Imagen que contiene Icono

Descripción generada automáticamenteDibujo en blanco y negro

Descripción generada automáticamente con confianza bajaINSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

INTELIGENCIA ARTIFICIAL

**PRÁCTICA 9**

**CLASIFICADORES EUCLIDIANO Y 1NN**

NOMBRE DEL ALUMNO: GARCÍA QUIROZ GUSTAVO IVAN

NOMBRE DEL PROFESOR: GARCÍA FLORIANO ANDRES

GRUPO:6CV3

FECHA DE ENTREGA DEL REPORTE: 31/10/2024

Índice

[1 Introducción 1](#_Toc182295547)

[2 Marco teórico 2](#_Toc182295548)

[2.1 Conjunto de datos de la flor iris 2](#_Toc182295549)

[2.1.1 Características del conjunto de datos de la flor iris 2](#_Toc182295550)

[2.2 Conjunto de datos de diagnóstico de cáncer de mama en Wisconsin 2](#_Toc182295551)

[2.2.1 Características del conjunto de datos del cáncer de mama de Wisconsin 3](#_Toc182295552)

[2.3 Conjunto de datos del vino 4](#_Toc182295553)

[2.3.1 Características del conjunto de datos del vino 4](#_Toc182295554)

[2.4 Algoritmo de K-vecino más cercano (KNN) 5](#_Toc182295555)

[2.4.1 Distancia euclidiana 6](#_Toc182295556)

[2.4.2 Distancia de Manhattan 6](#_Toc182295557)

[2.4.3 Distancia de Minkowski 7](#_Toc182295558)

[3 Desarrollo de la práctica. 8](#_Toc182295559)

[3.1 Implementación de los Clasificadores 8](#_Toc182295560)

[3.1.1 Clasificador Euclidiano 8](#_Toc182295561)

[3.2 Clasificador 1-NN (K-Nearest Neighbor con k=1) 9](#_Toc182295562)

[3.3 Métodos de Validación Implementados 10](#_Toc182295563)

[3.3.1 Hold Out 70/30 10](#_Toc182295564)

[3.3.2 10-Fold Cross-Validation 10](#_Toc182295565)

[3.3.3 Leave-One-Out 11](#_Toc182295566)

[4 Análisis de Resultados 12](#_Toc182295567)

[4.1 Resultados 1](#_Toc182295568)

[5 Conclusiones 4](#_Toc182295569)

[6 Referencias. 5](#_Toc182295570)

[7 Código 5](#_Toc182295571)

# Introducción

La clasificación es una tarea que consiste en asignar una categoría o etiqueta a un conjunto de datos de entrada. Existen diversos algoritmos de clasificación que se utilizan para abordar este tipo de problemas de clasificación, cada uno con sus propias características y enfoques. En esta práctica, nos enfocaremos en la implementación y evaluación de dos clasificadores clásicos: el Clasificador Euclidiano y el Clasificador 1-NN (1-Nearest Neighbor).

El Clasificador Euclidiano se basa en el cálculo de la distancia euclidiana entre un punto de prueba y los puntos de entrenamiento, asignando la clase del vecino más cercano. Por otro lado, el Clasificador 1-NN extiende este concepto al considerar los k vecinos más cercanos, aunque en esta práctica utilizaremos específicamente k=1.

Para el desarrollo de esta práctica, se utilizó el lenguaje de programación Python. Python es una opción popular en el campo del aprendizaje automático y el análisis de datos debido a su sintaxis clara, la amplia disponibilidad de bibliotecas y herramientas de soporte, y su facilidad de uso.

Para evaluar el desempeño de estos clasificadores, se implementarán tres métodos de validación comúnmente utilizados: Holdout 70/30, Validación Cruzada de 10 Pliegues y Leave-One-Out. Estos métodos permiten estimar la capacidad de generalización de los modelos y brindar una visión más completa de su rendimiento.

En esta práctica, se analizan los resultados del desempeño de los clasificadores en tres conjuntos de datos diferentes. Estos tres conjuntos de datos presentan diferentes niveles de complejidad y características, lo que permitirá evaluar el desempeño de los clasificadores en diversos escenarios. Estos tres conjuntos de datos son: flores iris, diagnóstico de cáncer de mama y vinos de diferentes regiones.

Iris Plant Dataset es un conjunto de datos clásico en el campo del aprendizaje automático, que contiene información sobre tres especies de flores iris. Cada muestra está descrita por cuatro características (longitud y anchura de los sépalos y pétalos) y pertenece a una de las tres clases (tipos) de iris. En total, este conjunto de datos cuenta con 150 muestras.

Breast Cancer Dataset un conjunto de datos proviene de exámenes de diagnóstico de cáncer de mama. Cada muestra representa las características de un tumor, y la tarea es clasificar si el tumor es benigno o maligno. El conjunto de datos cuenta con alrededor de 569 muestras.

Wine Dataset es un conjunto de datos contiene información química sobre vinos de diferentes regiones. El objetivo es clasificar los vinos en uno de los tres tipos (clases) diferentes. El conjunto de datos consta de 178 muestras.

# Marco teórico

## Conjunto de datos de la flor iris

El conjunto de datos flor Iris o conjunto de datos iris de Fisher es un conjunto de datos multivariante introducido por Ronald Fisher en su artículo de 1936, The use of multiple measurements in taxonomic problems (El uso de medidas múltiples en problemas taxonómicos) como un ejemplo de análisis discriminante lineal. A veces, se llama Iris conjunto de datos de Anderson porque Edgar Anderson coleccionó los datos para cuantificar la variación morfológica de la flor Iris de tres especies relacionadas.

