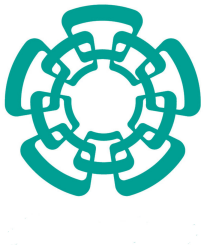


Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Unidad Cinvestav Tamaulipas

Método para la transformación de datos mixtos en tareas de aprendizaje automático

Tesis que presenta:

Gerardo Saúl Gausin Valle

Para obtener el grado de:

**Maestro en Ciencias
en Ingeniería y Tecnologías
Computacionales**

Dr. Iván López Arévalo, Co-Director
Dr. Edwyn Javier Aldana Bobadilla, Co-Director

Cd. Victoria, Tamaulipas, México. Diciembre, 2019

La tesis presentada por Gerardo Saúl Gausin Valle fue aprobada por:

_____

Dr. Hiram Galeana Zapién

Dr. José Luis González Compeán

Dr. Iván López Arévalo, Co-Director

Dr. Edwyn Javier Aldana Bobadilla, Co-Director

Cd. Victoria, Tamaulipas, México., 18 de Diciembre de 2019

The further one goes, the less one knows.
– Lao Tzu.

Agradecimientos

- Este trabajo se lo dedico a mis padres, hermanos y amigos que me han motivado a seguir adelante durante mis estudios. En especial, a mi amiga Elisa quien me ha aconsejado y apoyado en el transcurso de mi maestría.
- Agradezco la guía, paciencia y consejos que me proporcionaron mis asesores, Dr. Edwin Aldana Bodadilla y Dr. Iván López Árevalo.
- Agradecimientos al personal administrativo del CINVESTAV-Tamaulipas que durante mi estadía me apoyaron con trámites e instalaciones adecuadas. Adicionalmente, quiero agradecer al coordinador académico, Dr. Miguel Morales Sandoval por el apoyo que se me brindó para realizar un curso en la India.
- Agradezco a CONACyT por el apoyo económico que se me concedió, ya que me permitieron concentrarme exclusivamente en mis estudios. Asimismo, al CINVESTAV-Tamaulipas por la oportunidad que se me dió al estudiar en está institución.
- Agradezco también a los revisores de mi tesis, el Dr. Hiram Galeana Zapién y Dr. José Luis González Compeán.

Índice General

Índice de Figuras

Índice de Tablas

Índice de Algoritmos

Resumen

Método para la transformación de datos mixtos en tareas de aprendizaje automático

por

Gerardo Saúl Gausin Valle
Unidad Cinvestav Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2019
Dr. Edwyn Javier Aldana Bobadilla, Co-Director
Dr. Iván López Arévalo, Co-Director

Los métodos de aprendizaje automático más comunes resuelven problemas de clasificación y agrupamiento basados en conjuntos de datos en los que las características o propiedades del problema pertenecen a un espacio numérico. Sin embargo, muchos problemas a menudo incluyen datos donde coexisten características numéricas, nominales y ordinales. Dado que los datos nominales representan etiquetas codificadas por valores discretos sin un ordenamiento significativo, varias tareas de preprocesamiento son obligatorias. Si se ignoran estas tareas, podemos obtener efectos no deseados cuando se realiza métodos de aprendizaje automático. Los métodos como la codificación one-hot encoding son comúnmente usados para tratar este problema. El principal inconveniente de estos enfoque es la "maldición de la dimensionalidad" debido a una gran cantidad de atributos adicionales que se pueden introducir (dependiendo de los valores nominales distintos). En este trabajo proponemos un enfoque que codifica cada instancia en el conjunto de datos (con atributos numéricos y categóricos) como un código de caracteres numéricos. Este código induce un espacio discreto en el que se pueden realizar agrupación o clasificación, mejorando la efectividad lograda mediante la codificación tradicional. Esto se logra sin aumentar el número de atributos o características de las instancias en el conjunto de datos.

Abstract

Handling Mixed Type Data on Machine Learning Tasks

by

Gerardo Saúl Gausin Valle
Cinvestav Tamaulipas
Center for Research and Advanced Studies of the National Polytechnic Institute, 2019
Dr. Edwyn Javier Aldana Bobadilla, Co-advisor
Dr. Iván López Arévalo, Co-advisor

The most common machine learning methods solve classification and clustering problems based on datasets where the features of the problem belong to the numerical space. However, many problems often include data where numerical, nominal and ordinal features coexist. Since nominal data represent labels encoded by discrete values with no meaningful ordering, several pre-processing tasks are required. If these tasks are ignored, we can obtain biased effects when machine learning methods are applied. Methods as one-hot encoding are commonly applied to deal with this problem. The main drawback of these approaches is the "curse of dimensionality" due to the large number of additional attributes that can be introduced (depending on the distinct nominal values). We propose a novel approach that encodes each instance in the dataset (with numerical and categorical attributes) as a code of decimal characters that induces a discrete space. With this code clustering or classification can be performed improving the effectiveness achieved via traditional encoding without increasing the number of attributes of the instances in the dataset.

Nomenclatura

m	Número de atributos.
n	Número de instancias.
$\mathbb{S}$	Espacio de un conjunto de datos mixto.
$\mathbb{X}$	Conjunto de datos mixto.
$\overline{v}$	vector $(a_1, a_2, \ldots, a_m)$ en $\mathbb{S}^m$.
h	Número de cuantiles.
q_k	Cuantil con un intervalo de $[\underline{a_k}, \overline{a_k}]$ de tamaño δ.
$\underline{a_k}$	Límite inferior de intervalo de q_k.
$\overline{a_k}$	Límite superior de intervalo de q_k.
δ	Ancho del intervalo q_i.
$\mathbb{C}$	Conjunto de códigos.
RI	*Rand Index*.
ARI	*Adjusted Rand Index*.

1

Introducción

Mediante algoritmos de aprendizaje automático es posible programar una computadora para realizar tareas con conocimiento previo, representado en la forma de un conjunto de datos. Los conjuntos de datos son abstracciones de fenómenos de interés representadas como objetos, mismos que pueden describir un objeto del mundo real con sus respectivos atributos o características. Los atributos que se pueden representar son de tipo numérico, categórico o mixto (numérico y categórico). En los casos del mundo real, en su mayoría, se presentan conjuntos de datos de tipo mixto.

Por lo general los métodos de aprendizaje automático que se definen en la literatura asumen que un conjunto de datos es de tipo numérico o categórico, pero no mixto. A pesar de ello, dichos métodos no presentan una regla o procedimiento para lidiar con estas situaciones. Por lo que en este trabajo se busca definir una transformación de un conjunto de datos mixto que permita obtener resultados similares que a los que se obtendrían procesando un conjunto de datos original.

En este capítulo se describen los antecedentes y motivación de este trabajo, la hipótesis, los objetivos y la metodología de trabajo.

1.1 Antecedentes y motivación

El aprendizaje automático es una forma de programar una computadora para que aprenda a realizar ciertas tareas a partir de experiencias previas representadas en un conjunto de datos. Un conjunto de datos es la abstracción de fenómenos de interés representados en la forma de n objetos y un objeto es descrito en términos de atributos o características de tamaño m. En la Figura 1.1 se presenta un ejemplo de qué es un objeto, en este contexto se representa una persona con los atributos *altura*, *edad*, *peso*, *sexo* y *estado*.

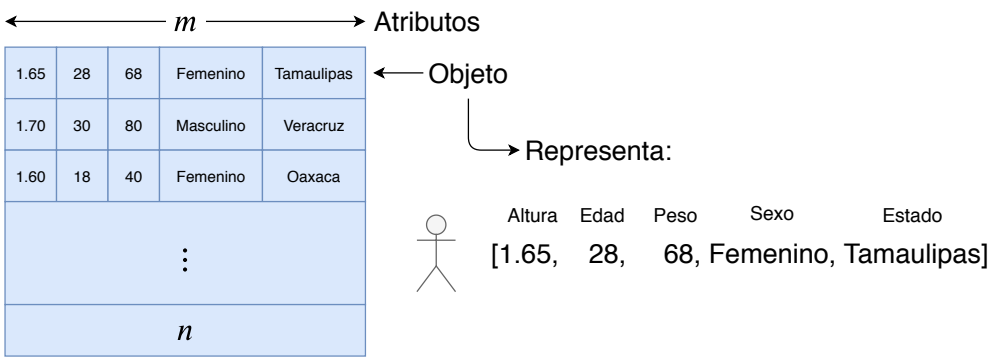


Figura 1.1: Abstracción de un conjunto de datos.

Sobre un conjunto de datos se pueden aplicar dos tareas que resaltan en el aprendizaje automático: clasificación y agrupamiento. Las características de los objetos se pueden presentar mediante tipos de dato numérico o categórico. Por ejemplo, en la Figura 1.2 se observa que un objeto representa una persona cuyos atributos de tipo numéricos son la *altura*, *edad* y *peso*, mientras que los atributos de tipo categóricos son *sexo* y *estado*.

Cuando los objetos presentan ambos tipos de dato se dice que el objeto es de tipo mixto. Las variables categóricas, como *sexo* y *estado*, se pueden presentar en la forma ordinal donde existe un orden de precedencia o en la forma nominal donde no existe dicho orden. Típicamente las variables categóricas en su forma nominal pueden ser reemplazadas por identificadores para reducir la complejidad de utilizar texto en las tareas de aprendizaje automático. En la Figura 1.3 se presenta

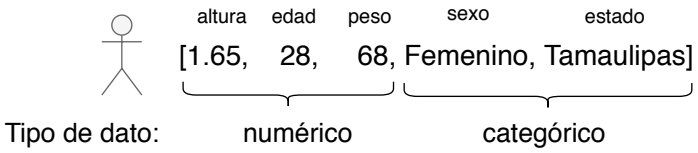


Figura 1.2: Tipos de dato que se pueden presentar en un objeto.

una posible equivalencia entre objetos, esto al reemplazar los valores de las variables categóricos por identificadores, donde sexo *Femenino* es 1 y sexo *Masculino* es 0; cada estado de México se representa con un caracter numérico con base a un orden alfabético.

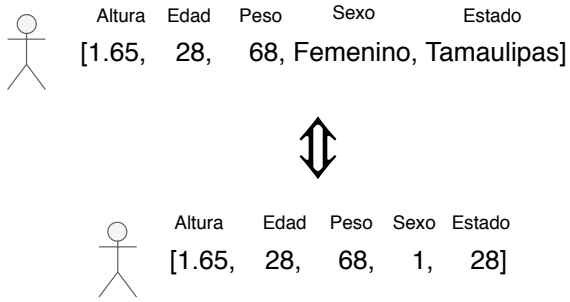


Figura 1.3: Equivalencia de atributos categóricos en su forma nominal.

Al realizar tareas de aprendizaje automático sobre los atributos de tipo de dato numérico existe un orden de precedencia. Por ejemplo, en la Figura 1.4 se presenta que el atributo numérico *altura* con valor 175 precede al valor 165. Sin embargo, para el caso del atributo categórico *estado*, en su forma nominal, no existe un orden de precedencia por lo que es incorrecto asumir que el valor *Zacatecas* (32) precede al valor *Tamaulipas* (28).

Ésto se vuelve aún más complicado cuando se presentan objetos de tipo de datos mixto. En la Figura 1.5 se ilustran dos objetos con características similares, la única diferencia es su valor en escala nominal del atributo *estado* donde el primer y segundo objeto tienen como valores *Aguascalientes* (1) y *Zacatecas* (32), respectivamente. En dicho ejemplo, no se puede decir que el primer objeto

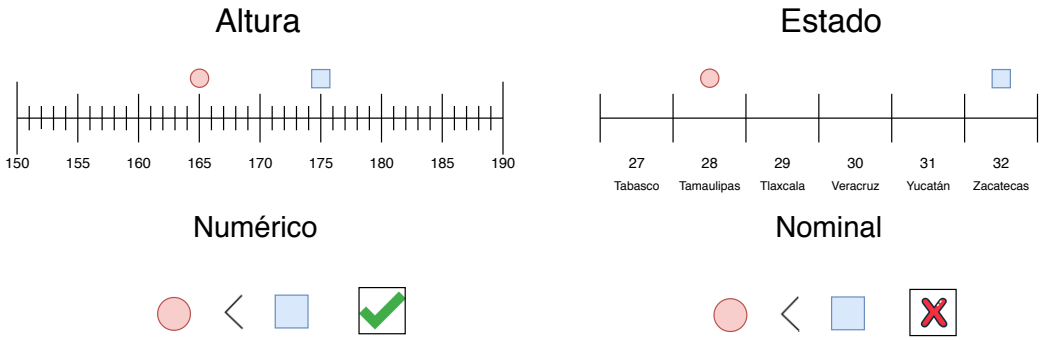


Figura 1.4: Orden de precedencia para los tipos de dato: numérico y categórico.

precede al segundo porque no existe un orden de precedencia.

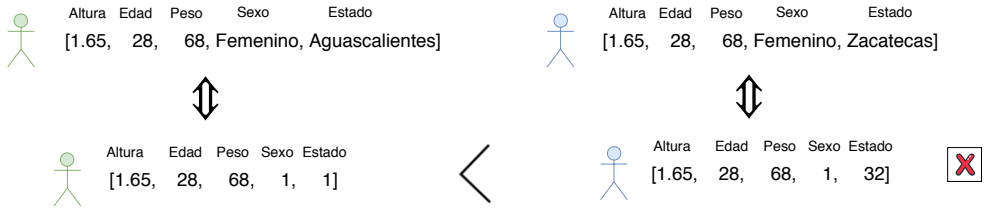


Figura 1.5: No es posible inducir un orden entre objetos de tipo de dato mixto.

También es incorrecto asumir o definir una medida de centralidad sobre los objetos de tipo de datos mixto. Como se puede observar en la Figura 1.6, una media de los dos objetos es incorrecto.

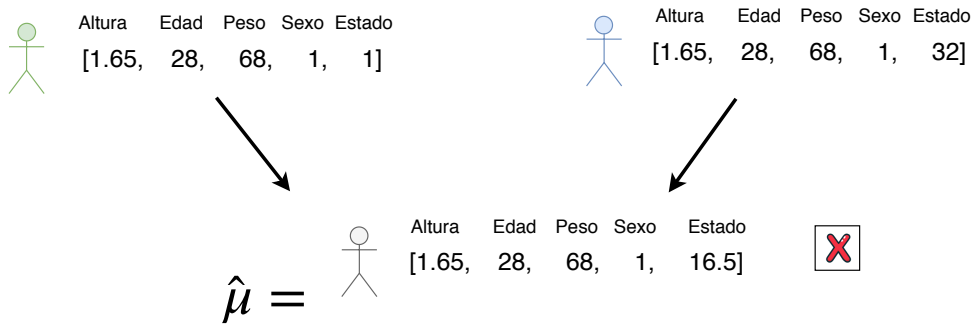


Figura 1.6: No es posible definir el concepto de medidas de centralidad.

Para lidiar con la problemática de los objetos de tipo de dato mixto existe una solución en la

literatura que se utiliza en una etapa de preprocesamiento llamado *one-hot encoding* [36]. El *one-hot encoding* es una codificación que permite convertir los atributos categóricos de los objetos, convirtiéndolos en atributos numéricos. Ésto es, por cada atributo categórico y dada la cantidad de valores únicos, éstos son convertidos a atributos numéricos donde el valor en el que incide el objeto es marcado con uno 1 y los demás como 0, tal como se ilustra en la Figura 1.7. Cuando se realiza una codificación a los objetos hay un aumento dimensional en función de la cantidad de los valores únicos pesentes en los atributos categóricos. En el ejemplo de la Figura 1.7 se contaba originalmente con 5 dimensiones, pero después de usar el *one-hot encoding* se obtuvieron 37 dimensiones: 3 dimensiones originalmente numéricos, 2 dimensiones correspondientes al número total de genéros y 32 dimensiones que corresponden al número total de estados. A este efecto se le conoce como *la maldición de la dimensionalidad* [48].

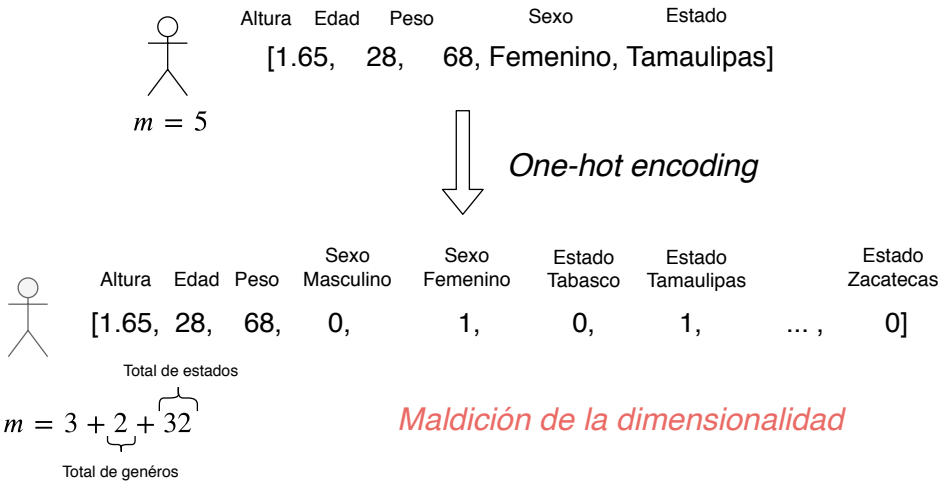


Figura 1.7: Preprocesamiento con la codificación *one-hot encoding* [36].

1.2 Hipótesis

Con base en lo presentado anteriormente, el presente trabajo plantea la siguiente hipótesis:

A partir de un conjunto de datos mixto, es posible obtener una transformación de los datos, que no implique un crecimiento dimensional de los mismos, y que a su vez permita realizar tareas de aprendizaje automático.

1.3 Objetivos

A partir de la hipótesis planteada, a continuación se presenta el objetivo general y objetivos específicos de este trabajo.

General

Obtener un método de transformación de conjuntos de datos mixtos que conserve las relaciones intrínsecas de los datos, con el propósito de realizar tareas de aprendizaje automático.

Específicos

- Definir la manera de representar datos en el espacio categórico.
- Definir un método para la transformación de conjuntos de datos mixtos.
- Crear un marco de experimentación que permita validar estadísticamente el desempeño del método propuesto.

1.3.1 Metodología propuesta

En esta sección se describe el proceso que se llevó acabo para cumplir con los objetivos planteados. En la Figura 1.8 se presentan las actividades que se describen a continuación.

1. **Revisión del estado del arte**. Fue necesario hacer una búsqueda de artículos publicados relacionados al problema de investigación. Con ello se esperaba encontrar las ventajas y desventajas de dichos trabajos y así complementar la propuesta de este trabajo.

2. **Redacción de la propuesta de tesis**. Se desarrolló el contenido del protocolo de tesis: descripción del proyecto, planteamiento del problema, objetivos generales y particulares del proyecto, metodología, cronograma de actividades, infraestructura, estado del arte y resultados esperados.

3. **Obtención del método de transformación de tipo de datos mixto**. Se desarrolló un método que transformara un conjunto de datos de tipo mixto a categóricos, donde en este nuevo espacio fuera posible mantener, dentro de lo posible, las relaciones del espacio euclídeo.

4. **Análisis formal de propiedades de los datos transformados**. Al ser transformado un conjunto de datos de tipo mixto a un espacio de categorías, se analizaron las propiedades de los objetos.

5. **Determinación de métrica de similitud**. Con respecto a un conjunto de datos transformado por el método propuesto, se buscó identificar una métrica de similitud que permitiera aprovechar las propiedades de dicha transformación.

6. **Búsqueda de conjuntos de datos mixtos**. Se buscó en repositorios públicos conjuntos de datos de tipo mixto para probar con la transformación propuesta y lo propuesto en la literatura.

7. **Selección de métodos de agrupamiento**. Dado que el conjunto de datos transformado se encontraba en el espacio de las categorías, se buscaron métodos de agrupamiento que permitieran evaluar las relaciones intrínsecas de los objetos. Además, se buscó un método de agrupamiento que permitiera evaluar lo propuesto en la literatura.

8. **Definición del diseño experimental**. Se especificaron las pruebas necesarias para que el método propuesto pudiera ser medido en función de la exactitud y el crecimiento dimensional.

9. **Análisis estadístico de resultados**. Se esperaba que al observar los resultados del método propuesto contra los métodos sugeridos en la literatura, se presentaran resultados mejorables a

favor del método propuesto. En caso contrario, se explicarían bajo qué casos no lo es y posibles mejoras.

10. **Redacción de las conclusiones**. Con base en la evaluación de los resultados se discutió el comportamiento de la transformación propuesta y una comparativa contra la literatura.

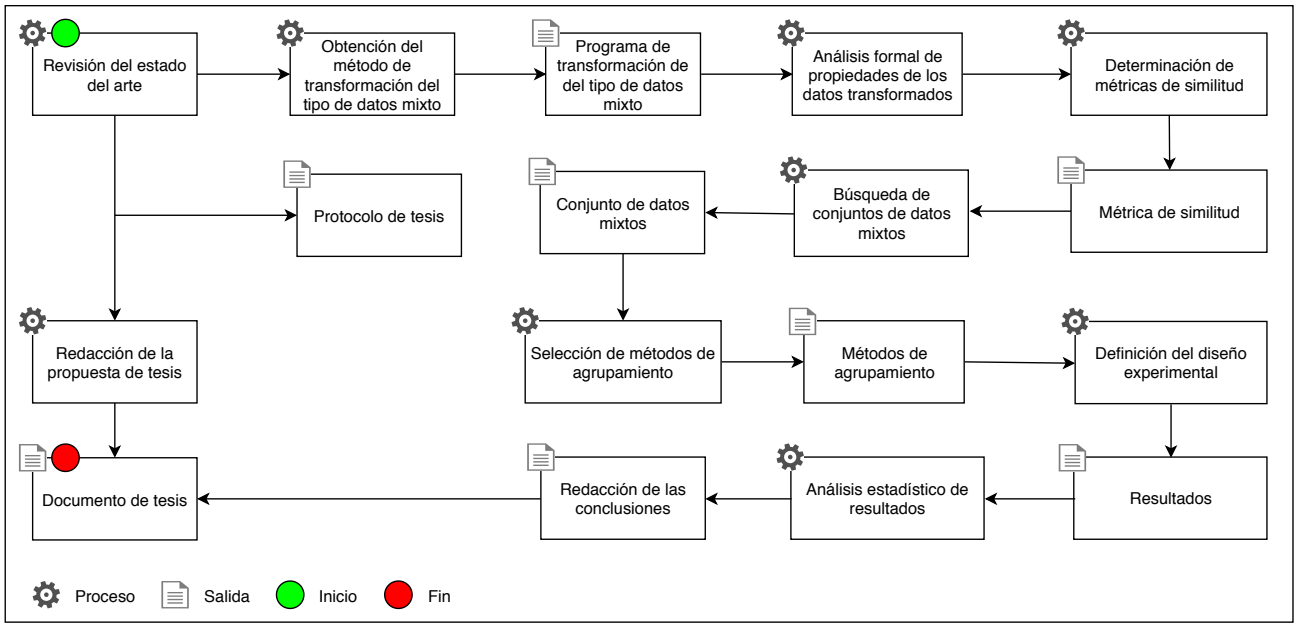


Figura 1.8: Diagrama de procesos.

