 e incluía funcionalidades importantes como varios tipos de datos y mecanismos para la gestión de errores. En 1994, Python 1.0 se presentó con mejoras significativas en la manipulación de listas, facilitando operaciones como la asignación, el filtrado y la reducción. El 16 de octubre de 2000, Python 2.0 salió al público, introduciendo características útiles como el soporte para Unicode y una forma más compacta de iterar sobre listas. Python 3.0, lanzado el 3 de diciembre de 2008, trajo consigo mejoras como una nueva función de impresión y un soporte mejorado para la división de números y la gestión de errores.

**Librerías de Python**

Para el sistema se necesitará varias bibliotecas de Python que ayudarán en diferentes aspectos del desarrollo.

**Math**

Es una biblioteca estándar de Python que proporciona funciones matemáticas básicas. Esta librería incluye una variedad de funciones para realizar cálculos matemáticos comunes, como operaciones trigonométricas, exponenciales, logarítmicas, y más. También contiene constantes matemáticas como π (math.pi) y e (math.e).

**Random**

Biblioteca estándar de Python. Se utiliza para generar números aleatorios y realizar operaciones relacionadas con la aleatoriedad, como seleccionar elementos al azar de una lista, barajar secuencias, y más.

# Procedimiento y descripción de las actividades realizadas.

Se describen detalladamente las tareas generadas de cada una de las actividades definidas previamente en el cronograma. Incluye figuras, tablas, capturas de pantallas y todos aquellos elementos que den soporte a la realización del proyecto. Respetando la metodología definida para el desarrollo.

## Requisitos funcionales

Los Requisitos Funcionales describen lo que el sistema debe hacer, es decir, las funciones y características específicas que debe tener para cumplir con las necesidades del usuario.

## Requisitos no funcionales

Los Requisitos No Funcionales describen cómo el sistema debe comportarse y sus cualidades, en lugar de describir funciones específicas. Estos requisitos se centran en la calidad del sistema.

## Análisis de la base de datos

# Resultados

Considera manuales de usuario, reportes del sistema, manual técnico, estadísticas.

# Conclusiones de Proyecto

Las Conclusiones deben iniciar con los objetivos y los alcances planteados en el anteproyecto para conocer lo que se logró. Se desprenden directamente de los Resultados. Describen las limitaciones encontradas durante el desarrollo del proyecto.

# Recomendaciones

Incluye todos los aspectos de mejora, alcances futuros para seguir desarrollando sobre lo que se ha detectado como necesario. Las acciones que se deben hacer con los resultados.

# Competencias desarrolladas y/o aplicadas

Son los conocimientos y habilidades adquiridas y aplicadas durante el desarrollo del proyecto (Trabajo en equipo, comunicación, aplicación de una metodología y/o plataformas).

**COMPETENCIAS DESARROLLADAS Y/O APLICADAS**

En esta sección se deberán mencionar, tanto las competencias genéricas como las específicas que el estudiante ha desarrollado durante la realización de la residencia profesional, así como las competencias adquiridas durante su estancia en el Instituto Tecnológico de Toluca que le permitieron llevar a cabo su proyecto de residencia profesional.

## Competencias Específicas

Son aquellas que en su desarrollo definen, una cualificación profesional concreta, al sujeto en formación; es decir: saberes, quehaceres y manejo de tecnologías propias de un campo profesional específico.

El dominio de estas competencias específicas aportaría, a los sujetos en formación, los conocimientos teóricos y las aplicaciones tecnológicas, propios de cada profesión ya que están vinculados a lo que se denomina “el saber hacer profesional”. (Tobón, 2008)

Manejo de Conocimientos relativos a la ciencia, la tecnología y las humanidades en un campo profesional específico.

Conocimiento de las prácticas profesionales que se desarrollan y la evolución y trascendencia de la profesión.

Utilización con precisión terminología, simbología e instrumentos; así como el uso de lenguajes, algoritmos, heurísticos.

Creación y diseño de aparatos, máquinas y de objetos tecnológicos en general.

Planificación, establecimiento, organización y desarrollo de procesos de distinta índole.

Selección y uso de materiales, máquinas y herramientas.

Montaje de aparatos e instalaciones; así como, dar solución a problemas productivos y tecnológicos.

Actuación profesional, trabajar en equipo, apreciar el conocimiento y los hábitos del trabajo, cultivar la precisión y la curiosidad; así como, el entusiasmo y el interés en su quehacer profesional.