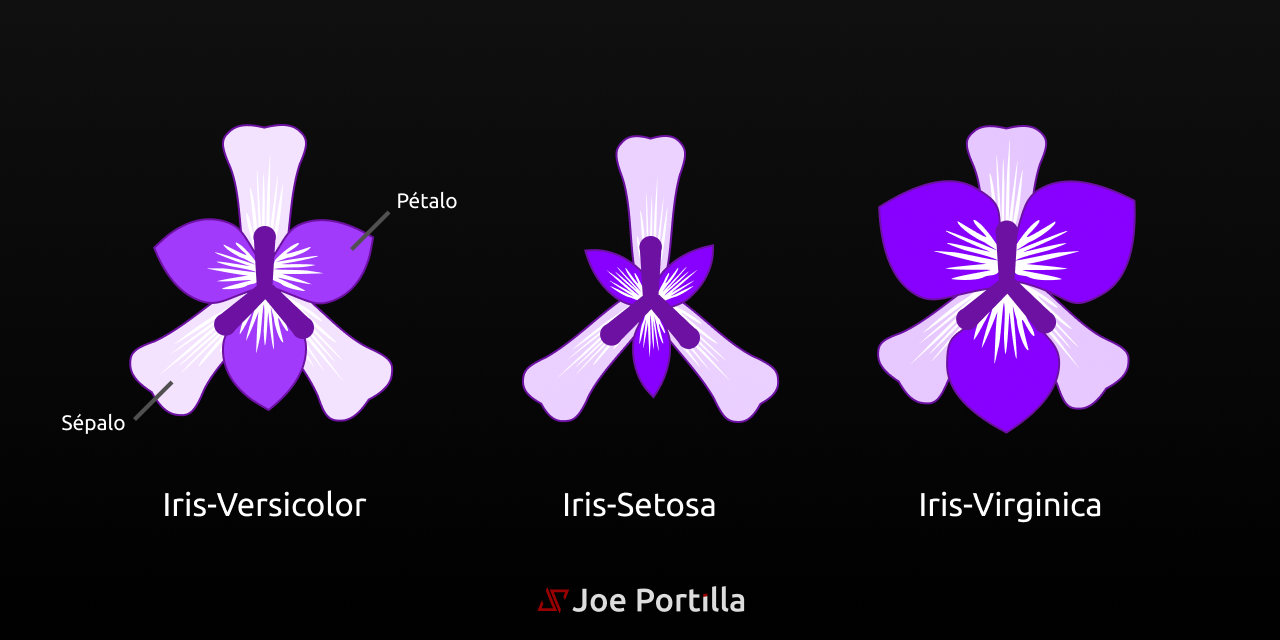


Figura Planta Iris.

### Características del conjunto de datos de la flor iris

El conjunto de datos contiene 50 muestras de cada una de tres especies de Iris (Iris setosa, Iris virginica e Iris versicolor). Se midió cuatro rasgos de cada muestra: el largo y ancho del sépalo y pétalo, en centímetros. Basado en la combinación de estos cuatro rasgos, Fisher desarrolló un modelo discriminante lineal para distinguir entre una especie y otra.

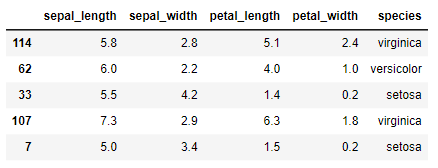


Figura Características del conjunto de datos de la flor iris.

## Conjunto de datos de diagnóstico de cáncer de mama en Wisconsin

El conjunto de datos de diagnóstico de cáncer de mama de Wisconsin es una reconocida colección de datos que se utiliza ampliamente en el aprendizaje automático y la investigación médica. Este conjunto de datos, que se origina a partir de imágenes digitalizadas de aspiraciones con aguja fina (AAF) de masas mamarias, facilita el análisis de las características de los núcleos celulares para ayudar en el diagnóstico del cáncer de mama.

El conjunto de datos de diagnóstico de cáncer de mama de Wisconsin es un conjunto de datos conocido que se utiliza habitualmente en el aprendizaje automático. El conjunto de datos fue seleccionado por el Dr. William H. Wolberg, W. Nick Street y Olvi L. Mangasarian. Contiene características calculadas a partir de imágenes digitalizadas de muestras de tejido mamario aspirado con aguja fina (FNA).

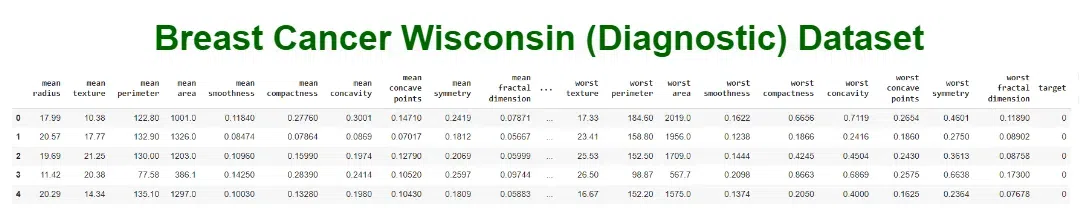


Figura Conjunto de datos de diagnóstico de cáncer de mama en Wisconsin.

### Características del conjunto de datos del cáncer de mama de Wisconsin

* Número de instancias: 569.
* Número de atributos: 30 atributos numéricos utilizados para la predicción, junto con una etiqueta de clase.
* Distribución de clases: 212 - Maligno, 357 – Benigno.