1.4 Organización del documento

Este documento de tesis está compuesto de seis capítulos los cuales se describen a continuación.

- El Capítulo 2 es el marco teórico de la tesis donde se encuentran los conceptos básicos y necesarios para entender este trabajo.
- El Capítulo 3 se presenta el estado del arte de la tesis donde se describen los trabajos relacionados a la problemática que se busca solucionar con este trabajo.

- El Capítulo 4 es donde se describe la conceptualización e implementación de la solución propuesta.
- El Capítulo 5 describe la experimentación que se realizó, bajo qué infraestructura fue realizada y con qué métricas fue evaluada. Por último, se presentan los resultados de dicha experimentación donde se describe y discuten los resultados.
- En el Capítulo 6 se presentan las conclusiones del trabajo y los resultados obtenidos. También se discute el trabajo futuro de la solución propuesta.

2

Marco Teórico

Para entender la problemática que aborda este trabajo y sus conceptos básicos, en este capítulo se describen cuáles son los tipos de dato que se pueden encontrar en un conjunto de datos, qué es una transformación de datos y las tareas de aprendizaje automático como clasificación y agrupamiento, así como sus métricas de evaluación.

2.1 Tipos de dato

En las tareas de aprendizaje automático usualmente se asume que un conjunto de datos se presenta exclusivamente de tipo de dato categórico o numérico. Una forma general de ilustrar la taxonomía de los tipos de dato se puede observar en la Figura 2.1.

Los datos de tipo categórico, también conocidos como *cualitativos*, tienen valores de la forma de nombres, códigos o símbolos mutuamente excluyentes y satisfacen una relación de igualdad y diferencia. Éstos se dividen en dos escalas:

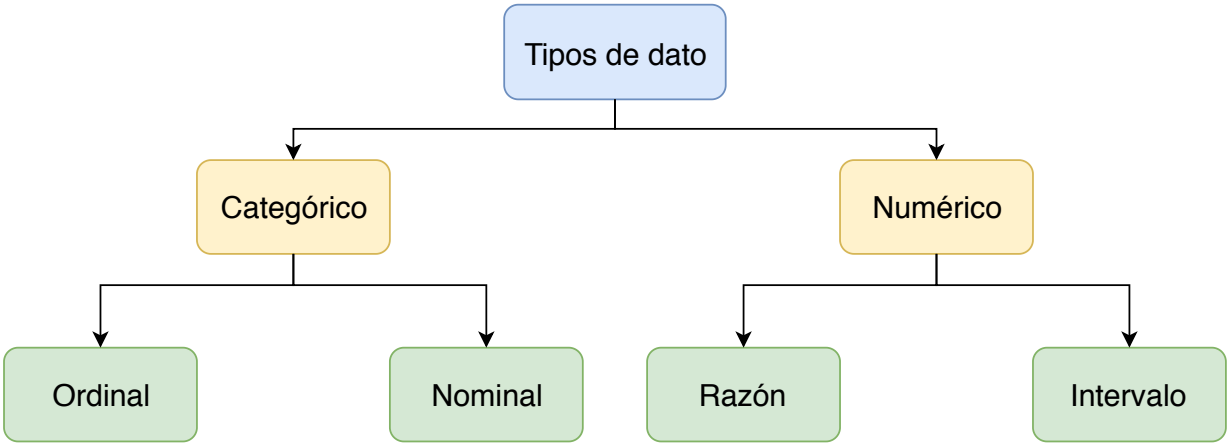


Figura 2.1: Taxonomía de tipos de dato.

- Nominal. Son etiquetas que sirven como identificadores que pueden presentarse en la forma de palabras, números, códigos o letras, de una serie de valores. Por ejemplo el género *Hombre* (0) y *Mujer* (1), ambos identifican diferentes individuos y tienen una relación de diferencia, en cambio si ambos individuos fueran *Hombre*, tendrían una relación de igualdad. Otros ejemplos de valores nominales son la placa serial de un auto, el número de un jugador de fútbol o la nacionalidad.

- Ordinal. Son etiquetas que preservan un orden de los identificadores y pueden presentarse en la forma de palabras, números, códigos o letras, de una serie de valores. Adicionalmente la escala ordinal cuenta con una relación de orden entre sus etiquetas. Por ejemplo la relación de orden se puede ver en el nivel educativo *Primaria* (0), *Secundaria* (1), *Preparatoria* (2) y *Universidad* (3), los cuales presentan un orden donde se tiene la Primaria (0) antes que *Secundaria* (1) y *Universidad* (3) es un valor con mayor peso que *Primaria* (0). Otros ejemplos de valores ordinales son la calidad del aire, la evaluación de satisfacción al cliente o el grupo sanguíneo.

Los datos de tipo numérico, también conocidos como *cuantitativos*, presentan valores de medición que satisfacen una relación de igualdad o desigualdad, así como una relacion de orden (propiedad

reflexiva, antisimétrica y transitiva). Éstos se dividen en dos escalas:

- Razón. Son valores representados en una forma numérica que tienen un orden, un valor exacto y cuentan con un cero absoluto. Por ejemplo la altura puede medirse en un rango de 0 a algún valor fijo, en el que no existen valores negativos. Otros ejemplos de valores de razón son el precio de un artículo, el peso o la duración en tiempo de un evento.

- Intervalo. Son valores representados en una forma numérica que tienen un orden, no cuentan con un cero absoluto y tienen la capacidad de ser diferenciados entre ellos. Por ejemplo la temperatura en grados Celsius que puede ser negativo (-10 °C) o positivo (40 °C). Otros ejemplos de valores de intervalo son los grados de temperatura (Fahrenheit, Kelvin, etc.), el estado de cuenta de banco o la métrica del coeficiente intelectual.

Con respecto a lo descrito anteriormente, las escalas presentan ciertas propiedades, medidas de tendencia central y operaciones aritméticas [80], lo que se resume en la Tabla 2.1.

Tabla 2.1: Propiedades de los tipos de dato.

	Nominal	Ordinal	Intervalo	Razón
Propiedades				
Hay un orden significativo		✓	✓	✓
Distancia significativa			✓	✓
Tiene un cero absoluto				✓
Medidas de tendencia central				
Moda	✓	✓	✓	✓
Mediana		✓	✓	✓
Media			✓	✓
Operaciones aritméticas				
Suma o resta			✓	✓
División o multiplicación				✓

En el análisis de datos los tipos de conjuntos de datos que se pueden encontrar son de tipo numérico, categórico o mixto; ésto se define acorde al tipo de dato que presentan los atributos del conjunto de datos. Los tipos de conjuntos de datos se describen como sigue:

- Numérico. Sus atributos pertenecen a escalas de intervalo y/o escalas de razón.
- Categórico. Sus atributos pertenecen a escalas nominales y/u ordinales.
- Mixto. Sus atributos presentan propiedades numéricas y categóricas.

Para realizar tareas de aprendizaje automático en conjuntos de datos mixtos, por lo general se realiza un preprocesamiento de los datos para unificar el tipo de dato a numérico o categórico. El preprocesamiento se define en la siguiente sección.

2.2 Preprocesamiento

El preprocesamiento de un conjunto de datos consiste en preparar adecuadamente los datos para que puedan ser procesados por métodos de aprendizaje automático [37]. Dicho proceso puede incluso mejorar el desempeño de las tareas de aprendizaje automático.

Existen diferentes técnicas de preprocesamiento que se pueden aplicar, aunque no todas se aplican:

(a) Selección de características, es un conjunto de técnicas utilizadas para reducir la cantidad de atributos presentes en un conjunto de datos por medio de un criterio de evaluación se seleccionan los atributos más relevantes [49, 54].

(b) Métodos de muestreo, son técnicas usadas para reducir la cardinalidad (número de filas) de un conjunto de datos, tomando la mejor representación estadística de la población. Algunos métodos de muestreo son *muestreo aleatorio simple*, *muestreo estratificado*, *muestreo conglomerados*, entre otros [3, 5, 30].

(c) Transformación de datos, es un conjunto de técnicas que permite el mapeo de objetos de un conjunto de datos a otra representación que permite resaltar propiedades que no eran visibles con anterioridad [70].

Una forma intuitiva de hacer el procesamiento de un conjunto de datos de tipo mixto es separando los atributos numéricos y categóricos, calcular sus respectivas métricas de similitud, para después unirlas en una única métrica. Sin embargo, de esta manera no queda del todo claro si la métrica tiene significado, o si ambas métricas se encuentran bajo la misma escala. Para resolver esta problemática se recurre a una tarea de transformación de datos, donde un conjunto de datos es transformado a una nueva representación exclusivamente numérica o categórica. Algunas técnicas de transformación de datos que destacan son escalamiento, limpieza de los datos y codificación.

Escalamiento

El escalamiento es un proceso que permite unificar variables que cuentan con diferentes unidades de medida [46]. El realizar un escalamiento se reduce la posibilidad de obtener un sesgo al realizar tareas de aprendizaje automático. En este ámbito, las técnicas de escalamiento más conocidas son:

- Z-score (también conocido como *estandarización*), que permite cambiar los valores de una variable de tal forma que la distribución normal de los datos sigue una media de 0 y una desviación estándar de 1. El Z-score se define como indica la Ecuación 2.1.

$$z = \frac{x_i - \mu}{\sigma}, \tag{2.1}$$

  donde x_i es un valor de una variable, μ es la media y σ es la desviación estándar de los valores en una variable.

- Min-max (también conocido como *normalización*), que permite cambiar los valores de una variable a un rango de valores $[0, 1]$. Dicho procedimiento es utilizado para escalar valores con una desviación estándar no alta, cuyos datos no pueden ser representados usando una distribución normal. Cabe resaltar que este método puede ser sesgado si hay valores atípicos.

Min-max se define como indica la Ecuación 2.2.

$$x'_i = \frac{x_i - \text{min}(x)}{\text{máx}(x) - \text{min}(x)}, \tag{2.2}$$

donde x_i es un valor de una variable, máx(x) el valor máximo de la variable x y min(x) el valor mínimo de la variable x.

Existen otras estrategias de escalamiento que son utilizadas dependiendo de la naturaleza de los datos. Ejemplos de ello son la escala logarítmica, la escala de raíz cúbica y la escala de raíz cuadrada, que son ampliamente usados para visualizar datos con valores altos. El escalamiento es una tarea importante del preprocesamiento porque permite que múltiples variables puedan aportar de forma equitativa sin sesgar el fenómeno de interés al hacer tareas de aprendizaje automático.

Limpieza de los datos

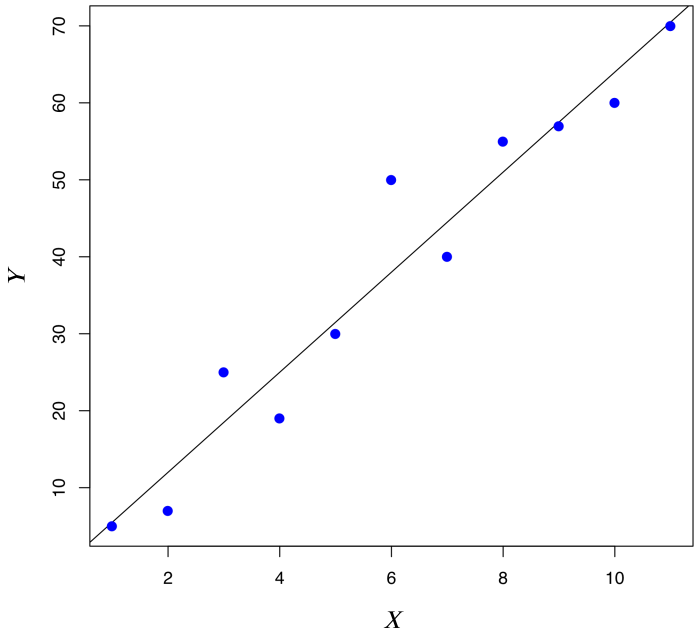
La limpieza de los datos es el proceso de eliminar atributos o sustituir valores vacíos, atípicos dados por errores de captura, ruido de algún instrumento de medición o falla del mismo, que no aportan información al análisis de un fenómeno de interés [64]. A continuación se describen dos enfoques de limpieza de los datos: por sustitución de valores vacíos o atípicos ó por eliminación de atributos.

Sustitución de valores vacíos o atípicos

Los valores vacíos o atípicos pueden sustituirse por:

- Un valor a discreción que es determinado con base al contexto del fenómeno de interés.
- Una regla definida por un diccionario o fórmula que permita estudiar el fenómeno de interés.
- Un valor basado en una medida de tendencia central como la media, la mediana o la moda.

También existe otro enfoque para la sustitución de valores vacíos, ésto por medio de la predicción de valores faltantes con métodos de regresión o interpolación. Un método de regresión permite la predicción de valores de una variable a partir de los valores de otra variable. La técnica más básica de regresión es la regresión lineal, que asume la existencia de una variable Y conocida como la *variable dependiente*, misma que se busca predecir; también se cuenta con otra variable X llamada *variable predictora*, que sirve como base para predecir valores de la variable Y [62]. Posteriormente, los datos de las variables X y Y se visualizan en términos de X con una línea recta, misma que se fija con el objetivo de obtener la mejor posición que atraviese todos los puntos. La línea recta permite predecir un valor en Y dado otro valor en X. Para determinar la mejor posición de la línea recta se calcula la suma de los errores cuadrados, con base a ello se decide ajustar la posición de la línea recta. Un ejemplo de una regresión lineal se presenta en la Figura 2.2, donde se visualiza una línea recta que permite predecir valores en Y dado otros valores en X.



X	Y
8	5
9	7
10	25
11	19
12	30
13	50
14	40
15	55
16	57
17	60
18	70

Figura 2.2: Regresión lineal de la variable Y con respecto a X.

La interpolación es otro procedimiento aplicado para la sustitución de valores vacíos, el cual es un proceso en el que dado un número de valores conocidos de una variable es posible inferir sus valores desconocidos [24]. Se asume que al menos se conocen dos valores y existe una taza de cambio entre ellos. Algunas técnicas ampliamente usadas son la interpolación por *splines* (lineales, cúbicos y cuadráticos) [40], *de Lagrange*, *de Hermite*, *de Taylor*, etcétera.

Eliminacion de atributos

La eliminación de atributos es el proceso de eliminar atributos que no aportan información al análisis de un fenómeno de interés, ésto permite reducir la redundancia de los datos. A continuación se describen las formas más utilizadas de eliminación de atributos:

- La forma de eliminar atributos más básica es identificando la proporción de los datos vacíos por atributo. Esto es, si la proporción de datos vacíos es mayor a la proporción de los datos que aportan información, se prosigue a eliminar el atributo. En situaciones donde se conoce el contexto del origen de los datos se puede llegar a usar una estrategia de sustitución de datos.

- Otra estrategia es proyectando las distancias de las instancias de un conjunto de datos, manteniendo los atributos que no alteren las distancias entre las instancias. Algunos métodos que permiten hacer una proyección de los conjuntos de datos son el *análisis de componentes principales* [45] (PCA), *mapa de Sammons* [77], *análisis lineal discriminante* [44] (LDA), entre otros.

- Los atributos también pueden ser eliminados por la identificación de aquellos que se encuentren correlacionados entre si. Ésto se debe a que un atributo puede explicar un mismo fenómeno de interés que al usar dos o más atributos.

- Otra forma de eliminar los atributos es a través de la ganancia de información usando etiquetas de clase [58]. De tal forma que al seleccionar un atributo para una etiqueta de clase se calcula la

ganancia de información. Posteriormente, los atributos con la menor ganancia de información son eliminados, con ello se reduce la incertidumbre presente en un conjunto de datos.

Codificación

La codificación es el proceso de transformar datos de un tipo dato o formato a otro que facilite el procesamiento o una mejor interpretación de los datos [6]. Por lo general el concepto de codificación se presenta en un contexto de transmisión de datos. Sin embargo, para el contexto de este trabajo, se utiliza la codificación para la representación de datos. Para ello se definen los casos de codificación para la representación de un conjunto de datos. Como se describe en la Sección 2.1, existen dos tipos de dato: el numérico y el categórico. Cuando se mezclan ambos tipos de dato en un conjunto de datos, a estos se les llama conjunto de datos de tipo mixto. En este tipo de contextos se llevan acabo las codificaciones, donde se busca convertir un conjunto de datos de tipo mixto al tipo numérico o categórico. Ésto permite que los datos sean procesados por tareas de aprendizaje automático o que sean analizados desde otra perspectiva. Con respecto a lo anterior, es importante destacar que solamente se han documentado en la literatura codificaciones que buscan transformar los atributos categóricos a numéricos de un conjunto de datos. Es posible que eso se deba a que la mayoría de las tareas de aprendizaje automático funcionan sobre el espacio numérico. Algunas técnicas que sobresalen en la codificación de datos de tipo categórico a numérico son la codificación one-hot encoding [6], la codificación binaria [39] y la codificación hash [87] que se discuten en el Capítulo 3.

2.3 Tareas de aprendizaje automático

Típicamente una tarea de aprendizaje se define o elige acorde a lo que se espera obtener con los datos que se usan a procesar. Las tareas de aprendizaje más conocidas son clasificación y agrupamiento, mismas que se describen a continuación.

Clasificación

La clasificación permite predecir las categorías a las que pertenecen las instancias de un conjunto de datos por medio del entrenamiento de un modelo. El proceso de clasificación se divide en dos etapas: primero se asume como entrada un conjunto de datos etiquetados, donde sus objetos permiten determinar un modelo de membresía. Por último, con la llegada de objetos que no estaban en el conjunto de datos original, éstos son etiquetados a través del modelo de membresía identificado [15]. Dicho proceso se ilustra en la Figura 2.3, donde se recibe como entrada un conjunto de datos etiquetados que determinan un modelo de membresía, mismo que es usado para etiquetar objetos desconocidos sin etiqueta.

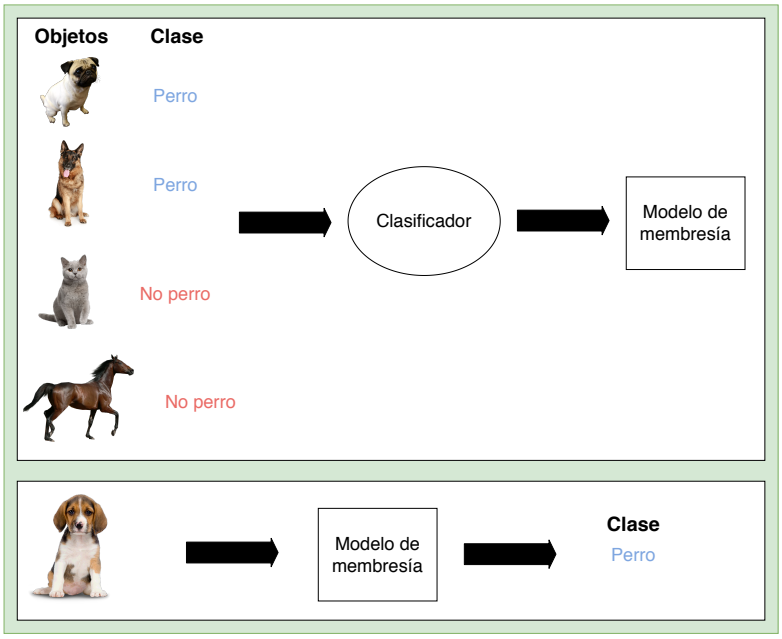


Figura 2.3: Proceso de clasificación de un conjunto de datos.

Cuando se implementa una tarea de clasificación sobre un conjunto de datos, éstos son divididos en dos subconjuntos: entrenamiento y prueba. El subconjunto de entrenamiento es usado para determinar el modelo de membresía. Mientras que el subconjunto de prueba es usado para evaluar el desempeño de clasificación del modelo de membresía.

Para medir el desempeño de un clasificador existen diferentes métricas que típicamente requieren de la creación de una matriz de confusión. La matriz de confusión es el conteo de los objetos clasificados con el modelo de membresía con respecto a las etiquetas de clase originales [32]. En la Figura 2.4 se observa una matriz de confusión, misma que se compone de los siguientes elementos:

- Verdadero Positivo (VP) es el total de objetos clasificados correctamente con respecto a la etiqueta de clase positiva (identificación de la clase a la cual un objeto pertenece).
- Verdadero Negativo (VN) es el total de objetos clasificados correctamente con respecto a la etiqueta de clase negativa (identificación de la clase a la cual un objeto no pertenece).
- Falso Positivo (FP) es el total de objetos clasificados incorrectamente con respecto a la etiqueta de clase positiva.
- Falso Negativo (FN) es el total de objetos clasificados incorrectamente con respecto a la etiqueta de clase negativa.

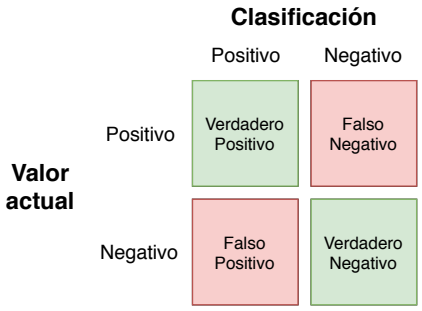


Figura 2.4: Matriz de confusión de las etiquetas originales contra lo clasificado.

Después, con respecto a la matriz de confusión se calculan las siguientes métricas [67]:

- Exactitud, mide la fracción del total de aciertos realizados por el modelo de membresía. Formalmente se define como indica la Ecuación 2.3.

