INDICE

|  |  |
| --- | --- |
| **Resumen** | 6 |
| Introducción | 8 |
| **Objetivo** | 9 |
| **Definición del problema** | 9 |
| Capitulo 1 | 10 |
| 1. Data Mining | 10 |
|  1.1 Generalidades | 10 |
|  1.2 ¿Qué es el Data Mining? | 12 |
|  1.2.1 Desarrollo de los Sistemas Data Mining | 13 |
|  1.2.2 ¿Qué hacen las herramientas Data Mining? | 14 |
|  1.2.3 El alcance del Data Mining | 16 |
|  1.2.4 ¿Cómo trabaja el Data Mining? | 18 |
|  1.3 Arquitectura del Data Mining | 20 |
|  1.4 ¿Por qué usar Data Mining? | 23 |
|  1.5 Data Mining versus Estadística | 25 |
|  1.6 Metodología del Data Mining | 26 |
|  |  |
| Capitulo 2 | 37 |
| 2. Técnicas de agrupamiento | 37 |
|  2.1 Introducción | 37 |
|  2.2 Clustering | 38 |
|  2.2.1 Requerimientos de un algoritmo de Clustering en Data Mining | 40 |
|  2.2.2 Métodos de Clustering | 44 |
|  2.2.2.1 Métodos de Particionamiento1 | 44 |
|  2.2.2.2 Métodos Jerárquicos | 46 |
|  2.2.2.3 Métodos Basados en densidad o distancias | 48 |
|  2.2.2.4 Métodos Basados en Modelos | 49 |
|  2.2.2.5 Métodos Basados en Grillas | 49 |
|  2.2.3 Tipos de Clasificación | 50 |
|  2.2.3.1 Aglomerativa / Divisiva | 50 |
|  2.2.3.2 Jerárquica / No jerárquica | 50 |
|  2.2.3.3 Monotética / Politética | 51 |
|  2.3 Método K-Medias | 53 |
|  2.3.1 Algoritmo K-Medias | 55 |
|  2.3.2 Representaciónb Gráfica del Algoritmo K-Medias | 58 |
|  2.3.3 Medidas de disimilitud entre cadenas | 59 |
|  2.4 Distancia de Edición | 60 |
|  2.5 Distancia de Levenshtein | 61 |
|  2.5.1 Implementación del Algoritmo de Levenshtein  | 62 |
|  |  |
| Capítulo 3 | 67 |
| 3.- Agrupación de las descripciones | 67 |
|  3.1 Muestreo | 67 |
|  3.1.1 Terminología | 67 |
|  3.1.2 Muestreo Probabilístico | 68 |
|  3.1.3 Muestreo aleatorio Simple | 68 |
|  3.1.4 Muestreo Sistemático | 69 |
|  3.1.5 Muestreo aleatorio Estratificado | 70 |
|  3.1.6 Muestreo por conglomerados | 71 |
|  3.2 Calculo del tamaño muestral | 72 |
|  3.2.1 Formula para el calculo del tamaño de muestra | 73 |
|  3.2.2 Obtención de los datos | 74 |
|  3.2.3 Análisis de los datos | 75 |
|  3.3 Criterio de agrupación | 76 |
|  |  |
| Capítulo 4 |  |
| **4. Anexos** |  |
|  4.1 Código fuente del proceso de agrupación de las descripciones | 82 |
|  4.2 Lista de artículos  | 85 |
|  4.2.1 Grupos categorizados | 95 |
|  4.3 Conclusiones y recomendaciones | 98 |
|  4.4 Record Linkage | 100 |
|  |  |
| Bibliografía | 107 |
|  |  |