El conjunto de datos incluye 30 características, entre ellas la media, el error estándar y los valores "peores" o "más grandes", calculados para cada imagen. Estas características encapsulan varios aspectos de las características de los núcleos celulares:

* Radio medio: Media de las distancias desde el centro a los puntos del perímetro.
* Textura media: desviación estándar de los valores de la escala de grises.
* Perímetro medio: Perímetro del tumor.
* Área media: Área del tumor.
* Suavidad media: variación en las longitudes de los radios.
* Compacidad media: Perímetro^2 / Área - 1.0.
* Concavidad media: Severidad de las porciones cóncavas del contorno.
* Puntos cóncavos medios: Número de porciones cóncavas del contorno.
* Simetría media: Simetría de los núcleos celulares.
* Dimensión fractal media: "Aproximación de la línea de costa" - 1

|  |  |
| --- | --- |
| **Clases** | 2 |
| **Muestras por clase** | 212(M),357(B) |
| **Muestras totales** | 569 |
| **Dimensionalidad** | 30 |
| **Características** | real, positivo |

Tabla Características del conjunto de datos del cáncer de mama de Wisconsin.

## Conjunto de datos del vino

El conjunto de datos Wine Recognition es un conjunto de datos de referencia clásico ampliamente utilizado en el aprendizaje automático para tareas de clasificación. Proporciona información valiosa sobre la clasificación del vino en función de varios atributos químicos.

El conjunto de datos original de Wine fue creado por Forina, M. et al, como parte del proyecto PARVUS, un paquete extensible para exploración, clasificación y correlación de datos, realizado en el Instituto de Análisis y Tecnologías Farmacéuticas y Alimentarias, Génova, Italia.

El conjunto de datos sobre vinos contiene los resultados de un análisis químico de vinos cultivados en tres regiones diferentes de Italia. En concreto, incluye 13 atributos derivados de mediciones de diversos componentes presentes en los vinos. Estos atributos suelen incluir factores como el contenido de alcohol, los niveles de acidez y las concentraciones de diferentes compuestos químicos, como fenoles y flavonoides. Estos atributos proporcionan información valiosa sobre la composición química de los vinos y pueden utilizarse para tareas de clasificación de vinos.

### Características del conjunto de datos del vino

El conjunto de datos de reconocimiento de Wine posee varias características clave que lo hacen ideal para tareas de clasificación y experimentación con aprendizaje automático. Estas características brindan información sobre la estructura, el tamaño y la naturaleza de los datos que contiene el conjunto de datos.

|  |  |
| --- | --- |
| **Número de instancias:** | 178 |
| **Número de atributos:** | 13 atributos numéricos, predictivos y la clase |
| **Información del atributo:** | * Alcohol * Ácido málico * Ceniza * Alcalinidad de la ceniza * Magnesio * Fenoles totales * Flavonoides * Fenoles no flavonoides * Proantocianinas * Intensidad del color * Matiz * OD280/OD315 de vinos diluidos * Prolina |

Tabla Características del conjunto de datos del vino.

Tres clases correspondientes al origen del vino:

1. Clase 1: Vinos de la primera región (denominados como "clase\_0")
2. Clase 2: Vinos de la segunda región (denominados como "clase\_1")
3. Clase 3: Vinos de la tercera región (denominados como "clase\_2")

El conjunto de datos de reconocimiento de vinos se utiliza habitualmente para tareas de aprendizaje supervisado, en particular algoritmos de clasificación. Los investigadores y profesionales suelen emplear técnicas de aprendizaje automático para crear modelos que puedan predecir con precisión el origen de los vinos en función de su composición química.

## Algoritmo de K-vecino más cercano (KNN)

El algoritmo K-Nearest Neighbors (KNN) es un método de aprendizaje automático supervisado que se emplea para resolver problemas de clasificación y regresión. Evelyn Fix y Joseph Hodges desarrollaron este algoritmo en 1951, que posteriormente fue ampliado por Thomas Cover. El artículo explora los fundamentos, el funcionamiento y la implementación del algoritmo KNN.

KNN es uno de los algoritmos de clasificación más básicos y esenciales en el aprendizaje automático. Pertenece al dominio del aprendizaje supervisado y encuentra una aplicación intensa en el reconocimiento de patrones, la minería de datos y la detección de intrusiones.

Es ampliamente descartable en escenarios de la vida real, ya que no es paramétrico, lo que significa que no hace suposiciones subyacentes sobre la distribución de datos (a diferencia de otros algoritmos como GMM, que suponen una distribución gaussiana de los datos dados). Se nos proporcionan algunos datos previos (también llamados datos de entrenamiento), que clasifican las coordenadas en grupos identificados por un atributo.

A modo de ejemplo, considere la siguiente tabla de puntos de datos que contienen dos características:

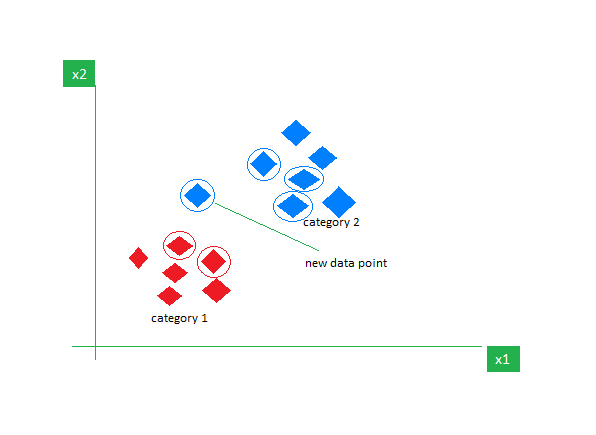


Figura Visualización del funcionamiento del algoritmo KNN.