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.3}$$

- Precisión, mide la fracción del total de aciertos positivos que fue correctamente identificado por el modelo de membresía. La precisión se define como indica la Ecuación 2.4.

$$\text{Precisión } = \frac{VP}{VP + FP}. \tag{2.4}$$

  Cuando se obtiene un valor de 1.0 en precisión, significa que no se producieron falsos positivos.

- Exhaustividad, mide la fracción del total de aciertos positivos que se identificaron correctamente por el modelo de membresía. Formalmente se define como indica la Ecuación 2.5.

$$\text{Exhaustividad } = \frac{VP}{VP + FN}. \tag{2.5}$$

  Cuando se obtiene un valor de 1.0 significa que no se producieron falsos negativos.

- Medida F_1 es el promedio armónico de la precisión y la exhaustividad. Sus valores pueden ser entre 0 (malo) y 1 (bueno). Se calcula como indica la Ecuación 2.6.

$$F_1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}. \tag{2.6}$$

  Dicha medida puede ser usada para reemplazar la medición de la precisión y la exhaustividad, sin comprometer ninguna de las métricas.

- Especificidad, mide la fracción del total de aciertos negativos que se identificaron correctamente por el modelo de membresía. Se define como se muestra en la Ecuación 2.7.

$$\text{Especificidad } = \frac{VN}{VN + FP}. \tag{2.7}$$

  Cuando se obtiene un valor de 0 significa que no se producieron verdaderos negativos.

De acuerdo al estado del arte, los métodos de clasificación más conocidos son: *clasificador*

bayesiano ingenuo [59], *vecino más cercano* [79] (kNN), *máquinas de vectores de soporte* [22] (SVM), *árboles de decisión* [68] y *redes neuronales* [76].

Agrupamiento

El proceso de agrupamiento consiste en que dado un conjunto de datos se buscan las relaciones de similitud que hay entre los objetos del conjunto de datos, de tal forma que surgan grupos de objetos con atributos similares [15], tal como se ilustra en la Figura 2.5.

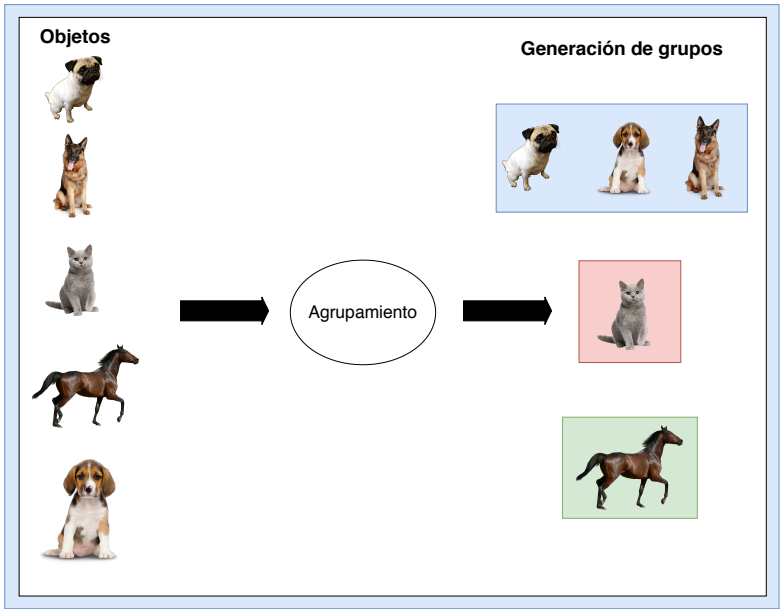


Figura 2.5: Proceso de agrupamiento de un conjunto de datos.

Por lo general las relaciones de similitud son establecidas por medio de métricas de distancia donde la métrica más utilizada es la *distancia Euclidiana*. Adicionalmente existen otras métricas como la *distancia Manhattan*, *Chebyshev*, *Cosenos*, *Hamming*, entre otros.

Para evaluar el desempeño del proceso de agrupamiento se pueden considerar dos escenarios:

(a) Cuando se tiene conocimiento de las etiquetas de clase del conjunto de datos se busca medir la similitud entre las etiquetas de clase obtenidas por la técnica de agrupamiento y las etiquetas de clase original, algunas métricas que lo permiten son:

- *Fowlkes-Mallows* [33] es una métrica basada en la media geométrica entre la precisión y exhaustividad, empleados en clasificación.
- *Adjusted Mutual Information* [84], mide el acuerdo que hay entre dos agrupaciones de etiquetas de clase por medio de la información mutua.
- *Rand Index* [71], mide en términos de porcentaje la similitud de dos agrupaciones considerando todos los pares de elementos que se encuentran dentro y fuera de un mismo grupo.
- *Adjusted Rand Index* [43], toma como base la métrica del *Rand Index*, donde ajusta dicha métrica al remover el valor esperado, ésto permite remover la aleatoriedad en los resultados, cambiando su escala de -1 a 1. Cuanto más cercano a 1, es mayor la similitud entre agrupaciones.

(b) Cuando un conjunto de datos no presenta etiquetas de clase, se calculan índices que permiten evaluar los grupos por medio su diámetro, tamaño, densidad, entre otras propiedades. Los índices más comunes son los siguientes:

- *Silhouette Index* [75], mide la similitud entre objetos dentro de un grupo comparado contra otros por cada muestra tomada. Dicha métrica da como resultado valores entre 1 y -1, donde 1 es el mejor y -1 es el peor. Los valores negativos indican que la muestra ha sido asignado al grupo equivocado.
- *Davies-Bouldin* [23], mide el promedio de la similitud para cada grupo con respecto a su grupo más similar. La similitud es medida en relación con las distancias dentro del grupo y con respecto a otros grupos. Entre más cercano a cero sea el resultado, significa que los grupos se encuentran más separados y menos dispersos.
- *Calinski-Harabasz* [51], mide la relación de la dispersión que hay dentro y fuera de los grupos.

- *Dunn Index* [29], mide la relación que hay entre la distancia más cercana de un punto que no pertenece a un mismo grupo contra la distancia más lejana dentro de un grupo. Los valores resultantes pueden estar en un rango de 0 a infinito. Cuanto mayor sea el resultado es mejor el agrupamiento.

De acuerdo al estado del arte, los métodos de agrupamiento más conocidos son: k-*means* [56], k-*modes* [42], *mean-shift* [34], *fuzzy* c-*means* [14], *agrupamiento espacial basado en densidad de aplicaciones con ruido* (DBSCAN) [31], entre otros.

3

Estado del Arte

En este capítulo se presenta el estado del arte sobre los tipos de conjuntos de datos que existen y cuál es su preprocesamiento desde el punto de vista de tareas de aprendizaje automático. También se presenta el concepto de transformación de datos, dentro del cual se explora una técnica llamada codificación. Se describen las codificaciones más utilizadas en la literatura y se discuten sus limitantes que se buscan solventar como parte de la solución propuesta en el Capítulo 4.

3.1 Introducción

El aprendizaje automático permite a una computadora aprender con base en experiencias previas. Estas experiencias son representadas en un conjunto de datos. Los conjuntos de datos contienen objetos del dominio que se modela, donde un objeto representa un fenómeno del mundo real y constituye una instancia en el conjunto de datos. Un ejemplo de un objeto puede ser la representación de una persona con características tales como la altura, edad, peso y profesión. Las características de un objeto se pueden presentar de tipo numérico, categórico y mixto (como se describe en la

Sección 2.1). Cuando un conjunto de datos presenta objetos con características exclusivamente de tipo numérico se dice que el conjunto de datos es de tipo numérico. Mientras que si un conjunto de datos presenta objetos con características exclusivamente de tipo categórico se dice que el conjunto de datos es de tipo categórico. Cuando un conjunto de datos presenta características de tipo numérico y categórico combinadas, se dice que el conjunto de datos es de tipo mixto. Típicamente en la mayoría de las tareas de aprendizaje automático se asume que los conjuntos de datos son de tipo numérico. Por lo que cuando se presentan conjuntos de datos de tipo categórico éstos usualmente requieren de un procedimiento que permita cambiar su representación a un espacio numérico; ésto con el fin de mejorar su representación. Lo mismo sucede al tratar conjuntos de datos mixtos, éstos requieren de un procedimiento que permita cambiar la representación a un espacio completamente numérico, sólo en algunas circunstancias categórico. El procedimiento requerido es una transformación de los datos (codificación) mediante la cual es posible representar conjuntos de datos de tipo categórico y mixto a conjuntos de datos netamente de tipo numérico o categórico. En las siguientes secciones se describen los tipos de conjuntos de datos, así como los métodos de aprendizaje automático que usualmente se utilizan acorde al tipo de dato.

3.1.1 Conjuntos de datos numéricos

Un ejemplo de un conjunto de datos de tipo numérico se ilustra en la Tabla 3.1, donde se tiene información del *peso*, *altura* y *edad* de personas, dichos atributos se encuentran en escalas de razón.

Tabla 3.1: Conjunto de datos numérico.

	Peso (kg)	**Altura (mts)**	**Edad**
1	50	1.73	25
2	70	1.60	29
3	60	1.60	30
4	45	1.42	17
5	70	1.80	17

Al contar con atributos numéricos por consiguiente existe un orden y una distancia entre los

objetos. Por ejemplo la primera y cuarta instancia de la Tabla 3.1 para el atributo de *peso* con un valor de 50 y 45, respectivamente, existe un orden porque 45 precede a 50, existe una distancia de 5 kg.

Una de las características de los atributos numéricos es que se pueden realizar operaciones aritméticas tales como suma o resta y división o multiplicación (para el caso de los valores que se encuentran en una escala de razón). Por ejemplo, el total del *peso* de todas las personas presentadas en la Tabla 3.1 es de 295 kg. O cambiar la unidad de medida de la altura de metros a pulgadas, lo cual implica una división o multiplicación.

Otra característica que tienen los atributos numéricos es que se pueden realizar cálculos estadísticos con sus valores, por ejemplo calcular medidas de tendencia central. Tomando de nuevo el ejemplo de la Tabla 3.1, se pueden calcular medidas de tendencia central y dispersión que permiten entender y modelar el comportamiento de los datos. Por ejemplo para el atributo *peso* su moda es 70, su mediana es 60, su media es 59 y cuenta con una desviación estándar de 11.40. Con esto podríamos entender que el peso promedio de las personas es de 59 kg y que puede presentarse con $\pm$ 11.40 kg a partir de la media.

Con las características que los conjuntos de datos numéricos presentan es posible realizar tareas de aprendizaje automático, tanto de clasificación como de agrupamiento y regresión. Los métodos de clasificación que son ampliamente usados en conjuntos de datos de tipo numérico son: *aprendizaje bayesiano* [63], k *vecinos más cercanos* (kNN) [9], *máquinas de vectores de soporte* (SVM) [22], *redes neuronales* [25], *árboles de decisión* [74] y *bosques aleatorios* [16]. Para el caso de las tareas de agrupamiento los métodos que destacan son: k*-means* [56], *mean-shift* [34], *fuzzy* c*-means* [14], k*-medoids* [47], *agrupamiento basado en densidades* (DBSCAN) [31] y *mapas auto-organizados* (SOM) [50].

3.1.2 Conjuntos de datos categóricos

Los conjuntos de datos categóricos presentan en sus atributos valores alfabéticos o alfanuméricos que se pueden presentar en escala nominal y/o escala ordinal. En la Tabla 3.2 se presenta un ejemplo de un conjunto de datos de tipo categórico, donde se tiene como información la *nacionalidad*, *profesión*, *nivel educativo* y *código postal* de personas. En el que los atributos *nacionalidad*, *profesión* y *código postal* se encuentran en la escala nominal, mientras que el atributo nivel educativo se encuentra en la escala ordinal.

Tabla 3.2: Conjunto de datos categórico.

	Nacionalidad	**Profesión**	**Nivel educativo**	**Código postal**
1	Mexicana	Ingeniero	Universitaria	87130
2	Alemana	Cocinero	Preparatoria	64460
3	Chilena	Médico	Universitaria	67170
4	Italiana	Cocinero	Secundaria	87000
5	Alemana	Mecánico	Primaria	87130

Con los datos de las instancias de un conjunto de datos categórico no se pueden realizar operaciones aritméticas, aunque sean caracteres numéricos (por ejemplo, el atributo *código postal* de la Tabla 3.2). En estos conjuntos de datos la única medida de centralidad admisible es la moda. En la Tabla 3.2 el atributo *nacionalidad* su valor central es *Alemana* y del atributo profesión el valor de centralidad es *Cocinero*. En el ejemplo de la Tabla 3.2 se puede apreciar que los valores del atributo *nivel educativo* tienen un orden: *Primaria*, *Secundaria*, *Preparatoria* y *Universitaria*, donde también se puede calcular el valor de centralidad de la moda: *Preparatoria*.

Dado que en los valores de tipo categórico no es posible realizar operaciones aritméticas, por ende tampoco es posible obtener medidas de centralidad como la media o mediana (con la excepción de los valores en escala ordinal). El análisis de dichos valores recae en un proceso de conteo del número de ocurrencias de los valores por categoría, a este proceso se le conoce como la generación de la tabla de frecuencias o distribución de frecuencias. En la tabla de frecuencias se calculan dos tipos de frecuencias: la *frecuencia absoluta* que es el número de ocurrencias de los valores por categoría

y la *frecuencia relativa* que es la frecuencia en términos de porcentajes del total de los valores. Adicionalmente se calculan las frecuencias acumuladas para *la frecuencia absoluta* y *la frecuencia relativa*, que implican la suma acumulada de inicio a fin de las frecuencias. Un ejemplo de la tabla de frecuencias se muestra en la Tabla 3.3, en la que se obtienen las frecuencias del atributo *profesión* de los datos de la Tabla 3.2.

Tabla 3.3: Tabla de frecuencias del atributo profesión.

Valores	Frecuencia Absoluta	Frecuencia Acumulada	Frecuencia Relativa (%)	Frecuencia Relativa Acumulada (%)
Ingeniero	1	1	20 %	20 %
Cocinero	2	3	40 %	60 %
Médico	1	4	20 %	80 %
Mecánico	1	5	20 %	100 %

Como se puede observar en la Tabla, con respecto a la frecuencia relativa, el 40 % de las instancias fueron de la profesión *Cocinero*, mientras que el 60 % se distribuye equitativamente con los valores *Ingeniero*, *Médico* y *Mecánico*. También, los valores de tipo categórico se les puede calcular su distribución de probabilidad por medio de una técnica de muestreo. Las distribuciones de probabilidad que usualmente se usan para datos de tipo categórico son la *distribución Bernoulli*, *binomial* y *multinomial* [1]. La *distribución Bernoulli* es la distribución de probabilidad que describe el resultado de que un evento ocurra (como éxito o fracaso), los datos que podrían ser descritos por esta distribución serían valores con dos categorías como las respuestas de una pregunta (por ejemplo, sí o no). La *distribución binomial* tiene como base la *distribución de Bernoulli*, con la diferencia de que cuenta la cantidad de éxitos por n cantidad de experimentos (éxito o fracaso de un evento). La *distribución multinomial*, también conocida como *distribución discreta* o *categórica* describe el resultado de un evento aleatorio en el que un evento sucede de k posibles salidas, cada uno con diferentes probabilidades. Los datos que podría describir la *distribución multinomial* serían valores con múltiples categorías, como la probabilidad de elegir el color *rojo* de los colores *verde*, *azul*, *blanco*, *rojo* y *amarillo*.

Dadas las propiedades descritas de los conjuntos de datos de tipo categórico es posible realizar tareas de clasificación con base en las frecuencias de las categorías con métodos como *naive bayes* [60] o con base en la entropía como ID3 [69]. Para tareas de agrupamiento existen más métodos como: k-*modes* [42], *rock* [38], *coolcat* [11], *cactus* [35], *limbo* [10] y agrupamiento basado en entropía [4].

3.1.3 Conjuntos de datos mixtos

Los conjuntos de datos mixtos presentan una combinación de atributos de tipo categórico y numérico, ésto se puede ejemplificar con la Tabla 3.4, donde se presentan tres atributos numéricos (*peso*, *altura*, *edad*) y tres atributos categóricos (*nacionalidad*, *profesión*, *nivel educativo*). En los conjuntos de datos de tipo mixto, usualmente no existe un balance que permita decir que un conjunto de datos sea más de tipo categórico o numérico, por lo general simplemente se le conoce como de tipo mixto. Cuando en un conjunto de datos se cuenta con la presencia de datos de tipo categórico y numérico o mixtos, éstos se pueden manipular de dos formas:

Tabla 3.4: Conjunto de datos mixto.

	Peso (kg)	Altura (mts)	Edad	Nacionalidad	Profesión	Nivel educativo
1	50	1.73	25	Mexicana	Ingeniero	Universitario
2	70	1.60	29	Americana	Cocinero	Preparatoria
3	60	1.68	30	Chilena	Médico	Universitario
4	45	1.42	17	Italiana	Cocinero	Universitario
5	90	1.80	37	Americana	Mecánico	Preparatoria

(a) Procesamiento directo de los datos [2, 42, 72, 78, 83]. En esta categoría los métodos procesan directamente los conjuntos de datos de tipo mixto, mismos que son aplicados directamente sobre las tareas de aprendizaje automático tales como clasificación y agrupación. Esto permite una manipulación más eficiente y directa de las instancias de un conjunto de datos acorde al tipo de dato de los atributos. Por lo general en este tipo de métodos se define una regla que permite la ponderación de los valores de tipo categórico, que por medio de alguna operación

aritmética éstos puedan ser ponderados en conjunto con los valores de tipo numérico, y así puedan ser diferenciados ante otras instancias. Sin embargo, no para todos los contextos de datos puede funcionar dicha ponderación. Un ejemplo es cuando se desconoce el orden de importancia de los valores de un atributo de tipo categórico, éstos se consideran en su forma nominal, por lo que no tienen un orden y la única medida de centralidad admisible es la moda o algún otro método que esté basado en frecuencias.

(b) Preprocesamiento a través de una transformación de datos [52, 66]. Dado un conjunto de datos de tipo mixto, éstos son transformados a un espacio diferente al original, el cual permita una manipulación más eficiente para que dichos datos sean procesados por tareas de aprendizaje automático en el espacio numérico o categórico.

En este trabajo se propone un método de transformación de datos (b) que permite realizar la transformación de conjuntos de datos mixtos usando una codificación de datos numéricos a categóricos para poder realizar tareas de aprendizaje automático. A continuación se presentan las propuestas de transformación de datos (codificación) más representativas.

3.2 Transformación de datos

La transformación de datos es la creación de un conjunto de nuevas características a partir del conjunto original de los datos por medio de un mapeo [49]. Esto permite una mayor discriminación entre objetos en un espacio diferente al original, donde es posible que se puedan denotar nuevos patrones no presentes en el espacio original de los datos. Dicha transformación es útil porque se pueden aplicar tareas de aprendizaje automático sobre este nuevo espacio y de esta forma obtener una mejor generalización del fenómeno de interés. Si bien existen diferentes técnicas de transformación de datos (Sección 2.2), el enfoque de este trabajo es específicamente sobre la técnica de codificación. En la codificación de datos existen dos enfoques: a) convertir datos numéricos a categóricos y b)

convertir datos categóricos a numéricos. Sin embargo, en la literatura no hay trabajos que documenten el enfoque del punto (a). Por lo tanto, a continuación se describen codificaciones de datos categóricos.

Codificación de datos categóricos

En la literatura existen diferentes codificaciones que son usadas para codificar valores categóricos a valores numéricos [52, 66]. El tipo de codificación es seleccionado acorde al procesamiento que se desea realizar sobre las instancias del conjunto de datos. Las variables categóricas típicamente se encuentran representadas por un número finito de posibles valores, donde no es posible hacer operaciones aritméticas o para el caso de la escala nominal, no tienen un orden (ver Sección 2.1). La mayoría de las tareas de aprendizaje automático tienden a usar representaciones numéricas, no categóricas, por lo que se recurre a métodos de codificación para representar atributos categóricos en el espacio numérico. En la Tabla 3.5 se presenta un conjunto de datos mixto que se utiliza para ejemplificar algunas de las codificaciones que se presentan más adelante en las siguientes subsecciones.

Tabla 3.5: Conjunto de datos mixto.

	Género	**Profesión**	**Salario**
1	Masculino	Ingeniero	20000
2	Femenino	Ingeniero	25500
3	Masculino	Médico	18000
4	Femenino	Arquitecto	27500
5	Masculino	Carpintero	15000
6	Masculino	Arquitecto	19000
7	Femenino	Mecánico	16000
8	Femenino	Mecánico	21500
9	Femenino	Arquitecto	27200
10	Masculino	Arquitecto	29300

Las codificaciones de datos categóricos a numéricos más comunes que se utilizan en la literatura se presentan a continuación [7]:

- La *codificación one-hot encoding* [36] (o también conocido como *dummy variables*) es la codificación base, cada atributo categórico es sustituido por atributos numéricos, donde el

número de atributos es igual al total de los valores del atributo categórico (como se describe en la Sección 2.2). Cada uno de los nuevos atributos indica la presencia (1) o ausencia (0) del valor categórico en el que incide. En la Tabla 3.6 se puede observar la codificación de los atributos *género* y *profesión* de la Tabla 3.5. Donde el atributo *género*, en la segunda instancia, tiene como valor en la columna *gen_0* $= 1$, que corresponde a la categoría *Femenino* y *gen_1* $= 0$ que corresponde a la categoría *Masculino*. Ésto también se puede observar en el atributo *profesión* donde *prof_1* $= 1$, correspondiente a la categoría *Ingeniero*. En la literatura se han propuesto métodos que hacen uso de la *codificación one-hot encoding*, como Du *et al*. [28] que utilizan dicha codificación para representar relaciones entre grafos que se encuentran en el espacio categórico. Li y Axhausen [55] proponen el uso de la *codificación one-hot encoding* para modelar series de tiempo de un conjunto de datos para la predicción del tráfico de servicios de transporte. Con la codificación resultante los datos se procesan mediante tareas de aprendizaje automático, aprendizaje profundo y modelos de series de tiempo. También Chu *et al.* [21] hacen uso de la *codificación one-hot encoding* para codificar un conjunto de datos con categorías, donde dichos datos representan pruebas de pozos petroleros. Posteriormente dichos datos son procesados por una técnica de clasificación basada en redes neuronales convolucionales.