## Competencias genéricas

Son aquellas que se pueden aplicar en un amplio campo de ocupaciones, condiciones y situaciones profesionales dado que aportan las herramientas intelectuales y procedimentales básicas que necesitan los sujetos para analizar los problemas, evaluar las estrategias, aplicar conocimientos a casos distintos y aportar soluciones adecuadas. (Tobón 2008)

|  |  |  |
| --- | --- | --- |
| **Instrumentales** | **Interpersonales** | **Sistémicas** |
| Capacidad de análisis y síntesis.  Capacidad de organizar y planificar.  Conocimientos generales básicos.  Conocimientos básicos de la carrera.  Comunicación oral y escrita en su propia lengua.  Conocimiento de una segunda lengua.  Habilidades básicas de manejo de la computadora.  Habilidades de gestión de información (habilidad para buscar y analizar información proveniente de fuentes diversas.  Solución de problemas.  Toma de decisiones. | Capacidad crítica y autocrítica.  Trabajo en equipo.  Habilidades interpersonales.  Capacidad de trabajar en equipo interdisciplinario  Capacidad de comunicarse con profesionales de otras áreas.  Apreciación de la diversidad y multiculturalidad.  Habilidad para trabajar en un ambiente laboral.  Compromiso ético. | Capacidad de aplicar los conocimientos en la práctica.  Habilidades de investigación.  Capacidad de aprender.  Capacidad de adaptarse a nuevas situaciones.  Capacidad de generar nuevas ideas (creatividad).  Liderazgo.  Conocimiento de culturas y  costumbres de otros países.  Habilidad para trabajar en forma autónoma.  Capacidad para diseñar y gestionar proyectos.  Iniciativa y espíritu emprendedor.  Preocupación por la calidad.  Búsqueda del logro. |

# Fuentes de información

López, E. S. (2006). Algoritmos de Agrupamiento en la Minería de Datos. Instituto Nacional de Astrofísica, Óptica y Electrónica. Recuperado de <https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/628/1/LopezES.pdf>

Hernández, E. (2006). Tesis de Algoritmos de Clustering. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional. Recuperado de <https://www.cs.cinvestav.mx/TesisGraduados/2006/tesisEdnaHernandez.pdf>

Amazon Web Services. (n.d.). What is Python? Recuperado de <https://aws.amazon.com/es/what-is/python/>

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645-678.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. John Wiley & Sons.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. Recuperado de <https://numpy.org/>

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56. Recuperado de <https://pandas.pydata.org/>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Recuperado de <https://scikit-learn.org/>

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. Recuperado de <https://matplotlib.org/>

Waskom, M. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. Recuperado de <https://seaborn.pydata.org/>

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261-272. Recuperado de <https://scipy.org/>

Ali, M., Jankowski, A., & Roy, A. (2020). PyCaret: An open source, low-code machine learning library in Python. arXiv preprint arXiv:2005.13547. Recuperado de <https://www.pycaret.org/>

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 2(3), 283-304.

Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values. Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, 21-34.

Siers, M. J., & Soria, E. M. (2017). Optimizing the K-Prototypes Clustering Algorithm Using Genetic Algorithms. Applied Soft Computing, 55, 196-207.

Wang, H., & Luan, Y. (2015). Clustering Mixed-Type Data Using K-Prototypes Algorithm. Journal of Bioinformatics and Computational Biology, 13(1), 1550004.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media.

Zaki, M. J., & Meira Jr, W. (2014). Data mining and analysis: Fundamental concepts and algorithms. Cambridge University Press.

# Anexos

Carta de autorización por parte de la empresa u organización para la titulación y otros si son necesarios.

**FORMATO GENERAL**

1. Tipo de fuentes: Arial
2. Tamaño: 12 texto, 13 subtítulos y 14 títulos.
3. Alineación: Justificada (Referencias alineadas a la izquierda)
4. Márgenes: superior 2.5 cm. , inferior 2.5 cm., izq. 3.0 cm y der. 2.5 cm.
5. Paginación: abajo a la derecha a partir de la Introducción
6. Interlineado: 1.5
7. Las citas textuales si tienen cinco o más renglones, utilizar sangría
8. Títulos, subtítulos y encabezados no llevan puntuación
9. Utilizar esquema de numerado decimal, máximo tres niveles. Ejem.

**3 Marco Teórico**

**3.1 Bases de datos**

**3.1.1 Oracle**

1. Los Anexos se ordenan con letras mayúsculas de la A a la Z.
2. Encabezados: Los títulos y subtítulos en negritas.
3. Los títulos inician en una nueva hoja