Ahora, dado otro conjunto de puntos de datos (también llamados datos de prueba), se asigna estos puntos a un grupo analizando el conjunto de entrenamiento.

**Métricas de distancia utilizadas en el algoritmo KNN**

Como sabemos, el algoritmo KNN nos ayuda a identificar los puntos o grupos más cercanos para un punto de consulta. Pero para determinar los grupos o puntos más cercanos para un punto de consulta necesitamos algunas métricas. Para este propósito, utilizamos las siguientes métricas de distancia:

### Distancia euclidiana

Esta no es otra cosa que la distancia cartesiana entre los dos puntos que se encuentran en el plano/hiperplano. La distancia euclidiana también se puede visualizar como la longitud de la línea recta que une los dos puntos en consideración. Esta métrica nos ayuda a calcular el desplazamiento neto realizado entre los dos estados de un objeto.



Figura Distancia euclidiana.

### Distancia de Manhattan

La métrica de distancia de Manhattan se utiliza generalmente cuando nos interesa la distancia total recorrida por el objeto en lugar del desplazamiento. Esta métrica se calcula sumando la diferencia absoluta entre las coordenadas de los puntos en n dimensiones.



Figura Distancia de Manhattan.

### Distancia de Minkowski

Podemos decir que la distancia euclidiana, así como la distancia de Manhattan, son casos especiales de la distancia de Minkowski.



Figura Distancia de Minkowski.

De la fórmula anterior podemos decir que cuando p = 2 entonces es la misma que la fórmula para la distancia euclidiana y cuando p = 1 entonces obtenemos la fórmula para la distancia de Manhattan.

Las métricas analizadas anteriormente son las más comunes cuando se trata de un problema de aprendizaje automático, pero también existen otras métricas de distancia, como la distancia de Hamming, que resultan útiles cuando se tratan problemas que requieren comparaciones superpuestas entre dos vectores cuyo contenido puede ser tanto booleano como de cadena.

# Desarrollo de la práctica.

El desarrollo de esta práctica se centró en los clasificadores Euclidiano y 1-NN, así como los diversos métodos de validación utilizados para conseguir los resultados. Los conjuntos de datos utilizados son:

1. Iris Plant Dataset: Conjunto de datos clásico con 3 clases de flores iris de 150 muestras con 4 características cada una.
2. Breast Cancer Dataset: Conjunto de datos de diagnóstico de cáncer de mama y es una clasificación binaria (maligno/benigno).
3. Wine Dataset: Conjunto de datos de clasificación de vinos y contiene datos químicos de vinos de diferentes regiones.

## Implementación de los Clasificadores

En esta sección se detalla la implementación de dos clasificadores fundamentales: el Clasificador Euclidiano y el Clasificador 1-NN. Ambos clasificadores fueron desarrollados utilizando Python con la biblioteca NumPy para operaciones numéricas eficientes.

### Clasificador Euclidiano

El Clasificador Euclidiano se implementó mediante la clase EuclideanClassifier, que contiene dos métodos principales: fit y predict. El código base de esta implementación es el siguiente:

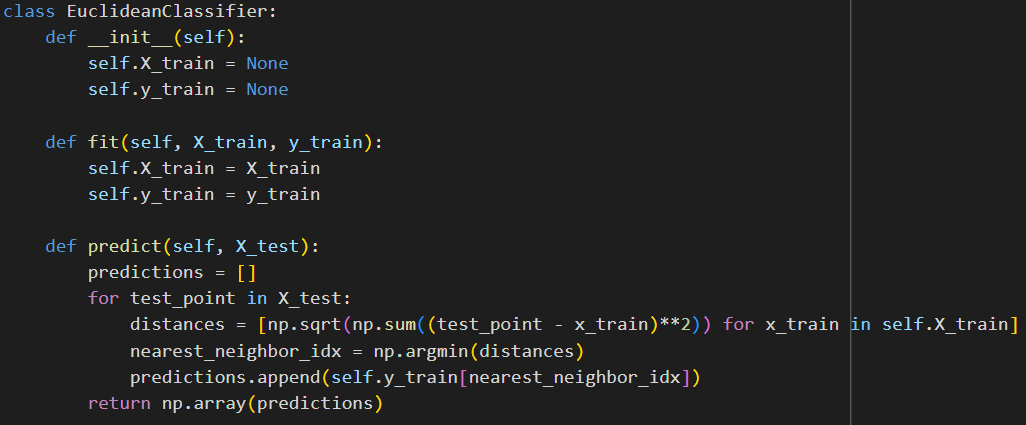


Figura Clasificador Euclidiano.

El método fit almacena los datos de entrenamiento en la instancia del clasificador. El método predict implementa la lógica de clasificación calculando la distancia euclidiana entre cada punto de prueba y todos los puntos de entrenamiento. La distancia euclidiana se calcula mediante la fórmula , implementada utilizando operaciones vectorizadas de NumPy para mejor rendimiento.