Tabla 3.6: Conjunto de datos usando la codificación *one-hot encoding*.

	gen_1	gen_2	prof_1	prof_2	prof_3	prof_4	prof_5	Salario
1	1	0	1	0	0	0	0	20000
2	0	1	1	0	0	0	0	25500
3	1	0	0	1	0	0	0	18000
4	0	1	0	0	1	0	0	27500
5	1	0	0	0	0	1	0	15000
6	1	0	0	0	1	0	0	19000
7	0	1	0	0	0	0	1	16000
8	0	1	0	0	0	0	1	21500
9	0	1	0	0	1	0	0	27200
10	1	0	0	0	1	0	0	29300

- La *codificación binaria* es una representación de los valores por medio del sistema binario,

donde cada dígito de dicha representación es una columna [39]. Usando el conjunto de datos de la Tabla 3.5 se puede obtener una codificación *binaria* de los atributos categóricos, tal como se presenta en la Tabla 3.7. En el atributo de *género* se requieren las categorías en su forma nominal donde *Masculino* $= 1$ y *Femenino* $= 2$, en el que su equivalente en la representación binaria es 01 y 10, respectivamente. Como el atributo *género* requiere de dos dígitos para codificar sus valores nominales se crean dos atributos *gen_0* y *gen_1*, donde sus valores son binarios. Lo anterior también se puede observar con el atributo *profesión*. En la literatura existen trabajos que hacen uso de la codificación como Rodriguez *et al*. [73], quienes proponen un método de clasificación embebida que usa la *codificación binaria* para modelar los atributos categóricos de un conjunto de datos. Otro trabajo que hace uso de la codificación binaria es el propuesto por Strasser *et al*. [81], quienes usan dicha codificación para modelar problemas de optimización.

Tabla 3.7: Conjunto de datos usando la codificación *binaria*.

	gen_0	gen_1	prof_0	prof_1	prof_2	prof_3	Salario
1	0	1	0	0	0	1	20000
2	1	0	0	0	0	1	25500
3	0	1	0	0	1	0	18000
4	1	0	0	0	1	1	27500
5	0	1	0	1	0	0	15000
6	0	1	0	0	1	1	19000
7	1	0	0	1	0	1	16000
8	1	0	0	1	0	1	21500
9	1	0	0	0	1	1	27200
10	0	1	0	0	1	1	29300

- La *codificación baseN* es una representación de los valores por medio de un sistema base N, donde N equivale al número de los valores que requieren ser codificados [18]. Por ejemplo, la codificación base uno es equivalente a la codificación del *one-hot encoding* o la codificación base dos es equivalente a la codificación *binaria*. En la Tabla 3.8 se presenta un ejemplo de la codificación *baseN* donde $N = 4$ del conjunto de datos de la Tabla 3.5. Como se puede

observar en el atributo *profesión*, éste cuenta con un total de cinco valores diferentes, donde su representación en base 4 se divide en tres atributos de tal forma que $1 = 1$, $2 = 2$, $3 = 3$, $4 = 10$ y $5 = 11$.

Tabla 3.8: Conjunto de datos usando la codificación *baseN*.

	gen_0	gen_1	prof_0	prof_1	prof_2	Salario
1	0	1	0	0	1	20000
2	0	2	0	0	1	25500
3	0	1	0	0	2	18000
4	0	2	0	0	3	27500
5	0	1	0	1	0	15000
6	0	1	0	0	3	19000
7	0	2	0	1	1	16000
8	0	2	0	1	1	21500
9	0	2	0	0	3	27200
10	0	1	0	0	3	29300

- La *codificación hash* fue creada con el objetivo de reducir dimensiones, sin embargo ésta también puede ser utilizada para codificar atributos categóricos por medio de una función llamada hash [87]. En el contexto de la codificación, una función hash [65] permite hacer un mapeo de las categorías de un atributo a un número fijo de valores conocidos como valores hash. Existen diferentes funciones hash entre las más populares se encuentran la familia de algoritmos de hash seguro (SHA-0, SHA-1 y SHA-2) y el algoritmo de resumen del mensaje (MD5). El proceso de esta codificación es el siguiente: dado un conjunto de valores de un atributo, éstos pasan por una función hash, dando como resultado un valor hash, después este valor se le aplica una operación módulo N con el fin de reducir el rango de posibles valores hash. Con base en la función hash que se utilice será el rango de posibles valores hash que puede tener, pueden ser del orden de millones. Una de las ventajas de usar esta codificación es que no mantiene un diccionario de los valores, por lo que su dimensionalidad no crece en tamaño con la aparición de nuevas categorías. En la Tabla 3.9 se presenta un ejemplo de la codificación *hash* de los atributos *genéro* y *profesión* de la Tabla 3.5. Para codificar dichos

datos se utilizó como función hash y una operación módulo $N = 8$, por lo que se tienen 8 columnas. En la literatura hay pocos trabajos donde hacen uso de la *codificación hash* para codificar, sin embargo destaca el propuesto por Chapelle *et al.* [20], quienes proponen el uso de la *codificación hash* para codificar métricas de publicidad y modelar dichos datos por medio de una regresión lineal.

Tabla 3.9: Conjunto de datos usando la codificación *hash*.

	col_0	col_1	col_2	col_3	col_4	col_5	col_6	col_7
1	0	0	0	0	1	0	1	0
2	0	0	1	0	0	0	1	0
3	1	0	0	0	1	0	0	0
4	0	0	1	0	0	0	0	1
5	0	0	0	0	1	1	0	0
6	0	0	0	0	1	0	0	1
7	0	0	1	0	0	1	0	0
8	0	0	1	0	0	1	0	0
9	0	0	1	0	0	0	0	1
10	0	0	0	0	1	0	0	1

También existen otras codificaciones [18, 61] que son usadas con menor frecuencia, tales como:

- La *codificación target encoder*, donde una variable categórica puede ser codificada con respecto a una variable objetivo en su forma numérica. Para cada elemento de una variable categórica se agrupan por valores y se obtiene un promedio con respecto a una variable objetivo. Cada promedio reemplaza a su correspondiente valor. Un ejemplo del uso de la codificación del *target encoder* se puede observar en la Tabla 3.10, en el que se codifican los tres atributos *género*, *profesión* y *salario* de la Tabla 3.5. En este ejemplo se codifica el atributo *género* con respecto a la variable objetivo *salario*. Para ello se agrupan las variables categóricas y se obtiene un promedio con respecto al *salario*. Para el género *Femenino* se obtiene un promedio de 23,540, mismo que reemplaza a la variable categórica *Femenino*. El mismo procedimiento sucede para la categoría *Masculino*, donde se obtiene un promedio de 20,260. Ésto también se repite para

el atributo *profesión*. Cerda y Varoquaux [19] sugieren que la mejor codificación que permite modelar rasgos genéticos es por *target encoder*, misma que usan para hacer predicción de genomas por medio de modelos de regresión lineal.

Tabla 3.10: Conjunto de datos usando la codificación *target encoder*.

	Género	**Profesión**	**Salario**
1	20260	22750	20000
2	23540	22750	25500
3	20260	18000	18000
4	23540	25750	27500
5	20260	15000	15000
6	20260	25750	19000
7	23540	18750	16000
8	23540	18750	21500
9	23540	25750	27200
10	20260	25750	29300

- La *codificación helmert encoding*, también conocida como *reverse helmert encoding*, reemplaza las variables categóricas por la media de un valor menos la media de todas los valores anteriores. Dado que requiere que los valores sean ordenados, este método sugiere que sean usados para valores categóricos en una escala ordinal. Zavras *et al.* [88] realizaron un estudio de evaluación sobre cómo la crisis económica afectó las necesidades de atención médica en Grecia; hacen uso de la *codificación helmert encoding* para codificar el nivel educativo e ingreso de la población. Otra propuesta es la de Dillon y Tinsley [26], quienes hicieron un estudio sobre cómo eventos cercanos (como huracanes o sismos) y mensajes de advertencia pueden afectar la toma de decisión de la población. Para ello aplican regresión lineal para determinar la probabilidad de las decisiones de la población, donde los datos categóricos son codificados por la *codificación helmert encoding*.

- La *codificación polynomial* [18] permite realizar un análisis de los valores categóricos, de forma lineal, cuadrática y cúbica. Donde típicamente se utilizan valores ordinales ya que los datos requieren tener niveles marginados. Potdar *et al.* [66] realizaron un estudio comparativo de

las codificaciones de variables categóricas que permiten hacer uso de redes neuronales como clasificadores, donde hacen uso de la *codificación polynomial* y determinaron que la codificación permite obtener una precisión adecuada para su caso de estudio.

- La *codificación sum* [66] reemplaza las variables categóricas por la media de una variable dependiente de un valor menos la media de la variable dependiente para todos los valores. Esta codificación también fue usada un estudio comparativo realizado por los autores, donde determinaron que la *codificación sum* es de las mejores codificaciones ya que permite aumentar la precisión con respecto su caso de estudio.

- La *codificación backward difference* es una forma de codificar por atributo, donde se realiza un conteo de los valores y se obtiene un promedio por valor. Dicho promedio se resta contra otro promedio de un valor anterior. Un ejemplo de ello se presenta en la Tabla 3.11, en el que se codifican los atributos género y profesión de la Tabla 3.5. Por ejemplo, en el atributo *género* se tiene que para codificar el valor *Masculino* se resta la media de la categoría anterior menos la media de la categoría actual, dando como resultado -0.50. Lo mismo sucede con el valor *Femenino*, en este caso obteniendo un valor de 0.50. También esta codificación fue utilizada en el estudio comparativo realizado por Potdar *et al.* [66]. En dicho estudio determinaron que las codificaciones con mejores resultados son *backward difference* y *sum* contra las codificaciones como *one-hot encoding*, *helmert encoding*, *polynomial encoding* y *binaria*.

- Una codificación experimental es la *codificación polar* de Barcelo-Rico y Diez [13], la cual asigna a cada valor un par de coordenadas polares dentro del círculo unitario. De tal forma que a partir de dicha representación se puedan calcular equi-distancias entre valores.

El propósito de realizar una transformación de datos por medio de una codificación es para representar de forma numérica datos de tipo categórico, en donde pueden tener o no un orden (codificaciones como sum, polynomial y helmert encoding asumen un orden), no se pueden realizar

Tabla 3.11: Conjunto de datos usando la codificación *backward difference*.

	gen_0	prof_0	prof_1	prof_2	prof_3	Salario
1	-0.50	-0.80	-0.60	-0.40	-0.20	20000
2	0.50	-0.80	-0.60	-0.40	-0.20	25500
3	-0.50	0.20	-0.60	-0.40	-0.20	18000
4	0.50	0.20	0.40	-0.40	-0.20	27500
5	-0.50	0.20	0.40	0.60	-0.20	15000
6	-0.50	0.20	0.40	-0.40	-0.20	19000
7	0.50	0.20	0.40	0.60	0.80	16000
8	0.50	0.20	0.40	0.60	0.80	21500
9	0.50	0.20	0.40	-0.40	-0.20	27200
10	-0.50	0.20	0.40	-0.40	-0.20	29300

operaciones aritméticas y se busca un balance para mantener la menor dimensionalidad posible. Dicho lo anterior, una vez que los datos se encuentran en una representación numérica, éstos pueden ser procesados por técnicas de aprendizaje automático especializados en el espacio numérico (ver la Subsección 3.1.1). A continuación se discuten las limitantes de los códigos presentados con anterioridad.

3.3 Discusión

Las codificaciones que se presentaron pueden ser de utilidad para representar de forma numérica datos categóricos, manteniendo en algunas codificaciones las relaciones intrínsecas del espacio original de los datos. Sin embargo, éstas cuentan con las limitantes de afectar el desempeño de las tareas de aprendizaje o hacer que los resultados sean susceptibles a una mala interpretación o sesgo. Las limitantes que se observaron son las siguientes:

1. Aumento dimensional. La mayoría de las codificaciones del estado del arte producen un aumento dimensional del conjunto de datos de entrada. Esto depende del número de valores que se tienen por atributo categórico y el número de atributos categóricos. El aumento dimensional implica un incremento del número de atributos, permitiendo almacenar mayor información. Cuanto

mayor sea la información almacenada, es mayor el ruido y la redundancia en los datos. En el aprendizaje automático, el aumento dimensional de los datos da como resultado un mayor esparcimiento de los datos (la distancia entre instancias incrementa). A dicho efecto se le conoce como *maldición de la dimensionalidad* [8]. Por ejemplo, si se desean hacer grupos de instancias con atributos similares en un conjunto de datos que tiene una alta dimensionalidad, ésto se vuelve complicado dado que las instancias con propiedades similares se encontrarían esparcidas o separadas por una distancia que no permite la identificación de grupos. Para respaldar el análisis de un conjunto de datos que presenta la maldición de la dimensionalidad se puede aplicar inferencia estadística. Usualmente se requiere de una mayor cantidad de datos, lo cual incrementa de forma exponencial con respecto al número de dimensiones. Con respecto a lo anterior, las codificaciones como *one-hot encoding*, *backward difference*, *helmert encoding*, *polynomial* y *sum* presentan, por definición, un aumento dimensional alto al contar con una gran cantidad de valores únicos por columna. Esto se puede observar en la Tabla 3.12, donde se presenta el crecimiento dimensional al codificar con las codificaciones presentadas el conjunto de datos mixto de la Tabla 3.5. En la Tabla se puede observar que el conjunto de datos original cuenta con tres dimensiones, las codificaciones con mayor crecimiento dimensional son *helmert encoding*, *one-hot encoding*, *hash* (N=8) y *binaria*. Mientras que las codificaciones con menor crecimiento dimensional fueron *target encoding*, *polynomial* y *sum*. El uso de las codificaciones presentadas sólo se recomienda cuando se presentan pocos valores por atributo categórico. Sin embargo, las codificaciones como la *sum*, *polynomial* y *target encoder* mantienen una dimensionalidad reducida a pesar de la cantidad de valores únicos.

2. Mala interpretación de los datos. Según la codificación elegida, se asignan valores que típicamente denotan la ausencia o presencia de un valor. Las codificaciones que funcionan bajo este esquema son *one-hot encoding*, *baseN*, *binaria* y *hash*. Un problema presente en dichas codificaciones es que cuando se codifican los valores categóricos a una forma numérica se altera

Tabla 3.12: Crecimiento dimensional de las codificaciones del estado del arte.

Codificación	Dimensionalidad
One-hot encoding	8
Binaria	7
BaseN (N=4)	6
Hash (N=8)	8
Target Encoder	3
Backward Difference	6
Helmert encoding	9
Polynomial	5
Sum	5

el principio de las propiedades que presentan las categorías, de tal forma que ahora existe un orden, se pueden realizar operaciones aritméticas o se puede calcular un valor de tendencia central sobre ellos. Adicionalmente, algunas codificaciones en su proceso de codificación dependen de una variable numérica, llamada variable dependiente o variable objetivo, por lo que es posible que las codificaciones no mantengan las mismas relaciones intrínsecas de los valores, tal es el caso de las codificaciones *target encoder*, *helmert encoding*, *sum* y *backward difference*.

4

Método

Como se ha mencionado, los conjuntos de datos mixtos son aquellos donde conviven datos numéricos y no-numéricos (ver sección 2.1), representando características de algún fenómeno de interés, con el propósito de ser analizado y explicado a través de técnicas matemáticas y computacionales. La mayoría de dichas técnicas trabajan sobre el supuesto de que dichas características son estrictamente numéricas (e.g. Análisis de Regresión, Redes Neuronales), por lo que aquellas características no-numéricas del conjunto deben tener un tratamiento especial, que representa evidentemente un desafío. En este sentido, existen importantes métodos que intentan, a través de transformaciones, proveer a los datos no-numéricos de propiedades numéricas. Esto con el fin de hacerlos susceptibles de análisis en un espacio estrictamente numérico sobre el que es posible realizar operaciones aritméticas y establecer métricas de similitud. Como se indicó en la Sección 3.3, existen importantes desventajas y restricciones a la hora de recurrir a dichos métodos. La propuesta presentada en este trabajo no pretende una transformación numérica de aquellos datos que no lo son. En su lugar, propone una transformación de los datos numéricos del conjunto a un espacio en

el que convivan con los datos no-numéricos del mismo conjunto. El método muestra que en este espacio "unificado" es posible establecer operaciones y métricas que permiten realizar tareas propias del análisis de datos tales como agrupamiento y clasificación.

Para efectos de la presentación del método se adoptará la siguiente de notación:

- $\mathbb{S}$: Espacio de todos los valores posibles de una variable o propiedad del fenómeno de interés. Esto es cadenas de símbolos representando valores no-numéricos y valores numéricos en $\mathbb{R}$. Formalmente, $\mathbb{S} = \{\{0-9, a-z, A-Z\}^+ \cup \mathbb{R}\}$.

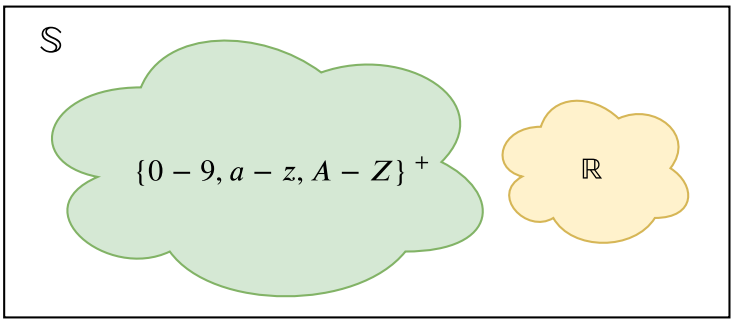


Figura 4.1: Espacio de un conjunto de datos mixto.

- $\hat{v} = (a_1, a_2, \ldots, a_m)$: Es una m-tupla o vector en $\mathbb{S}^m$ que contiene valores $a_i \in \mathbb{S}$ correspondientes a las propiedades o atributos del fenómeno de interés.

A partir de esta notación, se define un conjunto de datos mixto $\mathbb{X}$ como aquel conjunto conformado por N m-tuplas, expresado como:

$$\mathbb{X} = \begin{bmatrix} (a_{11}, a_{12}, \ldots, a_{1m}), \\ (a_{21}, a_{22}, \ldots, a_{2m}), \\ \vdots \\ (a_{N1}, a_{N2}, \ldots, a_{Nm}) \end{bmatrix}$$

Dado $\mathbb{X}$, la i-ésima propiedad del fenómeno de interés, estará conformada por los valores $a_{1i}, a_{2i}, ..., a_{Ni}$ (i-ésima columna de $\mathbb{X}$), los cuales serán del mismo tipo (numéricos o no-numéricos). Cuando dicho tipo es numérico (definido en $\mathbb{R}$), se hablará de una propiedad numérica, en otro caso, dicha propiedad será no-numérica. El siguiente ejemplo muestra una posible instancia de $\mathbb{X}$ con tres propiedades no-numéricas (*género*, *país* y nivel de Inglés) y una propiedad numérica (*estatura*).

$$\mathbb{X} = \begin{bmatrix} (\text{M}, \text{México}, \text{A1}, 1.68) \\ (\text{F}, \text{Colombia}, \text{B2}, 1.75) \\ \vdots \\ (\text{M}, \text{Brasil}, \text{C1}, 1.85) \end{bmatrix}$$

Se puede observar que las propiedades numéricas están definidas en $\mathbb{R}$, mientras las no-numéricas están definidas en $\{0-9, a-z, A-Z\}$ conforme a la definición de $\mathbb{S}$. Para efectos de la aplicación del método, se asume sin pérdida de generalidad, que las propiedades no-numéricas son previamente convertidas a códigos que representan valores nominales (no valores numéricos) los cuales corresponden a cadenas de símbolos únicamente sobre el conjunto $\{0-9\}$ preservando la definición de $\mathbb{S}$. Siguiendo con el ejemplo, $\mathbb{X}$ tendría posiblemente, los siguientes valores:

$$\mathbb{X} = \begin{bmatrix} (1, 5, 1, 1.68) \\ (0, 3, 2, 1.75) \\ \vdots \\ (1, 2, N, 1.85) \end{bmatrix}$$

Se requiere además que las propiedades estrictamente numéricas estén escaladas en el intervalo $[0, 1]$, por lo que finalmente, $\mathbb{X}$ será:

$$\mathbb{X} = \begin{bmatrix} (1, 5, 1, 0.45) \\ (0, 3, 3, 0.62) \\ \vdots \\ (1, 2, 5, 0.87) \end{bmatrix}$$

Como ya se mencionó, el método propuesto estará enfocado en transformar las propiedades numéricas de $\mathbb{X}$, de tal forma que puedan convivir con las propiedades no-numéricas en un mismo espacio, en el que sea posible definir algunas operaciones y relaciones de similitud. Este proceso de transformación será descrito en la siguiente sección.

4.1 Transformación de atributos numéricos

Para codificar los atributos numéricos se emplea un proceso de discretización. El rango de valores de cada atributo numérico (valores de la i-ésima columna en $\mathbb{X}$) es dividido en h intervalos denominados cuantiles. Cada cuantil contendrá una proporción de los valores del atributo o variable, tal como se ilustra la Figura 4.2.

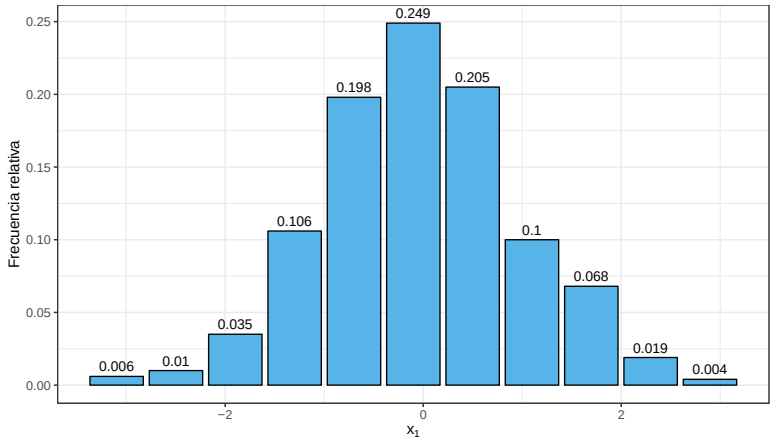


Figura 4.2: Ejemplo de discretización de una variable numérica en el intervalo $[-3.4, 3.2]$.

