Apéndice B: Marco Teórico

Práctica 1: Implementación del algoritmo de clasificación K-NN

1. Minería de datos y clasificación

La minería de datos es un área que busca extraer conocimiento útil a partir de grandes volúmenes de información. Dentro de este campo, la **clasificación** es una tarea de aprendizaje supervisado que consiste en asignar una etiqueta o clase a un objeto, utilizando como referencia un conjunto de datos previamente etiquetado. Su objetivo principal es construir un modelo capaz de predecir la clase de nuevos datos con base en patrones encontrados en el conjunto de entrenamiento.

2. Aprendizaje supervisado

El aprendizaje supervisado se caracteriza por utilizar ejemplos de entrada con su salida esperada (clase). Es decir, el modelo aprende una relación entre atributos y etiquetas a partir de datos conocidos. Posteriormente, dicha relación se utiliza para predecir clases en datos desconocidos. En esta práctica, las clases corresponden a especies de flores del conjunto Iris.

3. Algoritmo K-Nearest Neighbors (K-NN)

El algoritmo **K-Nearest Neighbors (K-NN)** es un método de clasificación basado en instancias. En lugar de generar un modelo matemático explícito durante el entrenamiento, K-NN almacena los datos etiquetados y realiza la clasificación al momento de recibir un nuevo objeto.

El proceso general de K-NN es:

1. Calcular la distancia entre el objeto desconocido y todos los objetos del conjunto de entrenamiento.
2. Ordenar las distancias de menor a mayor.
3. Seleccionar los K vecinos más cercanos.
4. Asignar la clase por votación mayoritaria.

4. Distancia euclidiana

La **distancia euclidiana** es una medida utilizada para cuantificar qué tan cerca se encuentran dos puntos en un espacio de características. Para dos objetos p y q con n atributos, se define como:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{1}$$

En esta práctica se trabajó con el conjunto de datos Iris, el cual posee $n = 4$ atributos numéricos, por lo que cada flor puede representarse como un punto en un espacio de cuatro dimensiones.

5. Votación mayoritaria

Después de seleccionar los K vecinos más cercanos, se determina la clase final mediante **votación mayoritaria**. Esto significa que la clase asignada será aquella que aparezca con mayor frecuencia entre los vecinos seleccionados. Matemáticamente:

$$\hat{y} = \arg\max_{c} \ \text{count}(c) \tag{2}$$

donde $\text{count}(c)$ es el número de vecinos que pertenecen a la clase c.

6. Exactitud (Accuracy)

Para evaluar el desempeño del clasificador se utilizó la métrica **exactitud** o *accuracy*, la cual indica la proporción de objetos correctamente clasificados respecto al total de objetos evaluados:

$$\text{Exactitud} = \frac{A}{B} \tag{3}$$

donde:

- A es el número de casos correctamente clasificados.
- B es el total de casos del conjunto de prueba.

Una exactitud más alta indica que el modelo clasifica correctamente una mayor cantidad de objetos.