## Clasificador 1-NN (K-Nearest Neighbor con k=1)

El Clasificador 1-NN se implementó a través de la clase KNNClassifier, que extiende el concepto del clasificador euclidiano pero con una estructura más generalizable. El código correspondiente es:

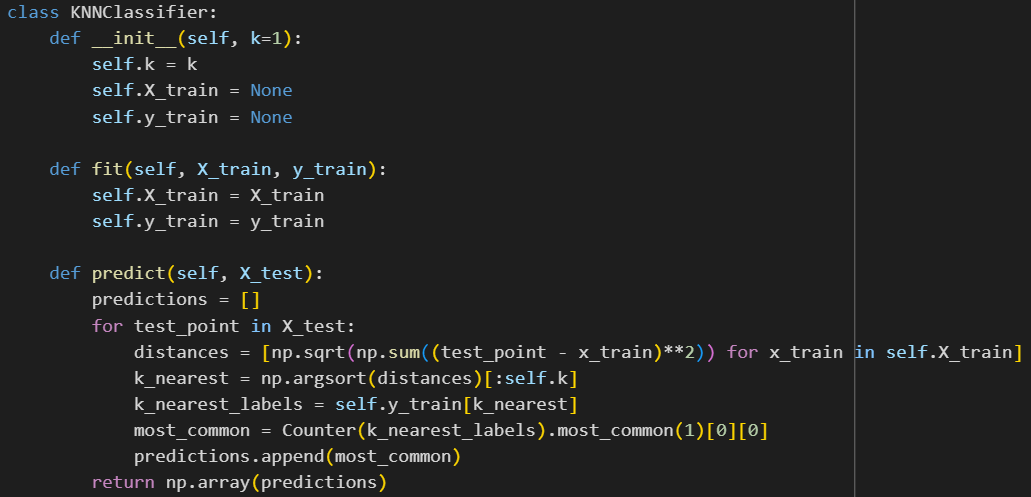
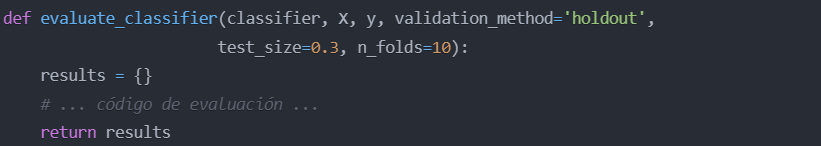


Figura Clasificador 1-NN.

La implementación del KNNClassifier incluye una variable k que determina el número de vecinos a considerar. Para este caso específico, se utiliza k=1, lo que significa que cada punto se clasifica según la clase de su vecino más cercano. El método predict calcula las distancias euclidianas, ordena los vecinos por distancia y selecciona la clase más común entre los k vecinos más cercanos utilizando la clase Counter de la biblioteca collections.

La diferencia principal entre ambas implementaciones radica en que el Clasificador Euclidiano siempre toma el vecino más cercano, mientras que el KNNClassifier está diseñado para manejar múltiples vecinos, aunque en esta práctica específica se configura con k=1. Ambos clasificadores utilizan la misma métrica de distancia euclidiana pero difieren en la estructura de su implementación para permitir diferentes niveles de flexibilidad en su uso.

La implementación de ambos clasificadores se complementa con la función evaluate\_classifier, que permite evaluar su desempeño utilizando diferentes métodos de validación:



Figura

Esta función proporciona una interfaz unificada para evaluar ambos clasificadores bajo diferentes esquemas de validación, permitiendo una comparación directa de su desempeño en distintos conjuntos de datos.

## Métodos de Validación Implementados

En esta sección se detalla la implementación de tres métodos de validación distintos: Hold Out 70/30, 10-Fold Cross-Validation y Leave-One-Out. Estos métodos se implementaron dentro de la función evaluate\_classifier, que acepta como parámetros el clasificador a evaluar, los datos de entrada, y el método de validación a utilizar.

### Hold Out 70/30

El método Hold Out divide el conjunto de datos en dos partes: 70% para entrenamiento y 30% para prueba. Este método realiza una única división de los datos, entrena el modelo con el conjunto de entrenamiento y evalúa su desempeño con el conjunto de prueba. Se utiliza estado aleatorio igual a 42 para asegurar la reproducibilidad de los resultados.

El código se realizó utilizando la función train\_test\_split de scikit-learn:

Texto

Descripción generada automáticamente

Figura Método Hold Out.

### 10-Fold Cross-Validation

La validación cruzada de 10 pliegues divide el conjunto de datos en 10 partes iguales, utilizando 9 partes para entrenamiento y 1 para prueba en cada iteración. Este método realiza 10 iteraciones de entrenamiento y prueba, donde cada muestra del conjunto de datos es utilizada exactamente una vez como dato de prueba. Las matrices de confusión de cada iteración se suman para obtener una matriz de confusión general del proceso.

El código se realizó usando KFold de scikit-learn:

Texto

Descripción generada automáticamente

Figura Método para la validación cruzada de 10 pliegues.

### Leave-One-Out

El método Leave-One-Out es un caso especial de validación cruzada donde el número de pliegues es igual al número de muestras en el conjunto de datos. En este método, cada muestra se utiliza una vez como dato de prueba mientras que el resto de las muestras se utilizan para entrenamiento. Se mantienen listas separadas para almacenar todas las predicciones y valores verdaderos, lo que permite calcular la matriz de confusión general al final del proceso.