El número de cuantiles h puede establecerse de manera discrecional (no recomendable) o usando algunos aproximadores la como *regla de Sturges* [82], *Rice* [53] o *Doane* [27]. En el caso de la regla de Sturges, el valor de h está definido como:

$$h = \lceil 1 + \log_2 N \rceil \tag{4.1}$$

donde N es el número de $m-$tuplas en $\mathbb{X}$, el cual deberá coincidir con el número de valores del atributo numérico que se va a discretizar. Por lo tanto, para todos los atributos numéricos en $\mathbb{X}$ a ser discretizados, el valor de h es el mismo. Por ejemplo, para un conjunto de datos $\mathbb{X}$ de 1,000 m-tuplas, todos los atributos numéricos serán divididos en 11 cuantiles.

Definición 4.1.1. Sea $\mathbb{X}[, i]$ el conjunto de valores o instancias del i-ésimo atributo numérico en $\mathbb{X}$. Un cuantil q_k es un intervalo $\left[\underline{a_k}, \overline{a_k}\right]$ de tamaño δ en el espacio de $\mathbb{X}[, i]$ donde, $\underline{a_k}$ y $\overline{a_k}$ son los límites inferior y superior de q_k y δ está dado por:

$$\delta = \frac{|\operatorname{máx}(\mathbb{X}[, i]) - \operatorname{mín}(\mathbb{X}[, i])|}{h} \tag{4.2}$$

El primer cuantil del i-ésimo atributo numérico $\mathbb{X}[, i]$ está dado por:

$$q_1 = [\operatorname{mín}(\mathbb{X}[, i]),\ \operatorname{mín}(\mathbb{X}[, i]) + \delta) \tag{4.3}$$

mientras que los cuantiles subsecuentes están dados por:

$$q_k = \begin{cases} [\overline{a}_{k-1},\ \overline{a}_{k-1} + \delta] \text{ si } k = h, \\ [\overline{a}_{k-1},\ \overline{a}_{k-1} + \delta) \text{ en otro caso.} \end{cases} \tag{4.4}$$

Para efectos ilustrativos, asúmase que $\mathbb{X}[, i]$ es una variable numérica con 1,000 instancias definida en el intervalo $[-3.40, 3.20]$, con $h = 11$ (de acuerdo a la regla de Sturges). De la Ecuación 4.2 se tiene que $\delta = 0.6$. Usando la Ecuación 4.3 y 4.4, se establecen los cuantiles de $\mathbb{X}[, i]$ como se

muestran en la Tabla 4.1.

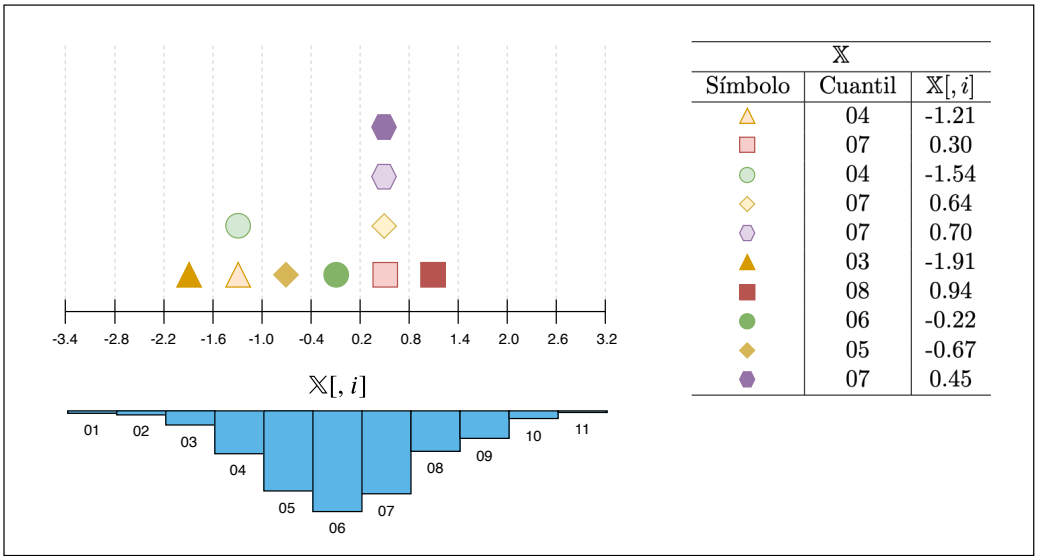
Tabla 4.1: Codificación de los cuantiles en $\mathbb{X}[,i]$.

$\mathbb{X}[,i]$		
Cuantil	Límite inferior	Límite superior
01	[-3.40	-2.80)
02	[-2.80	-2.20)
03	[-2.20	-1.60)
04	[-1.60	-1.00)
05	[-1.00	-0.40)
06	[-0.40	0.20)
07	[0.20	0.80)
08	[0.80	1.40)
09	[1.40	2.00)
10	[2.00	2.60)
11	[2.60	3.20]

Cada cuantil es identificado a través de un número secuencial (primera columna de la Tabla 4.1) que corresponderá en adelante, al *código de cuantil*. Entonces $\forall x_i \in \mathbb{X}[,i]$ será posible asignar el código del cuantil q_k cuyo intervalo $\left[\underline{a_k}, \overline{a_k}\right]$ puede contener el valor x_i. Por ejemplo, supóngase que $\{-1.21, 0.30, -1.54, 0.64, 0.70, -1.91, 0.94, -0.22, -0.67, 0.45\} \subset \mathbb{X}[,i]$. Estas instancias o valores pueden ser mapeados al espacio de los cuantiles con base en los intervalos que los definen. En la Figura 4.3 se ilustra el resultado de este proceso. De manera complementaria se incluye, a manera de histograma, la proporción de elementos contenidos en cada cuantil. En el ejemplo dichas proporciones obedecen a una distribución normal, aunque no siempre va ser así.

4.2 Dataset como conjunto de m-tuplas

A partir del proceso de transformación descrito, toda variable numérica en $\mathbb{X}$ es mapeada a un espacio conformado por los códigos de cuantil. Por su parte, las variables no-numéricas conservan su valor categórico o nominal, asumiendo que en este punto del proceso, dichas variables solo pueden tomar caracteres que satisfacen la expresión regular $\{0-9\}^+$.



$\mathbb{X}$		
Símbolo	Cuantil	$\mathbb{X}[, i]$
△	04	-1.21
□	07	0.30
○	04	-1.54
◇	07	0.64
⬡	07	0.70
▲	03	-1.91
■	08	0.94
●	06	-0.22
◆	05	-0.67
⬢	07	0.45

Figura 4.3: Codificación valores numéricos.

Para ilustrar lo anterior, sea $\mathbb{X}$ un conjunto de datos con las variables de altura, edad, peso, sexo y estado. Por su naturaleza numérica, la *altura*, *edad* y *peso* serán transformadas conforme al proceso descrito en la Sección 4.1, mientras que, las variables de *sexo* y *estado* conservaran su codificación original. Esto se muestra en la Figura 4.4.

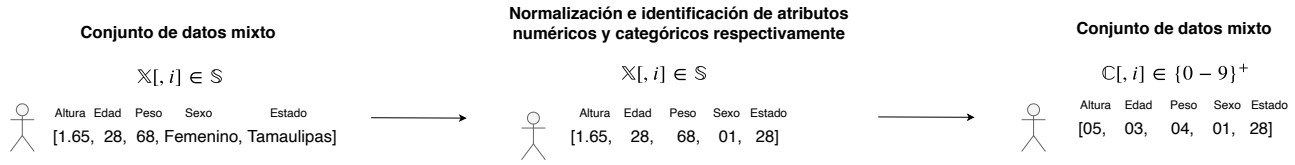


Figura 4.4: Ejemplo del método de transformación propuesta para una instancia en $\mathbb{X}$.

El conjunto de datos resultante estará conformado por m-tuplas con caracteres definidos exclusivamente en el conjunto $\mathbb{C} = \{0-9\}^+$, tal como se muestra en la Figura 4.5, donde los valores *femenino* y *masculino* se representan con 01 y 02 respectivamente. Los nombres de los estados han sido sustituidos por su número en el listado de todos los 32 estados de la República Mexicana (*Tamaulipas* $= 18$, *Veracruz* $= 30$, *Oaxaca* $= 20$).

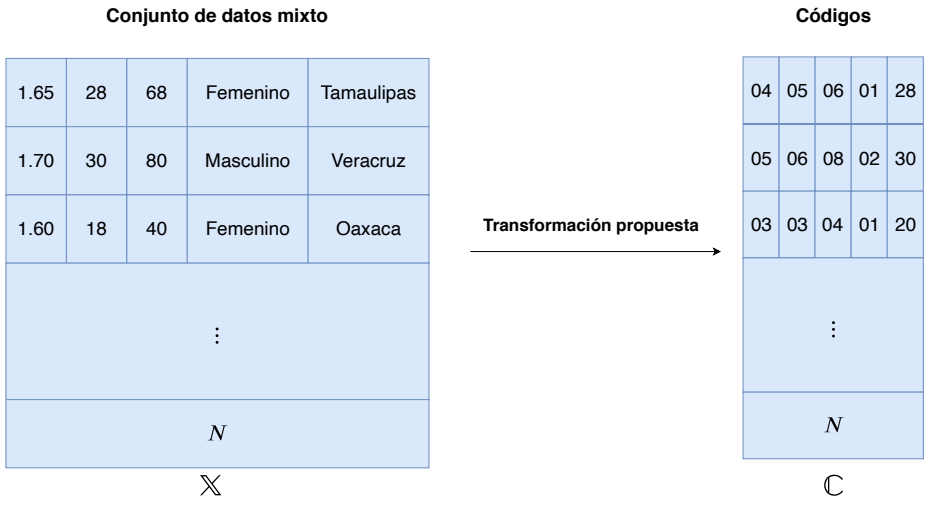


Figura 4.5: Transformación propuesta de los valores en $\mathbb{X}$.

4.3 Acerca del espacio de las m-tuplas

Hasta este punto, el conjunto de datos $\mathbb{X}$ se ha transformado en un conjunto de m-tuplas de la forma $\hat{v} = (a_1, a_2, ..., a_m)$ donde cada $a_i \in \mathbb{C}$. Dicho conjunto induce un espacio $\mathbb{C}^m$, denominado en adelante *espacio categórico* sobre el cual se enuncian las siguientes consideraciones:

1. Para toda m-tupla en $\mathbb{C}^m$, los valores a_i corresponden a códigos de naturaleza categórica o nominal sobre los cuales no es posible realizar operaciones aritméticas. En consecuencia, estadísticos como el valor medio o la desviación de un conjunto de m-tuplas carecen de sentido.

2. Debido a la mencionada naturaleza categórica, la similitud entre dos m-tuplas $\hat{u}$ y $\hat{v}$, no puede ser definida en términos de una medida de proximidad numérica tal como la distancia euclidiana.

Las medida de similitud es un elemento importante en tareas de clasificación y agrupamiento. Por lo que es de vital importancia definir una medida que permita definir la similitud entre las tuplas mencionadas. Al respecto, asúmase que $\mathbb{X}$ está definido en $\mathbb{C}^2$ con las siguientes 2-tuplas.

$$\mathbb{X} = \{[3,3],[2,2],[1,2],[1,1],[2,1],[3,1],[1,3],[2,3],[3,2]\} \tag{4.5}$$

Aunque por definición estas tuplas no tienen ningún orden, para efectos ilustrativos se procede a ordenarlas lexicográficamente (lo cual no es un orden numérico, estrictamente hablando), en cuyo caso:

$$\mathbb{X} = \{[1,1],[1,2],[1,3],[2,1],[2,2],[2,3],[3,1],[3,2],[3,3]\} \tag{4.6}$$

Para efectos ilustrativos, estas tuplas pueden ser representadas matricialmente como en la Figura 4.6.

11	12	13
21	22	23
31	32	33

Figura 4.6: Tuplas en $\mathbb{C}^2$

Para efectos del análisis, supóngase que se desea determinar las tuplas más cercanas a la tupla $\hat{v} = [1,1]$. Para esto, es necesario hacer uso de alguna métrica o medida de distancia. En la Figura 4.7 se presentan los elementos más cercanos (resaltadas con verde) a $\hat{v}$ (tupla $[1,1]$, resaltado con rojo) con base en tres métricas: 1) Euclidiana, 2) Chebyshev y 3) Hamming. Dado que la distancia

1) Euclidiana

11	12	13
21	22	23
31	32	33

2) Chebyshev

11	12	13
21	22	23
31	32	33

3) Hamming

11	12	13
21	22	23
31	32	33

Figura 4.7: Distancias entre la tupla $[1,1]$ y sus vecinos más próximos en $\mathbb{C}^2$.

Euclidiana representa un proximidad numérica, dicha proximidad carece de sentido en el contexto de

las tuplas definidas en el espacio categórico $\mathbb{C}^m$. En la siguiente sección se muestra que la distancia de Chebyshev resulta conveniente para realizar tareas de agrupamiento en el espacio $\mathbb{C}^m$, ya que dicha tarea es equivalente al agrupamiento en el espacio Euclidiano en el que se manejan típicamente los datos mixtos (después de tareas de procesamiento de valores categóricos).

Agrupamiento de m-tuplas

Para contar con un método de agrupamiento base se implementó una modificación del método k-*modes* usando la distancia de *Chebyshev*, en lugar de la distancia de *Hamming* (la que emplea el algoritmo base k-modes). Dicho método se denota más adelante como k-*modes-modificado*. La distancia de *Chebyshev* permite mantener una equidistancia entre los códigos. Por ejemplo, asumiendo la distancia de Chebyshev de los datos representados en la Figura 4.7, al tomar el código $[1, 1]$, éste tendría una equidistancia de 1 entre los códigos $[2, 1]$, $[2, 2]$ y $[1, 2]$. Para ello, en la Subsección 5.2 se presenta el algoritmo e ilustra su funcionamiento.

Adicionalmente al usar la distancia de *Chebyshev* es posible, para ciertos casos, obtener agrupaciones similares que al usar la distancia de *Hamming*. A continuación en la Figura 5.2 se muestra una comparativa de agrupamientos al usar ambas distancias. Para este ejemplo las agrupaciones son similares, por lo que para evaluar el desempeño de ambas métricas en tareas de agrupamiento, en el Capítulo 5 se propone una base experimental que evalúa bajo qué circunstancias la naturaleza de los datos es mejor que otra.

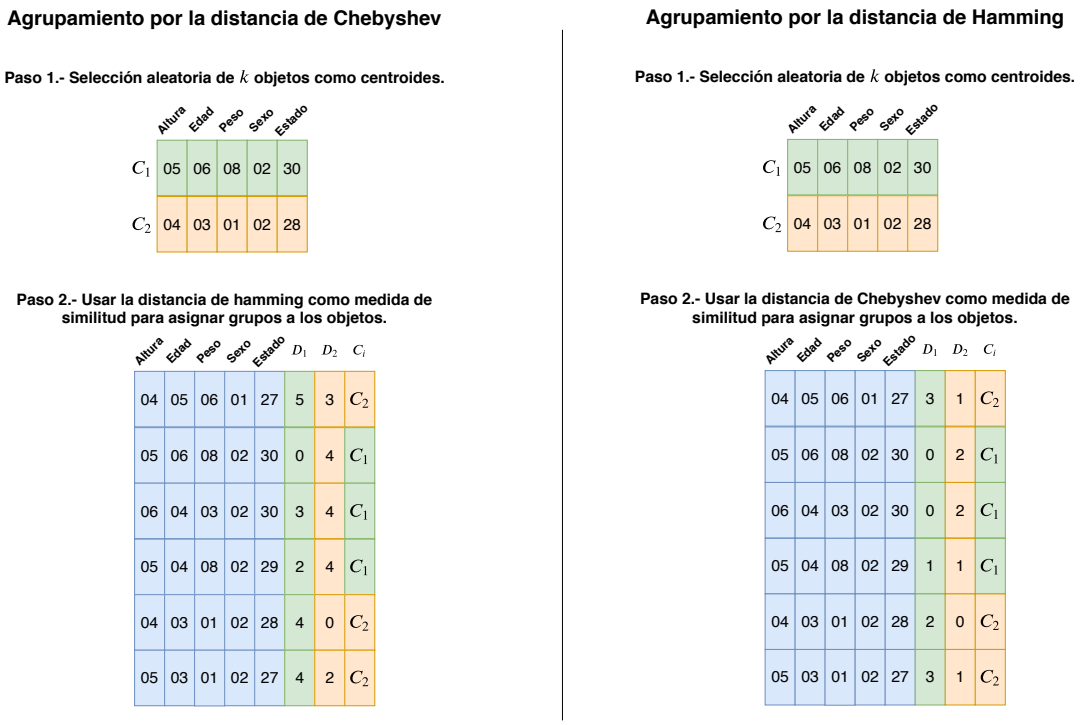


Figura 4.8: Comparativa de agrupamiento por distancia.

5

Experimentación y Resultados

En este capítulo se presenta la experimentación y resultados obtenidos con la implementación del método propuesto utilizando un repositorio de conjuntos de datos etiquetados (marca de clase conocida a priori). Dichos conjuntos de datos han sido clasificados de acuerdo al tipo de dato que contienen. Se distinguen tres grupos: categóricos, numéricos y mixtos. Para demostrar la aplicabilidad y viabilidad del método propuesto, éste se aplica en un caso de estudio de agrupación de datos. Se consideró la tarea de agrupamiento dado que su objetivo es buscar similitudes por medio de las propiedades de los objetos, por lo que las relaciones intrínsecas deberían ser mantenidas al codificar los datos por el método propuesto. En la Figura 5.1 se presenta el proceso de agrupamiento de un conjunto de datos mixto, donde un conjunto de datos es transformado con el método propuesto resultando en un conjunto de códigos $\mathbb{C}$.

En la experimentación los conjuntos de datos se procesan con métodos de agrupación como *coolcat* [11], *rock* [38], k-*modes* [42], k-modes-modificado y k-*means*. Los conjuntos de datos se transformaron dependiendo del método de agrupación aplicado. Esto con el objetivo de comparar

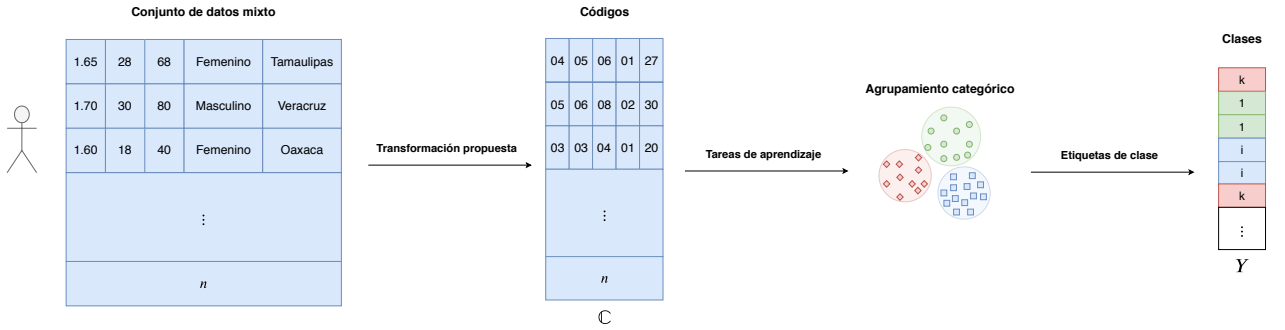


Figura 5.1: Método de agrupamiento para la codificación.

el desempeño de los métodos de agrupación. En esta experimentación no se intentó mejorar el desempeño de algún método de agrupación, sino demostrar que la codificación es útil y aplicable. Cada conjunto de datos se presentó a cada método de agrupamiento categórico después de realizar la transformación propuesta.

A manera de comparar el desempeño de los métodos de agrupación categóricos contra un método de agrupación numérica, los mismos conjuntos de datos fueron procesados mediante el método k-*means*, realizando previamente una transformación de los datos nominales mediante *one-hot encoding* (Figura 2.2).

A partir de los resultados obtenidos por los métodos de agrupamiento se calculó el índice ARI (ver la Subsección 5.3.5) relativo a la marca de clase de cada instancia del conjunto de datos con el fin de determinar una medida de desempeño del método de agrupamiento y, en consecuencia, determinar también el desempeño de la transformación propuesta.

5.1 Conjuntos de datos

Los conjuntos de datos que se utilizaron en la experimentación se muestran en la Tabla 5.1, donde se presentan por su nombre, tipo de dato, número de atributos numéricos y categóricos, total de atributos, total de instancias y número de clases conocidas a priori. Dichos conjuntos de datos fueron obtenidos del repositorio UCI [57].

Tabla 5.1: Conjuntos de datos mixtos, categóricos y numéricos empleados en la experimentación.

Nombre	Tipo	Atributos numéricos	Atributos categóricos	Total atributos	Total instancias	Clases
Wines	Numérico	13	0	13	178	3
Breast Cancer Wisconsin	Numérico	9	0	9	683	2
New Thyroid	Numérico	5	0	5	215	3
Lung Cancer	Numérico	56	0	56	32	3
Glass	Numérico	9	0	9	214	6
Soybean	Categórico	0	35	35	47	4
Zoo	Categórico	1	16	17	101	7
Dermatology	Categórico	1	33	34	366	6
Vote	Categórico	0	16	16	435	2
Mushroom	Categórico	0	22	22	8124	2
Autos	Mixto	15	10	25	205	7
Hepatitis	Mixto	6	13	19	155	2
Heart Statlog	Mixto	6	7	13	270	2
Heart Cleveland	Mixto	5	8	13	303	5
German Credit Data	Mixto	3	17	20	1000	2
Australian Credit Data	Mixto	6	8	13	690	2

A continuación se describen de los conjuntos de datos de tipo numérico y el preprocesamiento realizado.

- **Wines** es un conjunto de datos de tipo numérico con 13 atributos y 178 instancias que representan los resultados de un análisis químico de vinos provenientes de una misma región en Italia y que pertenecen a tres diferentes cultivos. Este conjunto de datos cuenta con tres etiquetas de clase diferentes, correspondientes a uno de los tres cultivos de vino. Las etiquetas de clase se dividen en 59, 71 y 48 instancias correspondientemente para la clase 1 (33,1 %), 2 (39,8 %) y 3 (26,9 %).

- **Breast Cancer Wisconsin** es un conjunto de datos de tipo numérico con 9 atributos y 699 instancias que representan los resultados de una prueba para detectar el cáncer de mama del Centro Médico Universitario del Instituto de Oncología de Ljubljana, Yugoslavia. Los atributos describen la presencia en una imagen de núcleos celulares y sus etiquetas de clase que anuncian si en la imagen se presenta tumor es maligno o benigno. En el preprocesamiento de este conjunto de datos se removieron 16 instancias que contenían valores incompletos, dejando un total de 683 instancias. Las etiquetas de clase se dividen en 446 instancias (65,4 %) como benigno y 237 instancias (34,6 %) como maligno.

