K-Nearest Neighbours

**Objective:**

The objective of this assignment is to implement and evaluate the K-Nearest Neighbours algorithm for classification using the given datasets

**Dataset:**

Need to Classify the animal type

**Tasks:**

1. Analyse the data using the visualizations

2. Preprocess the data by handling missing values & Outliers, if any.

3. Split the dataset into training and testing sets (80% training, 20% testing).

4. Implement the K-Nearest Neighbours algorithm using a machine learning library like scikit-learn On training dataset

5. Choose an appropriate distance metric and value for K.

6. Evaluate the classifier's performance on the testing set using accuracy, precision, recall, and F1-score metrics.

7. Visualize the decision boundaries of the classifier.

**Interview Questions:**

1. What are the key hyperparameters in KNN?

2. What distance metrics can be used in KNN?

Answer:

<https://colab.research.google.com/drive/1uvMrQEgNhhi2VqGExecOL_WEOJzKqFnn?usp=sharing>

Answer 1)

**K-Nearest Neighbors (KNN) Performance Parameters:**  
  
**Number of Neighbors (k):**  
• Crucial for determining the number of nearest neighbors for prediction.  
• Too small or too large can lead to overfitting or underfitting.  
 **Distance Metric:**  
• Common choices include Euclidean Distance, Manhattan Distance, Minkowski Distance, and Hamming Distance.  
  
**Weight Function:**  
• Determines how the distance of neighbors influences prediction.  
• Options include uniform, distance, and custom weights.  
  
**Algorithm for Nearest Neighbor Search:**  
• Options include brute force, KD-Tree, Ball Tree, and auto.  
  
**Leaf Size (for KD-Tree and Ball Tree):**  
• Influences the speed of tree construction and query time..  
  
Fine-tuning these hyperparameters can significantly affect KNN algorithm performance and accuracy.

Answer 2)

**K-Nearest Neighbors (KNN) Distance Metrics:**  
  
**Euclidean Distance:**  
• The straight-line distance between two points in Euclidean space.  
• Suitable for continuous variables and data in a Euclidean space.  
  
**Manhattan Distance (L1 Distance):**  
• Sums the absolute differences of their Cartesian coordinates.  
• Useful for high-dimensional data and grid-like structures.  
 **Minkowski Distance:**  
• A generalized metric that includes both Euclidean and Manhattan distances.  
• Used for categorical variables or binary data.  
  
**Hamming Distance:**  
• Measures the number of positions at which the corresponding symbols are different.  
• Primarily used for categorical variables or binary data.  
 **Mahalanobis Distance:**  
• Takes into account the correlations between variables and scales the distance based on the covariance matrix of the data.  
• Useful when dealing with multivariate data and when the data has correlations among different dimensions.  
  
**Cosine Similarity:**  
• Measures the cosine of the angle between two vectors.  
• Used for text data or high-dimensional sparse data.