El código se realizó utilizando LeaveOneOut de scikit-learn:

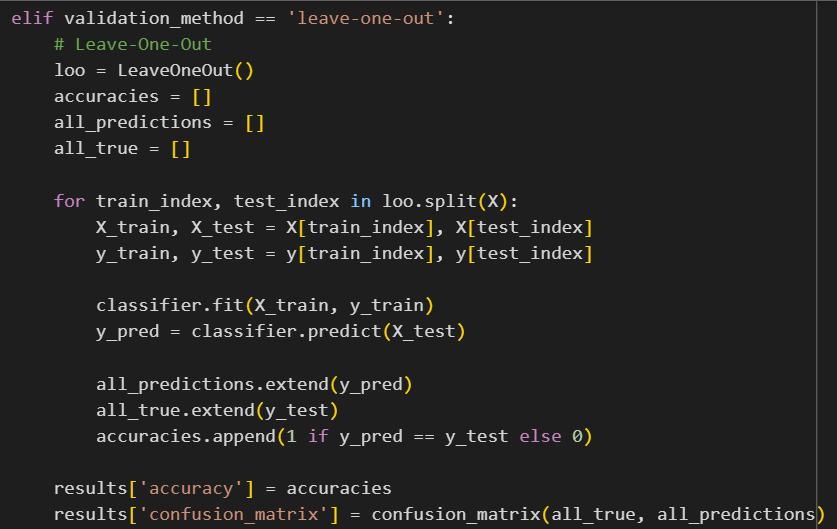


Figura Método para Leave-One-Out

El código completo de estos métodos de validación se integra en la función principal. Esta función coordina la evaluación de cada combinación de clasificador y método de validación sobre los tres conjuntos de datos seleccionados, proporcionando una estructura organizada para la comparación sistemática de los resultados.

* Iris Plant Dataset.
* Breast Cancer Dataset.
* Wine Dataset.

# Análisis de Resultados

Los resultados se obtuvieron para cada combinación de:

* 2 clasificadores (Euclidiano y 1NN)
* 3 métodos de validación
* 3 datasets

Para cada combinación se calculó:

* Accuracy (precisión)
* Matriz de confusión

Este análisis se dividirá por conjuntos de datos, observando el rendimiento de ambos clasificadores con los diferentes métodos de validación.

**Dataset Iris**

Para el conjunto de datos Iris, ambos clasificadores (Euclidiano y 1NN) mostraron resultados idénticos:

1. **Hold Out (70/30)**:
   * Accuracy: 1.0000 (100%)
   * La matriz de confusión muestra una clasificación perfecta: 19 muestras de la primera clase, 13 de la segunda y 13 de la tercera, todas correctamente clasificadas.
2. **10-Fold Cross-Validation y Leave-One-Out**:
   * Accuracy: 0.9600 (96%)
   * Matrices de confusión idénticas que muestran:
     + 50 clasificaciones correctas para la primera clase
     + 47 correctas y 3 errores para la segunda clase
     + 47 correctas y 3 errores para la tercera clase

**Dataset Breast Cancer**

Los clasificadores mostraron resultados idénticos en este conjunto de datos binario:

1. **Hold Out (70/30)**:
   * Accuracy: 0.9357 (93.57%)
   * La matriz de confusión muestra:
     + 56 verdaderos positivos y 7 falsos negativos
     + 104 verdaderos negativos y 4 falsos positivos
2. **10-Fold Cross-Validation**:
   * Accuracy: 0.9191 (91.91%)
   * La matriz muestra 182 y 341 clasificaciones correctas para cada clase
3. **Leave-One-Out**:
   * Accuracy: 0.9156 (91.56%)
   * Resultados similares a 10-Fold con ligeras variaciones en los errores

**Dataset Wine**

El conjunto de datos Wine mostró el rendimiento más bajo de los tres datasets:

1. **Hold Out (70/30)**:
   * Accuracy: 0.7963 (79.63%)
   * La matriz de confusión muestra más errores de clasificación distribuidos entre las tres clases
2. **10-Fold Cross-Validation**:
   * Accuracy: 0.7310 (73.10%)
   * Mayor cantidad de errores de clasificación, especialmente en la tercera clase
3. **Leave-One-Out**:
   * Accuracy: 0.7697 (76.97%)
   * Muestra una mejora ligera respecto a 10-Fold

Las siguientes tablas muestran un resumen de los valores obtenidos en cada clasificador y en cada metodo de validación. Es importante destacar que los métodos de validación no son comparables entre sí, por ejemplo, el resultado de aplicar Hold Out no puede compararse con el resultado de aplicar Leave One Out, o K Fold Cross Validation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Euclidiano | | | 1-NN | | |
|  | HO | KF CV | LOO | HO | KF CV | LOO |
| Iris | 1 | 0.96 | 0.96 | 1 | 0.96 | 0.96 |
| Breast Cancer | 0.9357 | 0.9191 | 0.9156 | 0.9357 | 0.9191 | 0.9156 |
| Wine | 0.7963 | 0.731 | 0.7697 | 0.7963 | 0.731 | 0.7697 |