- **New Thyroid** es un conjunto de datos de tipo numérico con 5 atributos y 215 instancias que representan una prueba de laboratorio para identificar si un paciente tiene hipertiroidismo, hipotiroidismo o se encuentra normal. El conjunto de datos proviene del Instituto Garavan de Sídney, Australia. Las etiquetas de clase se encuentran distribuidas en 150 instancias como normal (69,76 %), 35 instancias como hipertiroidismo (16,27 %) y 30 instancias como hipotiroidismo (13,95 %).

- **Lung Cancer** es un conjunto de datos de tipo numérico con 56 atributos y 32 instancias. No hay información sobre el significado de los atributos (por razones de privacidad), sin embargo cada instancia tiene asociada una etiqueta de clase que determina uno de tres tipos de cáncer de pulmón. En la fuente del conjunto de datos no se especifica el orden de las etiquetas de clase y se encuentran distribuidas en 9 (28,12 %), 13 (40,62 %) y 10 (31,25 %) instancias.

- **Glass** es un conjunto de datos de tipo numérico con 9 atributos y 214 instancias que representan el análisis de laboratorio de fragmentos de vidrio, esto para identificar el tipo de vidrio por medio de la presencia de ciertos elementos (en términos de porcentaje). Este conjunto de datos tiene como motivación el caso de estudio de la investigación forense. Dicho conjunto de datos proviene del Servicio de Ciencia Forense de Estados Unidos. En el preprocesamiento se removió el atributo *Id* (identificador de cada instancia), ya que es irrelevante para las tareas de análisis de datos. Las etiquetas de clase se encuentran distribuidas en 70 (32,71 %), 76 (35,51 %), 17 (7,94 %), 13 (6,07 %), 9 (4,20 %) y 29 (13,55 %) instancias.

A continuación se describen los conjuntos de datos de tipo categórico y el procesamiento realizado.

- **Soybean** es un conjunto de datos de tipo categórico con 35 atributos y 47 instancias, donde cada instancia está etiquetada con uno de cuatro tipos de plagas presentes en cultivos de soya: *Diaporthe Stem Canker* (10 instancias, 21,27 %), *Charcoal Rot* (10 instancias, 21,27 %), *Rhizoctonia Root Rot* (10 instancias, 21,27 %) y *Phytophthora Rot* (17 instancias, 36,17 %).

- **Zoo** es un conjunto de datos mixto pero es considerado como un conjunto de datos categórico dado que solamente se cuenta con un atributo numérico, mismo que se descartó, dejando un total de 16 atributos categóricos y 101 instancias. Cada instancia representa un animal y la presencia de ciertas características permiten identificar a qué subconjunto de animales pertenece. Se cuentan con 7 subconjuntos de animales (etiquetas de clase): 41 (40,59 %), 2 (19,80 %), 3 (4,95 %), 4 (12,87 %), 5 (3,96 %), 6 (7,92 %) y 7 (9,90 %) instancias.

- **Dermatology** es también un conjunto de datos mixto pero dado que cuenta con un atributo numérico, éste fue descartado para ser considerado como un conjunto de datos categórico. Este conjunto de datos cuenta con 33 atributos, de los cuales se descartarán 6 por contener valores constantes o altamente correlacionados y 366 instancias donde cada instancia describe un historial patológico de un paciente con respecto a tipos de enfermedades *eritemato-escamosas* (término médico usado en dermatología). Las etiquetas de clase se distribuyen en 6 clases: *psoriasis* (112 instancias, 30,60 %), *seboreic dermatitis* (61 instancias, 16,66 %), *lichen planus* (72 instancias, 19,67 %), *pityriasis rosea* (49 instancias, 13,38 %), *chronic dermatitis* (52 instancias, 14,20 %) y *pityriasis rubra* (20 instancias, 5,46 %).

- **Vote** es un conjunto de datos de tipo categórico con 16 atributos y 435 instancias, de las cuales se descartarán 203 instancias por contener valores incompletos. Este conjunto de datos contiene votos del congreso de Estados Unidos de 1984 y se encuentran etiquetados por dos partidos políticos: republicanos (144 instancias, 62,06 %) y demócratas (88 instancias, 37,93 %).

- **Mushroom** es un conjunto de datos de tipo categórico con 22 atributos y 8,124 instancias, en el que cada instancia describe las características de un hongo. Cada instancia cuenta con una etiqueta de clase sobre si es comestible o venenoso. Las etiquetas de clase son dos: comestible con 4,208 instancias (51,79 %) y venenoso con 3,916 instancias (48,20 %).

Por último se presenta la descripción de los conjuntos de datos de tipo mixto y el preprocesamiento realizado.

- **Autos** es un conjunto de datos de tipo mixto con 15 atributos numéricos, 10 atributos categóricos y 193 instancias (después de eliminar 12 instancias por contener valores incompletos). Este conjunto de datos por cada instancia describe algunas características de un automóvil, se encuentra etiquetada con una clase que describe qué tan riesgoso es para una aseguradora dicho automóvil. Las etiquetas de clase se distribuyen en 7: 3 con 3 instancias (1,55 %), 2 con 22 instancias (11,39 %), 1 con 63 instancias (32,64 %), 0 con 51 instancias (26,42 %), -1 con 31 instancias (16,06 %) y -2 con 23 instancias (11,91 %).

- **Hepatitis** es un conjunto de datos mixto con 6 atributos numéricos, 13 atributos categóricos y 80 instancias (después de eliminar 75 instancias con variables incompletas). Este conjunto de datos representa características de un paciente, se encuentra etiquetada con dos clases: die (13 instancias, 16,25 %) y live (67 instancias, 83,75 %).

- **Heart Statlog** es un conjunto de datos mixto con 6 atributos numéricos y 7 atributos categóricos con 270 instancias que representan características de un paciente con problemas del corazón. Dicho conjunto de datos se encuentra etiquetado con la presencia (150 instancias, 55,55 %) o ausencia (120 instancias, 44,44 %) de un problema del corazón.

- **Heart Cleveland** es un conjunto de datos mixto de una Clínica en Cleveland (Estados Unidos) que contiene 5 instancias numéricas y 8 instancias categóricas, misma que cuenta con 297 instancias (de las cuales se descartaron 6 instancias por contener valores incompletos). Este conjunto de datos describe problemas del corazón de un paciente. Cada instancia se encuentra etiquetada con una clase en escala del 0 al 4, donde 0 es la ausencia y 4 es la presencia de un problema en el corazón. Las etiquetas de clase se encuentran distribuidas como 0 con 163 instancias (54,88 %), 1 con 52 instancias (17,50 %), 2 con 35 instancias (11,78 %), 3 con 34 instancias (11,44 %) y 4 con 13 instancias (4,37 %).

- **German Credit Data** es un conjunto de datos mixto con 3 atributos numéricos y 17 atributos

categóricos, cuenta con 1,000 instancias que describen características personas que solicitan un crédito. Por razones de privacidad no se indica el significado de cada atributo. Cada instancia es etiquetada con una clase: bueno (700 instancias, 70 %) y malo (300 instancias, 30 %).

- **Australian Credit Data** es un conjunto de datos mixtos con 6 atributos numéricos y 8 atributos categóricos con un total de 690 instancias. Cada instancia caracteriza la aprobación (383 instancias, 55,50 %) o rechazo (307 instancias, 44,49 %) del crédito que se le otorga a una persona. Por motivos de privacidad no se proporciona más contexto sobre la información de los atributos.

5.2 Implementación

Los métodos de agrupamiento de datos categóricos que se consideraron para esta experimentación son *rock* [38], *coolcat* [11], k-*means* [56], k-*modes* [42] y una modificación del k-*modes* que se explica más adelante 5.2. La implementación de los métodos de agrupamiento se realizó en el lenguaje de programación R, dado que la mayoría de los métodos se encuentran disponibles en librerías implementadas por los mismos autores o se encuentran corroboradas por una comunidad de software libre. La forma en que están implementados los métodos se describe a continuación:

- *Rock*: Implementado en la función rockCluster del paquete cba [17] en R.
- *Coolcat*: Implementado en la función coolcat del paquete coolcat [12] en R.
- k-*means*: Implementado en la función kmeans, mismo que se encuentra definido dentro de las librerías básicas de R.
- k-*modes*: Implementado en la función kmodes del paquete klaR [86] en R.

Modificación del método de agrupamiento k-*modes*

Para contar con un método de agrupamiento base se implementó una modificación del método k-*modes* usando la distancia de *Chebyshev*, en lugar de la distancia de *Hamming*. Dicho método se denota más adelante como k-*modes-modificado*. La distancia de *Chebyshev* permite mantener una equidistancia entre los códigos. La distancia de *Chebyshev* se define formalmente como indica la Ecuación 5.1.

$$d(p, q) = \max_i \left(|p_i - q_i| \right), \tag{5.1}$$

donde p y q son dos vectores.

En la Figura 5.2 se ilustra el funcionamiento del método k-*modes-modificado*, el cual se describe a continuación:

1. Dado un conjunto de códigos C, se seleccionan k objetos como representantes de los grupos denominados como *centroides*.

2. Se calculan las distancias de los objetos hacia los k centroides empleando la distancia de *Chebyshev*. El objeto con la distancia mínima a un centroide es asignado al grupo correspondiente.

3. Por cada grupo de objetos se calculan los nuevos centroides. Se calcula la moda de cada grupo, mismos que son definidos como los nuevos centroides.

4. Iterar los pasos 2 y 3 hasta k número de veces.

En el método el número k iteraciones es parametizable, por lo que dicho valor puede ser elegido a discreción.

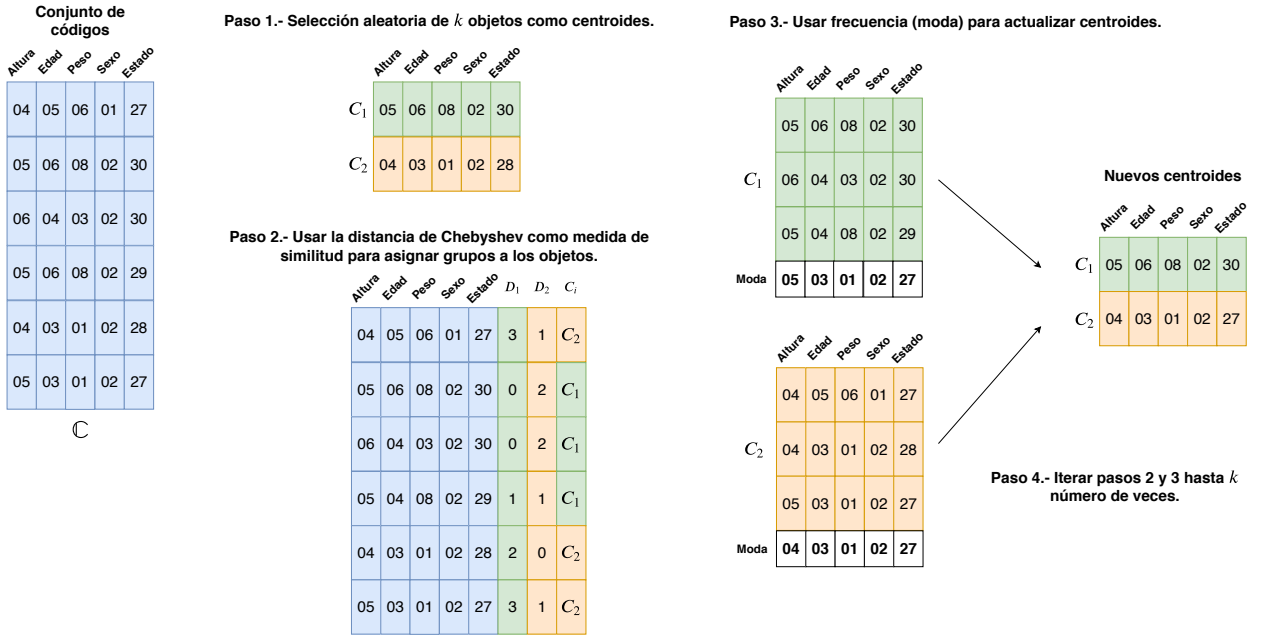


Figura 5.2: Modificación del *k-modes*, reemplazando la distancia de *Hamming* por *Chebyshev*

5.3 Diseño experimental

Para conocer el desempeño generalizado de la implementación se creó un ambiente de experimentación, el cual se explica a continuación.

5.3.1 Preprocesamiento de los conjuntos de datos

El proceso experimental se describe a continuación por tipo de dato de los conjuntos de datos que se describieron con anterioridad. Asimismo se describe su preprocesamiento (tal como se ilustra en la Figura 5.3).

- Conjuntos de datos categórico

    1. Preprocesamiento *one-hot encoding*: Se codificaron los atributos categóricos mediante el método *one-hot encoding*. Se obtuvo una nueva representación en un espacio numérico, donde luego fueron procesados por el método de agrupamiento numérico k-*means* [56].

2. Preprocesamiento transformación propuesta: El conjunto de datos fue procesado directamente por los métodos de agrupamiento categórico *coolcat* [11], *rock* [38], k-*modes* [42] y k-*modes-modificado*.

- Conjuntos de datos mixto

1. Preprocesamiento *one-hot encoding*: Se codificaron los atributos categóricos mediante el método *one-hot encoding*, los atributos numéricos se mantuvieron tal cual. Como resultado se obtuvo una representación en el espacio numérico, la cual fue procesada por el método de agrupamiento numérico k-*means* [56].
2. Preprocesamiento transformación propuesta: Con la transformación propuesta se transformaron los atributos numéricos y los atributos categóricos se pasaron a su forma nominal. El resultado fue una representación de códigos en un espacio categórico en escala nominal, el cual fue procesado por los métodos de agrupamiento categórico *coolcat* [11], *rock* [38], k-*modes* [42] y k-*modes-modificado*.

- Conjuntos de datos numérico

1. Preprocesamiento *one-hot encoding*: El conjunto de datos fue procesado directamente por el método de agrupamiento numérico k-*means* [56].
2. Preprocesamiento transformación propuesta: Con la transformación propuesta se transformaron los atributos numéricos. El resultado fue una representación de códigos en un espacio categórico en escala nominal, mismo que fue procesado por los métodos de agrupamiento categórico: *coolcat* [11], *rock* [38], k-*modes* [42] y k-*modes-modificado*.

5.3.2 Normalidad en la experimentación

Cada método con cada conjunto de datos se ejecutó 100 veces de acuerdo al Teorema del Límite Central [85]. Para obtener un comportamiento generalizado del método propuesto fue necesario

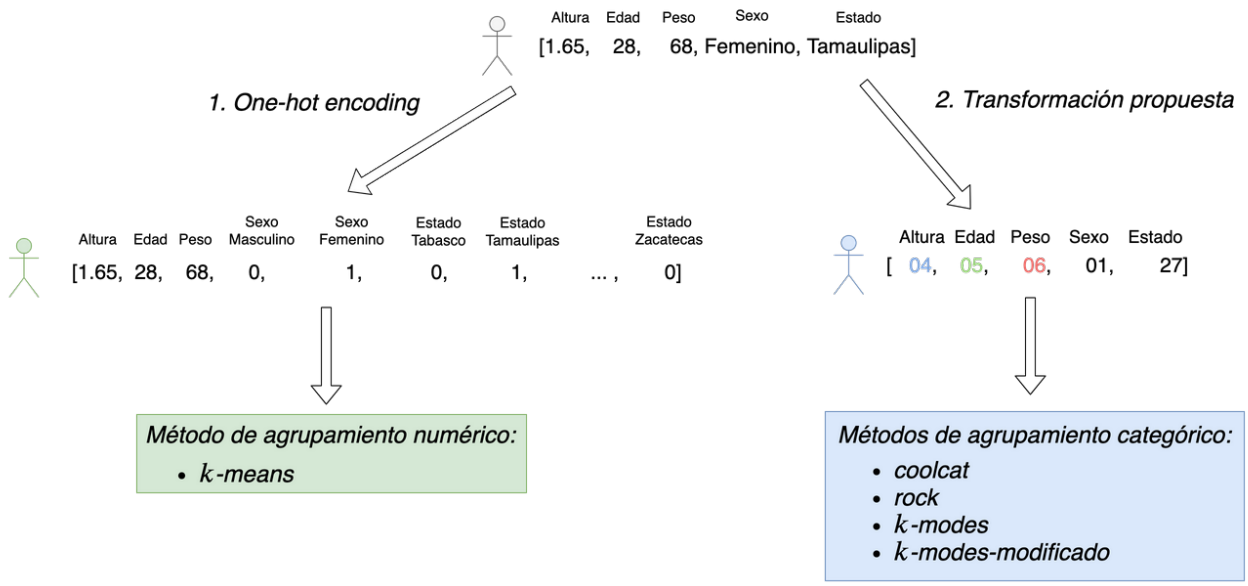


Figura 5.3: Proceso de la experimentación.

contar con una muestra de un número adecuado de elementos para experimentar. El Teorema del Límite Central asegura que conforme el número de elementos se vuelve más grande, el resultado se aproximará una distribución normal con una media μ y una desviación estándar σ. Por lo que con dicho teorema se asegura que el resultado del experimento no fue obtenido a partir de efectos aleatorios, sino por efectos probabilísticos. En la práctica, se aconseja un número de experimentos mayor o igual a 30, lo cual es suficiente para obtener una aproximación a una distribución normal [41]. Cuando el número de experimentos tiene una suficiente aproximación, la muestra promedio sigue aproximadamente una distribución normal con una media μ' y una desviación estándar de la muestra promedio, de la forma como indica la Ecuación 5.2.

$$SE = \sigma' = \frac{\sigma}{\sqrt{n}} \tag{5.2}$$

donde n es el número de experimentos y σ es la desviación estándar de la muestra.

5.3.3 Pruebas de hipótesis de diferencia de medias

A los resultados del método de agrupamiento categórico que mejor se desempeñó se le realizó una prueba de hipótesis de diferencia de medias contra el resultado del método k-*means*. Lo anterior permitió identificar cuál de ambos métodos se desempeñó mejor, con cuáles tipos de conjuntos de datos lo hizo y si la diferencia entre ambos resultados es estadísticamente significativa (que su resultado no se deba a efectos aleatorios). El nivel de significancia que se consideró para esta experimentación fue de un $\alpha = 0.05$.

5.3.4 Infraestructura empleada

Las características del equipo de cómputo donde se realizó la experimentación es la siguiente:

- Procesador 2.6 GHz Intel Core i5.
- Memoria 8GB 1600 MHz DDR3.
- Sistema operativo macOS Mojave versión 10.14.

Las herramientas de software utilizadas para la experimentación fueron:

- RStudio versión 1.1.383 con R versión 3.5.3.
- MySQLServer versión 14.14 distribución 5.7.21.

5.3.5 Métricas de evaluación

Las métricas de evaluación que se utilizaron en esta experimentación fueron el crecimiento dimensional y el ARI. El crecimiento dimensional, para el contexto de este trabajo, es una comparativa de cómo las codificaciones utilizadas aumentan el número de dimensiones de un conjunto de datos con respecto al número de dimensiones que tenían originalmente.

Para contextualizar qué es la métrica ARI, primero se explica la base de dicha métrica, la cual se basa en la métrica *Rand Index* [71].

Dado dos agrupamientos de un conjunto de datos con etiquetas de clase, a éstos se le pueden medir su exactitud por medio de la métrica *Rand Index* (RI por sus siglas en inglés). RI es una medida que permite calcular la exactitud de un método de agrupamiento con respecto a sus etiquetas de clase originales. Dicha métrica recibe como entrada dos grupos de etiquetas: las etiquetas de clase originales (C_1) y las etiquetas de clase obtenidas por el método de agrupamiento (C_2). Con ello se busca identificar el número de pares de elementos con los que ambos grupos concuerdan (denotados más adelante como $a+b$), como se ilustra en la Figura 5.4. Un método de agrupamiento no produce como salida las mismas etiquetas de clase del conjunto de datos original, no obstante, denota a los grupos encontrados con una etiqueta que mapea a la etiqueta original.

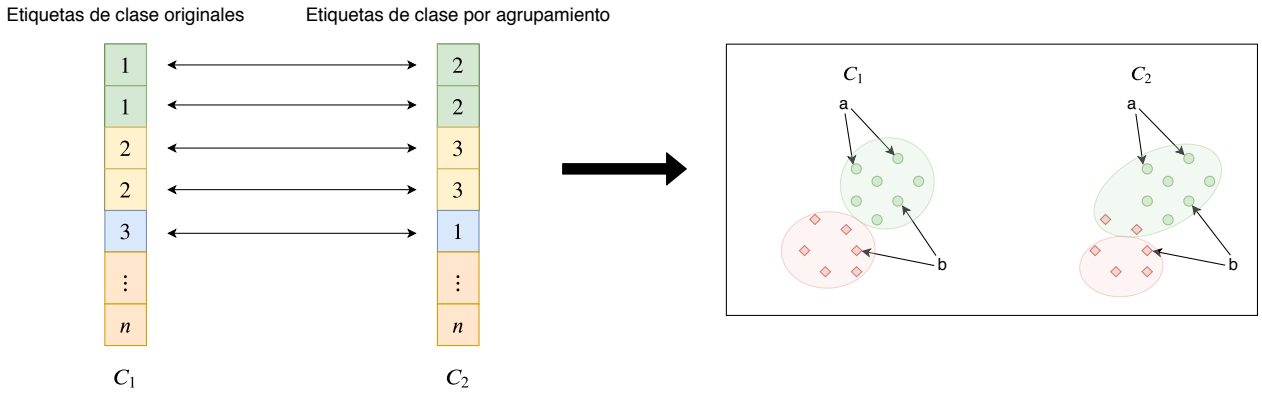


Figura 5.4: Identificación del número de pares de elementos con los que ambos grupos concuerdan (a y b).

Una vez identificado el número de pares de elementos con los que concuerdan, el RI se define como indica la Ecuación 5.3.

$$\text{RI}\,(C_1, C_2) = \frac{a+b}{\binom{m}{2}}, \tag{5.3}$$

donde a es el número de pares de elementos que se encuentran en un mismo grupo, b el número de pares de elementos que se encuentran en grupos diferentes y $\binom{m}{2}$ es el número de pares de

elementos que se pueden elegir en m instancias. Como resultado se obtiene un porcentaje de la similitud que hay entre subconjuntos, donde 0 denota que ningún grupo de etiquetas de clase fue identificado correctamente y 1.0 denota que todos los grupos de etiquetas de clase fueron identificados correctamente.

Una desventaja que se presenta al usar el RI es que al aumentar la cantidad de grupos hay una mayor probabilidad de que un par de elementos se encuentren en grupos diferentes. En dicho escenario ambos grupos concuerdan, por lo que el resultado se ve afectado por un sesgo de aleatoriedad. Para lidiar con dicho defecto existe una métrica llamada *Adjusted Rand Index* [43] (ARI por sus siglas en inglés), la cual mejora el RI en el sentido que toma en cuenta la aleatoriedad. La diferencia que hay entre RI y ARI es que en el ARI se consideran todos los pares de elementos que hay en todos los grupos, mientras que en RI se considera solamente el número de pares de elementos con los que concuerdan ambos grupos. El ARI se define como indica la Ecuación 5.4.

$$\text{ARI} = \frac{RI - \mathbb{E}(\text{RI})}{1 - \mathbb{E}(\text{RI})} \tag{5.4}$$

donde $\mathbb{E}(\text{RI})$ es el valor esperado de RI que, por medio de probabilidad, indica qué valores esperar a largo plazo y 1 representa el valor máximo que se puede obtener del RI.

Para calcular el ARI se crea una tabla de contingencia, como la que se ilustra en la Tabla 5.2, donde X y Y representan los dos conjuntos de etiquetas de clase (el original y el calculado por el método de agrupamiento), n_{ij} el número de pares de elementos que tienen en común ambos grupos, por último se obtiene una sumatoria de las filas y columnas denotadas como a_i y b_j respectivamente.

Tabla 5.2: Tabla de contingencia para calcular el ARI.

$X \setminus Y$	Y_1	Y_2	$\cdots$	Y_j	Sumatoria
X_1	n_{11}	n_{12}	$\cdots$	n_{1j}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2j}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	a_i
Sumatoria	b_1	b_2	$\cdots$	b_j	

Con base a la tabla de contingencia, el ARI se define de la siguiente forma como indica la Ecuación 5.5.

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_j \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}. \tag{5.5}$$

A continuación se presenta un ejemplo del cálculo del ARI. Supóngase que se tienen dos grupos de etiquetas de clase $X = \{3, 2, 1, 3, 1, 2, 1, 1, 3, 2, 3, 2\}$ y $Y = \{3, 3, 1, 3, 2, 2, 1, 2, 2, 1, 1, 3\}$ después de un proceso de agrupación, donde se tienen tres clases diferentes en ambos grupos $(1, 2, 3)$. La tabla de contingencia con respecto a los grupos de etiquetas de clase se muestra en la Tabla 5.3, donde se presenta el número de ocurrencias de etiquetas en ambos grupos. Por ejemplo, en la intersección de X_1 y Y_1 se cuenta el número de veces que la etiqueta 1 se repite en la misma posición en ambos grupos, en este caso es 2. Otro ejemplo, es en la intersección X_3 y Y_2, donde se cuenta el número de veces que la etiqueta 3 y 2 se repiten en la misma posición en ambos grupos, siendo éste 1. Esto se repite por cada intersección.

Tabla 5.3: Ejemplo de la tabla de contingencia para calcular el ARI.

$X \setminus Y$	Y_1	Y_2	Y_3	Sumatoria
X_1	2	2	0	4
X_2	1	1	2	4
X_3	1	1	2	4
Sumatoria	4	4	4	

Tomando como referencia la Ecuación del ARI, a partir de la tabla de contingencia se pueden hacer los cálculos $\sum_{ij}\binom{n_{ij}}{2}$, $\sum_i \binom{a_i}{2}$ y $\sum_j \binom{b_j}{2}$. Siguiendo con el ejemplo anterior, ésto se resuelve de la siguiente forma:

$$\sum_{ij}\binom{n_{ij}}{2} = \binom{2}{2} + \binom{2}{2} + \binom{0}{2} + \binom{1}{2} + \binom{1}{2} + \binom{2}{2} + \binom{1}{2} + \binom{1}{2} + \binom{2}{2} = 6 \tag{5.6}$$

$$\sum_{i}\binom{a_i}{2} = \binom{4}{2} + \binom{4}{2} + \binom{4}{2} = 18 \tag{5.7}$$

$$\sum_j \binom{b_j}{2} = \binom{4}{2} + \binom{4}{2} + \binom{4}{2} = 18 \tag{5.8}$$

Sustituyendo los valores en la fórmula del ARI, se obtiene que:

$$\text{ARI} = \frac{6 - [324]/\binom{12}{2}}{\frac{1}{2}[36] - [324]/\binom{12}{2}} = 0.083 \tag{5.9}$$

El resultado indica que las etiquetas de clase de ambas agrupaciones tienen una similitud de 0.083 donde entre más cercano a 1.0 es mayor el parecido entre las etiquetas de clase de ambas agrupaciones. En este ejemplo la similitud es muy baja entre el grupo de clases original y el calculado por el método de agrupación.

5.4 Resultados

En la esta subsección se presentan los resultados de las pruebas de significancia estadística obtenidos del proceso de experimentación. Para efectos de visualización estos resultados se agruparon de acuerdo al tipo de conjunto de datos utilizado: categórico, mixto y numérico.

Los resultados presentados corresponden a la experimentación para la evaluación de dos métricas: ARI y el crecimiento dimensional. Para el ARI, se obtuvo la media de la distribución muestral μ', cada muestra correspondió al conjunto de datos completo. Cada muestra se ejecutó 100 veces, que por el TLC (Teorema del Límite Central) converge al parámetro μ de una distribución normal de medias, por lo que $\mu' = \mu$ o en adelante simplemente μ.

5.4.1 Comparativa del ARI para datos de tipo categórico

En la Tabla 5.4 se presentan los resultados del ARI para los conjuntos de datos de tipo categórico. Se destaca el nombre del método, la media estimada (μ), la desviación estándar estimada (σ), los valores mínimo y máximo donde cae el 95 % de los resultados, usando un nivel de significancia (α)

del 5 %. El símbolo $*$ al lado del método k-*means* significa que se utilizó la codificación *one-hot encoding* como técnica de transformación.

Los métodos que obtuvieron la mejor exactitud fueron k-*means* con 0.53 y k-*modes* con 0.43. En esta experimentación k-*means* le ganó a k-*modes*, pero a costo de aumentar la dimensionalidad; ésto no sucede al usar la transformación propuesta. A pesar de ello se obtuvieron resultados aproximados al usar k-*modes* con respecto al algoritmo k-*means*.

Tabla 5.4: Resultados del ARI para el conjunto de datos categórico.

Método	μ	σ	mín	máx
coolcat	0.00	0.01	-0.01	0.01
k-*modes*	**0.43**	**0.32**	**0.05**	**0.81**
k-*modes-modificado*	0.15	0.16	-0.05	0.34
rock	0.02	0.03	-0.01	0.05
k-*means* $(*)$	**0.53**	**0.35**	**0.12**	**0.94**

Al usar la codificación *one-hot encoding* hubo un aumento dimensional de hasta cinco veces el tamaño original de un conjunto de datos de tipo categórico. Mientras que al usar la transformación propuesta, el número de dimensiones se mantiene como el conjunto de datos original. Ésto se puede observar en la Figura 5.5, donde el conjunto de datos que aumentó más su dimensionalidad fue *Mushroom*, incrementando 5.5 veces su tamaño original. El conjunto de datos que menos aumentó su dimensionalidad fue *Vote*, incrementando dos veces su tamaño original. El aumento dimensional se da al utilizar la codificación *one-hot encoding* debido a la cantidad de valores que pueden tener los atributos categóricos. En el conjunto de datos *Mushroom* su dimensionalidad aumenta porque por cada atributo categórico cuenta con una gran cantidad de valores diferentes. En el caso de *Vote* cada atributo categórico contiene una respuesta binaria ("Y" o "N"), por lo que su dimensionalidad no aumentó tanto como *Mushroom*. Cuanto mayor sea la cantidad de valores por atributo categórico, mayor es la dimensionalidad final. Dada la naturaleza de los datos de tipo categóricos, al usar la codificación *one-hot encoding* siempre habrá un aumento dimensional.

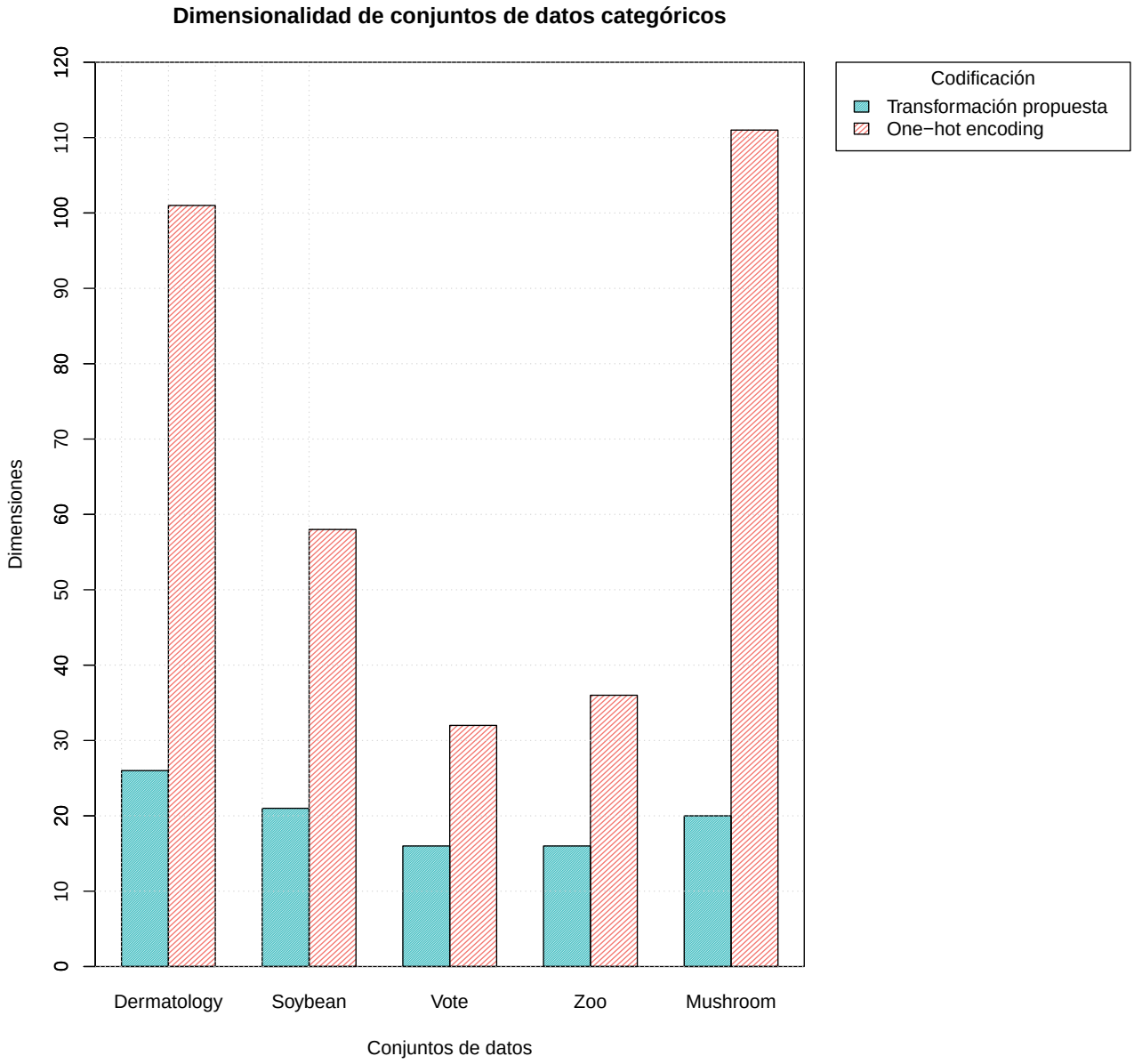


Figura 5.5: Crecimiento dimensional de los conjuntos de datos categóricos.

5.4.2 Comparativa del ARI para datos de tipo mixto

En la Tabla 5.5 se presentan los resultados del ARI para la experimentación de los conjuntos de datos de tipo mixto con un nivel de significancia (α) del 5 %. En los resultados el método que destaca con mejor desempeño es k-*modes* con 0.12, con el peor resultado se tiene a *coolcat* con 0.0

y k-*means* (método de referencia) con 0.02.

Tabla 5.5: Resultados del ARI para el conjunto de datos mixto.

Método	μ	σ	mín	máx
coolcat	**0.00**	**0.01**	**-0.01**	**0.01**
k-*modes*	**0.12**	**0.12**	**-0.02**	**0.26**
k-*modes-modificado*	0.02	0.06	-0.05	0.09
rock	0.02	0.04	-0.02	0.06
k-*means* $(*)$	**0.02**	**0.04**	**-0.02**	**0.07**

Desde la perspectiva dimensional, en la Figura 5.6 se puede apreciar el contraste del número de dimensiones con las que se trabaja al utilizar la codificación *one-hot encoding* y la transformación propuesta. En la Figura los conjuntos de datos con mayor crecimiento dimensional fueron *Autos*, *Australian Credit Data* y *German Credit Data*, aumentando aproximadamente tres veces sus dimensiones. El conjunto de datos con menor crecimiento dimensional fue *Heart Statlog*, el cual aumentó 1.7 veces sus dimensiones.

5.4.3 Comparativa del ARI para datos de tipo numérico

Es conocido que el método k-*means* es el método base para realizar tareas de agrupamiento en el espacio numérico. En la Tabla 5.6 muestran los resultados de ARI para los datos de tipo numérico con un nivel de significancia (α) del 5 %. Como es de esperarse, el método con mejor desempeño de ARI fue k-*means* con 0.56. Los dos siguientes mejores métodos fueron k-*modes-modificado* y k-*modes* con 0.25 y 0.21, respectivamente. Dado que la transformación propuesta cambia la representación de los atributos numéricos a categorías, existe una pérdida de precisión numérica. Por lo que es evidente que k-*means* haya sido el mejor de los métodos de agrupamiento.

5.4.4 Comparativa del ARI k-*means* contra k-*modes*

En los resultados obtenidos en la experimentación anterior se observó que los métodos con mejores ARI fueron k-*means* y k-*modes*. Posteriormente se buscó que, por medio de inferencia estadística, se

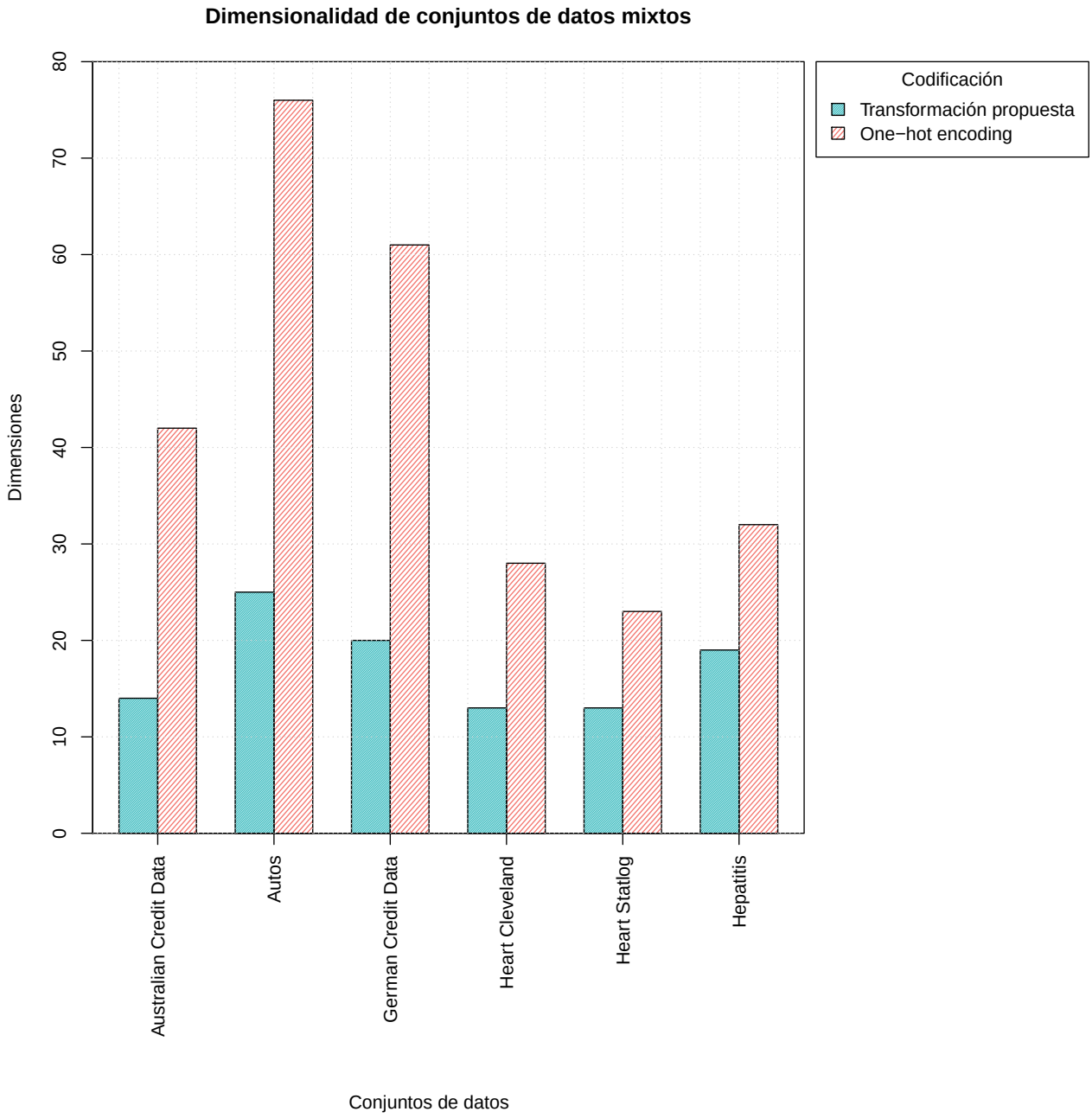


Figura 5.6: Crecimiendo dimensional de los conjuntos de datos mixtos.

pudiera identificar cuál de los métodos es mejor y con qué probabilidad. En la Figura 5.7 se presentan en términos de porcentaje cuál método es mejor para conjuntos de datos de tipo categórico. Con una confiabilidad del 95 %, un límite inferior del -1.02 y un límite superior del 0.82, el algoritmo k-*means*

Tabla 5.6: Resultados del ARI para conjuntos de datos numérico.

Método	μ	σ	mín	máx
coolcat	0.01	0.01	-0.00	0.01
k-*modes*	**0.21**	**0.24**	**-0.08**	**0.50**
k-*modes-modificado*	**0.25**	**0.23**	**-0.02**	**0.53**
rock	0.01	0.01	0.00	0.03
k-*means* $(*)$	**0.56**	**0.32**	**0.17**	**0.94**

fue mejor que k-*modes* con una probabilidad del 52.63 %.

Método	μ	σ
k-means	0.53	0.34
k-modes	0.43	0.32

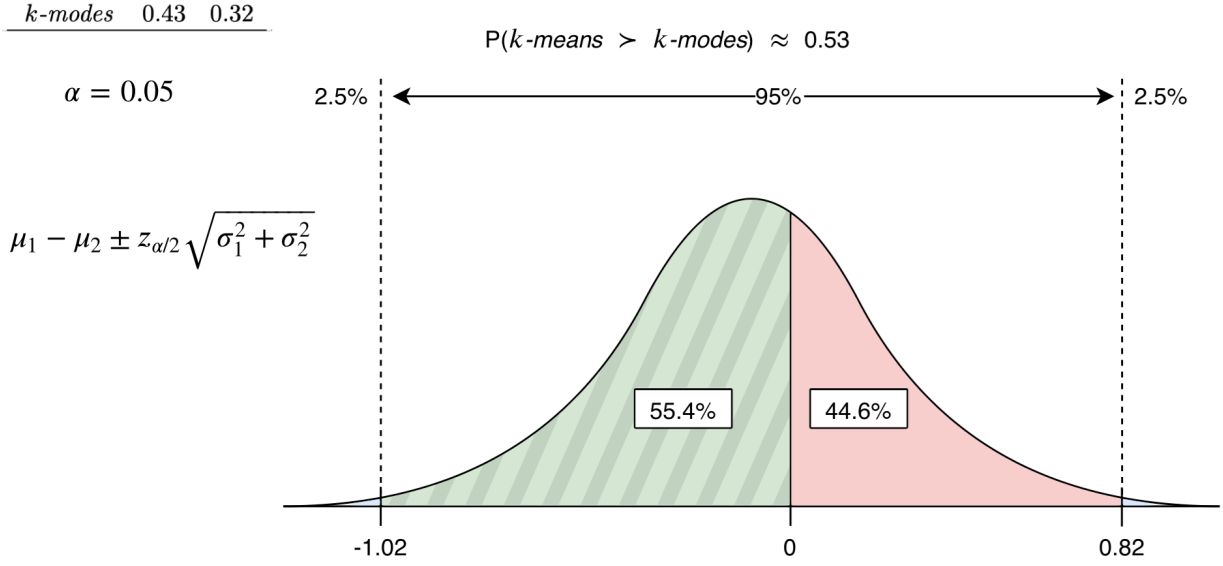


Figura 5.7: Resultados ARI k-*means* vs k-*modes* para datos de tipo categórico.

En la Figura 5.8 se ilustra que, para el caso de los datos de tipo mixto, se infiere que k-*modes* fue mejor que k-*means*. Esto con una probabilidad del 66.77 % y una confiabilidad del 95 %, un límite inferior de -0.14 y un límite superior de 0.34.