Figura 14 Tabla de valores de Accuracy

|  |  |  |  |
| --- | --- | --- | --- |
|  | Euclidiano | | |
|  | HO | KF CV | LOO |
| Iris | |  |  |  |  | | --- | --- | --- | --- | |  | **Setosa** | **Versicolor** | **Virginica** | | **Setosa** | 19 | 0 | 0 | | **Versicolor** | 0 | 13 | 0 | | **Virginica** | 0 | 0 | 13 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Setosa** | **Versicolor** | **Virginica** | | **Setosa** | 50 | 0 | 0 | | **Versicolor** | 0 | 47 | 3 | | **Virginica** | 0 | 3 | 47 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Setosa** | **Versicolor** | **Virginica** | | **Setosa** | 50 | 0 | 0 | | **Versicolor** | 0 | 47 | 3 | | **Virginica** | 0 | 3 | 47 | |
| Breast Cancer | |  |  |  | | --- | --- | --- | |  | **P** | **N** | | **P** | 56 | 7 | | **N** | 4 | 104 | | |  |  |  | | --- | --- | --- | |  | **P** | **N** | | **P** | 182 | 30 | | **N** | 16 | 341 | | |  |  |  | | --- | --- | --- | |  | **P** | **N** | | **P** | 182 | 30 | | **N** | 18 | 339 | |
| Wine | |  |  |  |  | | --- | --- | --- | --- | |  | **Clase 1** | **Clase 2** | **Clase 3** | | **Clase 1** | 17 | 0 | 2 | | **Clase 2** | 3 | 16 | 2 | | **Clase 3** | 1 | 3 | 10 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Clase 1** | **Clase 2** | **Clase 3** | | **Clase 1** | 52 | 3 | 4 | | **Clase 2** | 5 | 52 | 14 | | **Clase 3** | 3 | 19 | 26 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Clase 1** | **Clase 2** | **Clase 3** | | **Clase 1** | 52 | 3 | 4 | | **Clase 2** | 5 | 54 | 12 | | **Clase 3** | 3 | 14 | 31 | |

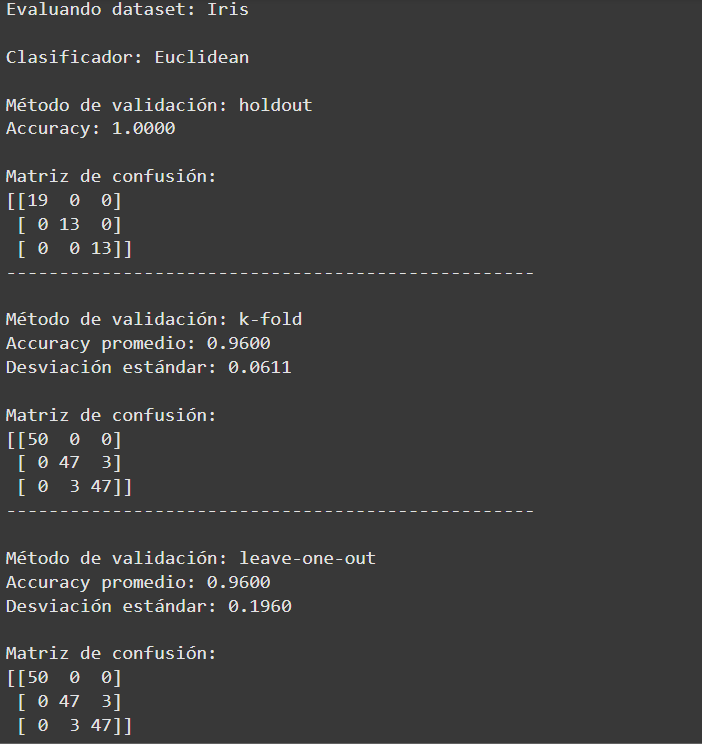
Tabla Tabla de valores de las Matrizes de confusión (Euclidiano).

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1-NN | | |
|  | HO | KF CV | LOO |
| Iris | |  |  |  |  | | --- | --- | --- | --- | |  | **Setosa** | **Versicolor** | **Virginica** | | **Setosa** | 19 | 0 | 0 | | **Versicolor** | 0 | 13 | 0 | | **Virginica** | 0 | 0 | 13 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Setosa** | **Versicolor** | **Virginica** | | **Setosa** | 50 | 0 | 0 | | **Versicolor** | 0 | 47 | 3 | | **Virginica** | 0 | 3 | 47 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Setosa** | **Versicolor** | **Virginica** | | **Setosa** | 50 | 0 | 0 | | **Versicolor** | 0 | 47 | 3 | | **Virginica** | 0 | 3 | 47 | |
| Breast Cancer | |  |  |  | | --- | --- | --- | |  | **P** | **N** | | **P** | 56 | 7 | | **N** | 4 | 104 | | |  |  |  | | --- | --- | --- | |  | **P** | **N** | | **P** | 182 | 30 | | **N** | 16 | 341 | | |  |  |  | | --- | --- | --- | |  | **P** | **N** | | **P** | 182 | 30 | | **N** | 18 | 339 | |
| Wine | |  |  |  |  | | --- | --- | --- | --- | |  | **Clase 1** | **Clase 2** | **Clase 3** | | **Clase 1** | 17 | 0 | 2 | | **Clase 2** | 3 | 16 | 2 | | **Clase 3** | 1 | 3 | 10 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Clase 1** | **Clase 2** | **Clase 3** | | **Clase 1** | 52 | 3 | 4 | | **Clase 2** | 5 | 52 | 14 | | **Clase 3** | 3 | 19 | 26 | | |  |  |  |  | | --- | --- | --- | --- | |  | **Clase 1** | **Clase 2** | **Clase 3** | | **Clase 1** | 52 | 3 | 4 | | **Clase 2** | 5 | 54 | 12 | | **Clase 3** | 3 | 14 | 31 | |

Tabla Tabla de valores de las Matrizes de confusión (1-NN).

## Resultados

Conjunto de datos de la flor iris

 Texto

Descripción generada automáticamente

Figura Resultados del conjunto de datos de flores iris

Conjunto de datos de diagnóstico de cáncer de mama.