Por último, en la Figura 5.9 se presenta que el mejor método para los datos de tipo de dato numérico fue k-*means* con una probabilidad del 68.30 % y una confiabilidad del 95 %. Si bien este resultado era evidente, el propósito fue mostrar las ventajas y desventajas de la transformación propuesta. Por lo anterior, no se recomienda utilizar la transformación propuesta para conjuntos de

Método	μ	σ
k-means	0.02	0.036
k-modes	0.12	0.12

$\alpha = 0.05$

$\mu_1 - \mu_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2}$

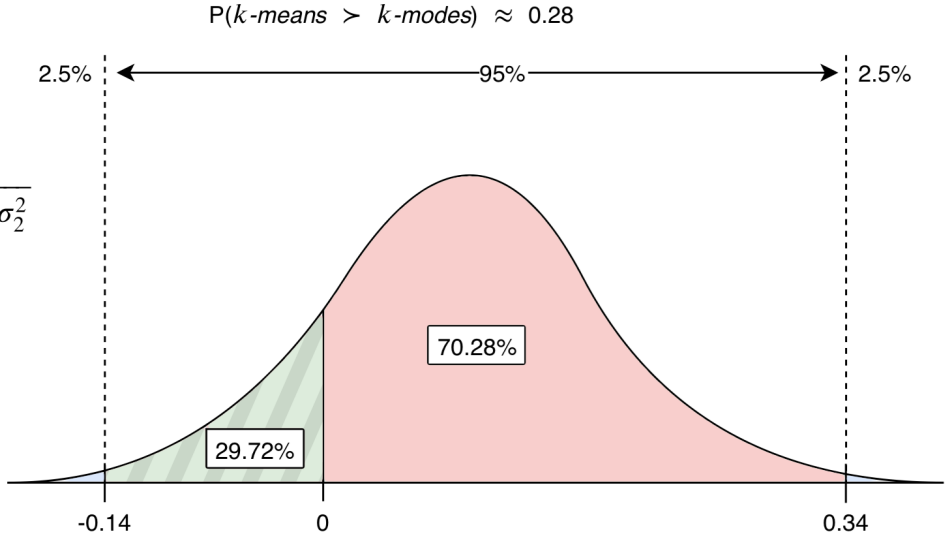


Figura 5.8: Resultados ARI k-*means* vs k-*modes* para datos de tipo mixto.

datos de tipo netamente numérico.

Método	μ	σ
k-means	0.56	0.32
k-modes	0.21	0.24

$\alpha = 0.05$

$\mu_1 - \mu_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2 + \sigma_2^2}$

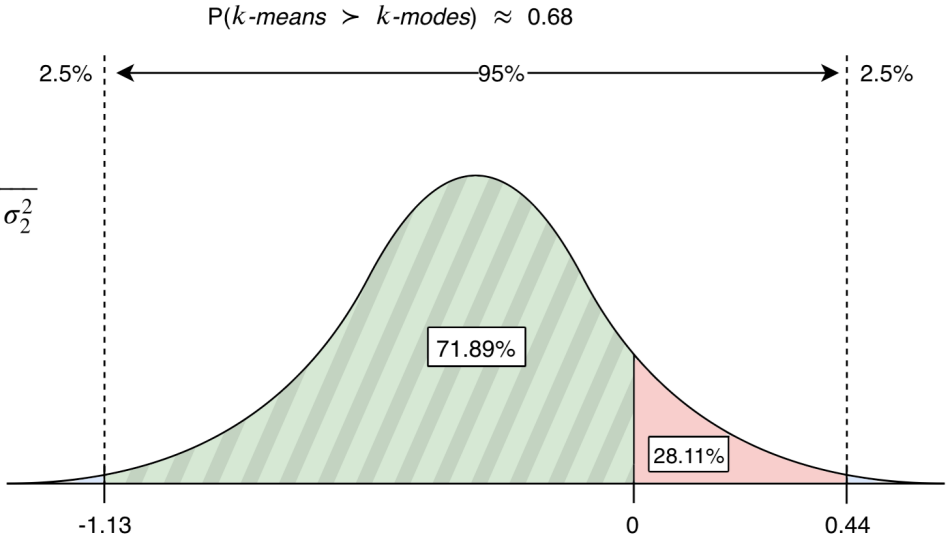


Figura 5.9: Resultados ARI k-*means* vs k-*modes* para datos de tipo numérico.

5.5 Discusión

En los resultados de la comparativa del ARI para datos de tipo categórico (ver Tabla 5.4), se puede ver que al utilizar la transformación propuesta y *one-hot encoding* se obtienen resultados similares con respecto a la métrica ARI. Esto se puede observar al usar el método *k-modes* y *k-means*, respectivamente con su codificación. Por lo tanto, cualquiera de los dos métodos de agrupación podría utilizarse para procesar un conjunto de datos categórico después de realizar la transformación de datos con el método propuesto. Sin embargo, al usar el *one-hot encoding* se observa que la dimensionalidad del conjunto de datos crece (ver Figura 5.6), esto afecta el procesamiento en memoria y procesador, lo cual eleva el costo computacional al procesar conjuntos de datos más grandes.

Al observar los resultados de la comparativa del ARI para datos de tipo mixto (ver Tabla 5.5) se puede ver que k-*modes* con la transformación propuesta resultó ser mejor que k-*means* con la codificación *one-hot encoding*. Esto se puede deber a que en la representación numérica realizada por el *one-hot encoding* se pierden las propiedades intrínsecas del espacio euclídeo, mientras que la transformación propuesta las mantiene. Adicionalmente, si se utiliza la codificación *one-hot encoding* ocurre un incremento en las dimensiones (ver Figura 5.6), lo cual aumenta el costo computacional en memoria y procesador. Por lo que creemos que usar la transformación propuesta y k-*modes* es más recomendable que k-*means* con la codificación *one-hot encoding* para conjuntos de datos mixtos.

Al analizar los resultados de la comparativa del ARI para los conjuntos de datos de tipo numérico (ver Tabla 5.6), se puede ver que si bien k-*means* resulta ser el mejor para este tipo de dato, los métodos k-*modes* y k-*modes-modificado* con la transformación propuesta siguen obteniendo mejores en comparación a los métodos de agrupamiento categórico. Para este caso no hay un aumento en las dimensiones ya que los datos de tipo numérico no requieren ser transformados por *one-hot encoding*.

Consideramos que la causa de que el método de agrupamiento k-*modes-modificado* presente un mejor desempeño respecto al ARI para los conjuntos de datos de tipo numérico (ver la Tabla 5.6), se debe a que usa la distancia de *Hamming* (donde la distancia obtenida es el número de atributos

diferentes entre objetos). Por lo que al usar el método de transformación propuesto, el número de categorías diferentes por atributo se define por la regla de *Sturges*. Mientras que el número de atributos categóricos de los conjuntos de datos de tipo mixto y categórico tienden a tener una mayor cantidad de categorías diferentes. En consecuencia a ello, es posible que la distancia de *Hamming* siempre mantenga la distancia máxima entre objetos para dichos conjuntos, evitando una cercanía.

6

Conclusiones y Trabajo Futuro

En este trabajo se propuso un método de transformación de datos de tipo mixto con el propósito de obtener resultados similares a los que se obtendrían al procesar un conjunto de datos con respecto a otra técnica de transformación como el *one-hot encoding*, pero sin incrementar el número de dimensiones. Con la experimentación se demostró que la transformación propuesta obtiene mejores resultados en tareas de agrupamiento de conjuntos de datos mixtos que al usar el *one-hot encoding*. Por lo que se recomienda el uso de la transformación propuesta únicamente para el procesamiento de datos de tipo mixto.

6.1 Resumen

Para el propósito de este trabajo de tesis se planteó la forma de representar un conjunto de datos de tipo mixto para tareas de aprendizaje automático. Se identificó que la mayoría de las tareas de aprendizaje automático procesan conjuntos de datos en su forma numérica, por lo que es necesario una forma de transformar los conjuntos de datos mixtos a una forma numérica, o categórica que es

la forma menos convencional y menos documentada en la literatura. La técnica más conocida para realizar transformación de datos es la codificación *one-hot encoding*, la cual transforma cada atributo categórico en múltiples atributos numéricos dependiendo del número de valores diferentes de cada atributo. Dicha codificación presenta como desventaja que produce un incremento en el número de dimensiones del conjunto de datos, aumentando su tiempo de procesamiento y almacenamiento en memoria, por lo que consideramos que es ineficiente. Existen otros métodos para la transformación de datos en el estado del arte, pero por practicidad se considera que *one-hot encoding* es la codificación más común. Las codificaciones existentes también presentan el mismo problema del crecimiento dimensional, además presentan un sesgo en el proceso de codificación ya que dependen de los valores de otro atributo en el mismo conjunto de datos. Para superar a estas problemáticas se propuso un método de transformación de datos de tipo numérico que permite una manipulación más eficiente en un espacio de categorías. Este método, para el caso de un conjunto de datos de tipo mixto o numérico, codifica los atributos numéricos por medio de un proceso de discretización mediante cuantiles. Como resultado se obtiene un conjunto de datos de tipo categórico (valores en escala nominal), donde cada instancia se representa con un código compuesto por identificadores numéricos correspondientes a cada atributo y no existe un aumento dimensional. Al conjunto de códigos se le puede aplicar diferentes tareas de aprendizaje automático categórico, para el propósito de este trabajo se abordó el caso de estudio de agrupamiento. Los resultados mostraron que al usar el método de transformación propuesto se pueden obtener mejores resultados que al usar la codificación *one-hot encoding* con respecto a la métrica ARI para conjuntos de datos mixtos.

6.2 Discusión

Durante la experimentación se identificó que al procesar los conjuntos de datos de tipo categórico con el método k-*modes* usando la transformación propuesta contra el k-*means* usando *one-hot encoding*, se obtienen resultados similares. Sin embargo, al usar la codificación *one-hot encoding*

hubo un incremento dimensional de los datos, por lo que su procesamiento en tiempo y memoria aumentan.

En la aplicación del método propuesto a conjuntos de datos de tipo mixto para procesar los datos con k-*modes* y k-*means*, se pudo observar que el desempeño de k-*modes* supera estadísticamente a k-*means* usando la codificación propuesta, sin aumentar la dimensionalidad; por lo que es recomendable usar la transformación propuesta sobre *one-hot encoding* para conjuntos de datos de tipo mixto.

Con la experimentación realizada se demostró que en agrupamiento se pueden obtener resultados similares o inclusive mejores al usar la transformación propuesta. Creemos que es posible obtener resultados similares al realizar la experimentación sobre tareas de clasificación. En este trabajo no se consideraron tareas de clasificación porque cada conjunto de datos requeriría de un proceso de selección de características, y debido a la alta dimensionalidad de los conjuntos de datos, el proceso de selección sería subjetivo al individuo que realiza la selección, lo cual daría como resultado un sesgo de los resultados. No es objetivo de este trabajo realizar tareas ajenas a la transformación de datos.

Para el caso de los conjuntos de datos numéricos, el método *one-hot encoding* usando k-*means* resultó ser el que mejor desempeño tuvo estadísticamente en relación a la transformación propuesta, por lo que no se recomienda el uso de la transformación propuesta para este tipo de conjuntos de datos.

6.3 Contribuciones

Después de haber desarrollado, implementado y evaluado el método propuesto para transformación de datos se obtuvieron las siguientes contribuciones.

- Un método de transformación de un conjunto de datos de tipo numérico o mixto a una representación categórica para que dicho conjunto de datos pueda ser procesado por tareas de aprendizaje automático de naturaleza categórica.

- Un estudio comparativo de agrupamiento de conjuntos de datos de tipo numérico, categórico y mixto usando la transformación propuesta contra el método *one-hot encoding* con métodos de agrupamiento categórico y numérico, respectivamente.

- Un prototipo que implementa el método propuesto de transformación de datos.

6.4 Limitantes

Si bien se obtuvieron resultados alentadores del método propuesto, éste también tiene limitantes. A continuación se describen las limitantes que fueron identificadas de la solución propuesta:

- La transformación propuesta únicamente sobresale en desempeño (con respecto al ARI) sobre conjuntos de datos de tipo mixto.

- La transformación propuesta solamente se puede usar para conjuntos de datos de tipo numérico o mixto. Esto debido a que la transformación propuesta codifica datos de tipo numérico a categorías.

- La transformación propuesta puede recibir como entrada conjuntos de datos de tipo numérico o mixto y da como salida un conjunto de datos de tipo categórico, por lo que sólo se pueden utilizar técnicas de aprendizaje automático de naturaleza categórica.

- El método propuesto sólo se evaluó con la tarea agrupamiento a los conjuntos de datos transformados.

6.5 Trabajo Futuro

Aunque se obtuvieron resultados aceptables y alentadores con el trabajo desarrollado, aún quedan algunos pendientes:

- Experimentar con tareas de clasificación. En este trabajo solamente se demostró que la transformación propuesta tiene un buen desempeño utilizando métodos de agrupamiento, sin embargo no se ha demostrado para las tareas de clasificación. Se planea como trabajo futuro la realización de una experimentación considerando tareas de clasificación.

- Estudiar y explorar la utilidad del método k-*modes-modificado* implementado. La versión actual del método emplea la medida de distancia *Chebyshev*, hace falta analizar las propiedades del método para aprovechar esa adaptación.

- Realizar un estudio del costo computacional al emplear el método de transformación de datos propuesto. Por el momento sólo se analizó la proporción de tamaño que incrementa un conjunto de datos cuando se aplica *one-hot encoding*. Falta demostrar el ahorro en costo computacional que se lograría en términos de uso de recursos computacionales de memoria, disco y procesador.

Bibliografía

[1] Agresti, A. and Kateri, M. (2011). *Categorical data analysis*, pages 206–208. Springer Berlin Heidelberg.

[2] Ahmad, A. and Dey, L. (2011). A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognition Letters*, 32(7):1062–1069.

[3] Aldana-Bobadilla, E. and Alfaro-Pérez, C. (2015). Finding the optimal sample based on shannon's entropy and genetic algorithms. In Sidorov, G. and Galicia-Haro, S. N., editors, *Advances in Artificial Intelligence and Soft Computing*, pages 353–363, Cham. Springer International Publishing.

[4] Aldana-Bobadilla, E. and Kuri-Morales, A. (2015). A clustering method based on the maximum entropy principle. *Entropy*, 17(1):151–180.

[5] Aldana-Bobadilla, E., Lopez-Arevalo, I., and Molina Villegas, A. (2017). A novel data reduction method based on information theory and the eclectic genetic algorithm. 21:803–826.

[6] Alkharusi, H. (2012). Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education*, 4(2):202.

[7] Allison, P. D. (2012). *Logistic Regression Using SAS: Theory and Application, Second Edition*. SAS Institute Inc., Cary, NC, USA, 2nd edition.

[8] Altman, N. and Krzywinski, M. (2018). The curse (s) of dimensionality. *Nat Methods*, 15:399–400.

[9] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

[10] Andritsos, P., Tsaparas, P., Miller, R. J., and Sevcik, K. C. (2004). Limbo: Scalable clustering of categorical data. In *International Conference on Extending Database Technology*, pages 123–146. Springer.

[11] Barbará, D., Li, Y., and Couto, J. (2002). Coolcat: An entropy-based algorithm for categorical clustering. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, pages 582–589, New York, NY, USA. ACM.

[12] Barbará, D., Li, Y., and Couto, J. (2013). *coolcat: Clustering algorithm Coolcat.* R package version 0.2-21.

[13] Barcelo-Rico, F. and Diez, J.-L. (2012). Geometrical codification for clustering mixed categorical and numerical databases. *Journal of Intelligent Information Systems*, 39(1):167–185.

[14] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203.

[15] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[16] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[17] Buchta, C. and Hahsler, M. (2019). *CBA: Clustering for Business Analytics*. R package version 0.2-21.

[18] Carey, G. (2003). Coding categorical variables. `http://psych.colorado.edu/~carey/Courses/PSYC5741/handouts/Coding%20Categorical%20Variables%202006-03-03.pdf`. (Accessed on 08/07/2019).

[19] Cerda, P., Varoquaux, G., and Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, 107(8-10):1477–1494.

[20] Chapelle, O., Manavoglu, E., and Rosales, R. (2015). Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):61.

[21] Chu, H., Liao, X., Dong, P., Chen, Z., Zhao, X., and Zou, J. (2019). An automatic classification method of well testing plot based on convolutional neural network (cnn). *Energies*, 12(15):2846.

[22] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[23] Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.

[24] Davis, P. J. (1975). *Interpolation and approximation*. Courier Corporation, New York, NY.

[25] Dawson, C. and Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modeling. *Hydrological Sciences Journal*, 43:47–66.

[26] Dillon, R. L. and Tinsley, C. H. (2016). Near-miss events, risk messages, and decision making. *Environment Systems and Decisions*, 36(1):34–44.

[27] Doane, D. P. (1976). Aesthetic Frequency Classifications. *The American Statistician*, 30(4):181–183.

[28] Du, M., Jiang, J., Jiang, Z., Lu, Z., and Du, X. (2019). PRTIRG: A knowledge graph for people-readable threat intelligence recommendation. In *International Conference on Knowledge Science, Engineering and Management*, pages 47–59. Springer.

[29] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

[30] Elfil, M. and Negida, A. (2016). Sampling methods in clinical research; an educational review. *EMERGENCY*, 4.

[31] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231.

[32] Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

[33] Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

[34] Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.

[35] Ganti, V., Gehrke, J., and Ramakrishnan, R. (1999). Cactus-clustering categorical data using summaries. In *KDD*, volume 99, pages 73–83.

[36] Garavaglia, S. and Sharma, A. (1998). A smart guide to dummy variables: Four applications and a macro. In *Proceedings of the Northeast SAS Users Group Conference*, page 43.

[37] García, S., Luengo, J., and Herrera, F. (2015). *Data preprocessing in data mining*. Springer.

[38] Guha, S., Rastogi, R., and Shim, K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345 – 366.

[39] Gujarati, D. N. (2009). *Basic econometrics*. Tata McGraw-Hill Education.

[40] Hall, C. A. and Meyer, W. W. (1976). Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory*, 16(2):105–122.

[41] Hogg, R. V., Tanis, E., and Zimmerman, D. (2014). Probability and statistical inference (9th edition). pages 200–206.

[42] Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34. Singapore.

[43] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

[44] Izenman, A. J. (2013). Linear discriminant analysis. In *Modern multivariate statistical techniques*, pages 237–280. Springer.

[45] Jolliffe, I. T. (1986). Principal Component Analysis and Factor Analysis. In *Principal Component Analysis*, pages 115–128. Springer.

[46] Juszczak, P., Tax, D., and Duin, R. P. (2002). Feature scaling in support vector data description. In *Proc. ASCI*, pages 95–102. Citeseer.

[47] Kaufman, L. and Rousseeuw, P. (1987). Statistical data analysis based on the l1norm. *Clustering by means of medoids*, pages 405–416.

[48] Keogh, E. and Mueen, A. (2017). *Curse of Dimensionality*, pages 314–315. Springer US, Boston, MA.

[49] Khalid, S., Khalil, T., and Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *2014 Science and Information Conference*, pages 372–378.

[50] Kohonen, T. and Somervuo, P. (1998). Self-organizing maps of symbol strings. *Neurocomputing*, 21(1):19–30.

[51] Kozak, M. (2012). "A dendrite method for cluster analysis" by caliński and harabasz: A classical work that is far too often incorrectly cited. *Communications in Statistics — Theory and Methods*, 41(12):2279–2280.

[52] Kuri-Morales, A. F. (2015). Categorical encoding with neural networks and genetic algorithms. In *WSEAS Proceedings of the 6th International Conference on Applied Informatics and. Computing Theory*, pages 167–175.

[53] Lane, D. (2003). Histograms. http://onlinestatbook.com/2/graphing_distributions/histograms.html. (Accessed on 02/27/2019).

[54] Lei, S. (2012). A feature selection method based on information gain and genetic algorithm. In *2012 International Conference on Computer Science and Electronics Engineering*, volume 2, pages 355–358.

[55] Li, A. and Axhausen, K. W. (2019). Comparison of short-term traffic demand prediction methods for transport services. *Arbeitsberichte Verkehrs-und Raumplanung*, 1447.

[56] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

[57] Machine Learning Repository, U. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/index.php. (Accessed on 06/13/2019).

[58] Mani, K. and Kalpana, P. (2015). A filter-based feature selection using information gain with median based discretization for naive bayesian classifier'. *International Journal of Applied and Engineering Research*, 10(82):280–285.

[59] Maron, M. E. (1961). Automatic indexing: An experimental inquiry. *J. ACM*, 8(3):404–417.

[60] Metsis, V., Androutsopoulos, I., and Paliouras, G. (2006). Spam filtering with naive bayes - which naive bayes?

[61] Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *SIGKDD Explor. Newsl.*, 3(1):27–32.

[62] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

[63] Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

[64] Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage.

[65] Paar, C. and Pelzl, J. (2010). Hash functions. In *Understanding Cryptography*, pages 293–317. Springer.

[66] Potdar, K., Pardawala, T. S., and Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175(4):7–9.

[67] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[68] Quinlan, J. R. (1983). *Learning Efficient Classification Procedures and Their Application to Chess End Games*, pages 463–482. Springer Berlin Heidelberg, Berlin, Heidelberg.

[69] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[70] Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13.

[71] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[72] Ren, M., Liu, P., Wang, Z., and Pan, X. (2016). An improved mixed-type data based kernel clustering algorithm. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 1205–1209. IEEE.

[73] Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1619–1630.

[74] Rokach, L. and Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications*, volume 69. World scientific.

[75] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.

[76] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

[77] Sammon, J. W. (1969). A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18(5):401–409.

[78] Sangam, R. S. and Om, H. (2018). An equi-biased k-prototypes algorithm for clustering mixed-type data. *Sādhanā*, 43(3):37.

[79] Silverman, B. W. and Jones, M. C. (1989). An important contribution to nonparametric discriminant analysis and density estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, 57(3):233–238.

[80] Stevens, S. S. et al. (1946). On the theory of scales of measurement.

[81] Strasser, S., Goodman, R., Sheppard, J., and Butcher, S. (2016). A new discrete particle swarm optimization algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pages 53–60. ACM.

[82] Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153):65–66.

[83] Van de Velden, M., Iodice D'Enza, A., and Markos, A. (2019). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3):e1456.

[84] Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1073–1080, New York, NY, USA. ACM.

[85] Voit, J. (2013). *The statistical mechanics of financial markets*, pages 97–99. Springer Science & Business Media.

[86] Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klar analyzing german business cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

[87] Weinberger, K., Dasgupta, A., Langford, J., Smola, A., and Attenberg, J. (2009). Feature hashing for large scale multitask learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1113–1120, New York, NY, USA. ACM.

[88] Zavras, D., Zavras, A. I., Kyriopoulos, I.-I., and Kyriopoulos, J. (2016). Economic crisis, austerity and unmet healthcare needs: the case of greece. *BMC health services research*, 16(1):309.