Texto

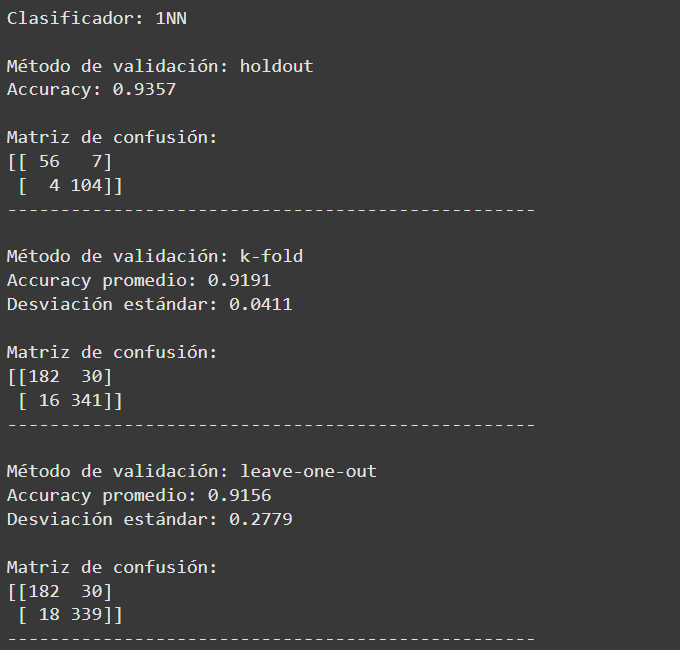
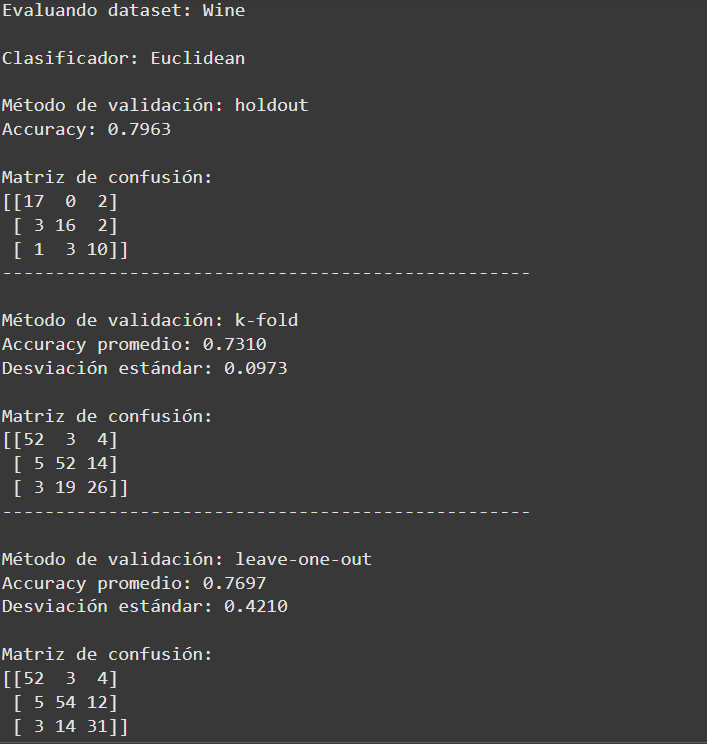
Descripción generada automáticamente

Figura Resultados de diagnóstico de cáncer de mama.

Conjunto de datos de clasificación de vinos.

Texto

Descripción generada automáticamente

Figura Resultados del conjunto de datos de clasificación de vinos.

# Conclusiones

Los clasificadores Euclidiano y 1-NN mostraron ser muy similares en su funcionamiento y resultados, lo cual es lógico ya que ambos se basan en el mismo principio de distancia entre puntos. La principal diferencia radica en que el clasificador 1-NN está diseñado de manera más flexible para permitir el uso de más vecinos, aunque en esta práctica se utilizó específicamente k=1.

En cuanto a los métodos de validación, cada uno mostró sus ventajas y desventajas. El método Hold Out 70/30 resultó ser el más rápido de ejecutar, pero proporciona menos información sobre el comportamiento real del clasificador, ya que solo usa una división de los datos. Por otro lado, el método de validación cruzada de 10 pliegues ofreció una visión más completa del desempeño del clasificador, ya que prueba el modelo con diferentes combinaciones de datos de entrenamiento y prueba. El método Leave-One-Out, aunque es el más completo en términos de uso de datos, resultó ser el más costoso en tiempo de ejecución, especialmente en conjuntos de datos grandes.

Los tres conjuntos de datos utilizados (Iris, Breast Cancer y Wine) permitieron probar los clasificadores en diferentes escenarios. El conjunto de datos Iris, siendo más pequeño y bien balanceado, mostró resultados consistentes en todos los métodos de validación. Los conjuntos de datos de Breast Cancer y Wine, al ser más grandes y complejos, permitieron una evaluación más rigurosa de los clasificadores.

# Referencias.

* *UCI machine learning repository*. (s/f). Uci.edu. Recuperado el 28 de octubre de 2024, de https://archive.ics.uci.edu/dataset/53/iris
* Wikipedia contributors. (s/f). *Conjunto de datos flor iris*. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Conjunto\_de\_datos\_flor\_iris&oldid=159237500
* <https://www.geeksforgeeks.org/breast-cancer-wisconsin-diagnostic-dataset/>
* <https://www.geeksforgeeks.org/wine-dataset/>

# Código

* <https://github.com/GUSTAVOIVANGQ/AI/tree/main/8_Datasets>