*Investigate and Implement KNN Classifier*

1st Ankush Laxman Patil   
ankush.patil.stud@fra-uas.de

2nd Ayan Borthakur  
ayan.borthakur@stud.fra-uas.de

3rd Nasir Ishaq  
Nasir.ishaq@stud.fra-uas.de

*Abstract— Machine learning (ML) has grown over a wide range of sectors such as finance, education, communication, transportation, retail, and healthcare. In response to this, researchers have developed a variety of algorithms to analyze a large amount of data and derive quantifiable insights from that data. Subsequently, ML assisted enterprises to make impactful decisions. As many organizations work on different domains so they have a collection of the dataset in text, image, audio, video, and numeric formats so machine learning is used to create predictive models for speech recognition, natural language processing, computer vision, and recommendation systems.* *In this paper, we have proposed the implementation of the most widely used k-Nearest Neighbor (kNN) machine learning algorithm. This algorithm helps to solve the problems related to the classification of data based on pre-trained labels and predict the outcome for unlabeled data input. The developed k-NN model is also integrated with the Neocortex API in order to get the input dataset. The model gets a stream of sequences as input and then it classifies whether the predicated sequence has a match or mismatch with the input data sequence with 80% accuracy. Additionally, the design procedure, challenges, and enhancements to improve model accuracy are discussed in the paper.*

*Keywords— K-Nearest Neighbors; Machine Learning; Neocortex API; classifies; Unlabeled data set.*

# Introduction

The aim of machine learning is to create statistical models that allow computers to learn from data without being to be explicitly programmed. More precisely, these statistical models and algorithms generate the "learn" method, which helps the computer in predictions about the future based on the trained data. In general, supervised, unsupervised, and reinforcement learning are the three main types of machine learning.

### Supervised learning: The supervised machine learning algorithms learn from input variables and a pre-labeled dataset to predict an output variable.

### Unsupervised Learning: Unsupervised Learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

### Reinforcement Learning: Reinforcement Learning: This kind of algorithm learns by error-prone trial-and-error, and it is extensively used when an agent interacts with its environment in order to accomplish a certain task. Reinforcement learning involves learning a policy or decision-making rule that maximizes a reward function.

In addition to these main types, there are also other machine learning algorithms, such as semi-supervised learning, which combines supervised and unsupervised learning, and transfer learning, which leverages pre-trained models for new tasks. In this paper, we are more focused on The k-nearest neighbor (kNN) classifier which comes under the supervised machine learning algorithm andis widely used for pattern recognition and classification problems.

The kNN algorithm works by first selecting a value for K, which is the number of nearest neighbors to consider when making a prediction. Given a new observation, the algorithm calculates the distance between the new observation and all the existing observations in the training set. The most commonly used distance metrics are Euclidean distance, Manhattan distance, and Minkowski distance. It then selects the k observations that are closest to the new observation. Then the class is determined by a majority vote among the neighbors. it is simple and effective but the choice of value for k can significantly affect the performance of the algorithm.

# Literature Review

In this section, the kNN

we will understand the KNN in detail and understand the design procedure of different researchers

The selection of K is an important parameter in the algorithm, and its value determines the complexity and accuracy of the classifier.

In a study by Zhang et al. (2019), they proposed an adaptive KNN algorithm based on the density of the data points. The algorithm dynamically adjusts the value of K based on the density of the data points in the local region. They showed that the adaptive KNN algorithm outperformed the traditional KNN algorithm on several benchmark datasets.

Zhang et al. (2019) proposed an adaptive KNN algorithm based on the density of the data points in their study. Based on the density of the data points in the immediate area, the algorithm dynamically modifies the value of K. On a number of benchmark datasets, they demonstrated that the adaptive KNN algorithm performed better than the conventional KNN technique.

In another study by Yu et al. (2020), they proposed a weighted KNN algorithm that assigns different weights to the K nearest neighbors based on their distance from the query point. The weights are computed using a Gaussian kernel function, and the classifier is trained using a cross-validation technique. They showed that the weighted KNN algorithm outperformed the traditional KNN algorithm on several benchmark datasets.

Liu et al. (2021) proposed a hybrid KNN algorithm that combines KNN with the random subspace method. The algorithm randomly selects a subset of features and applies KNN to the reduced feature space. The process is repeated multiple times, and the results are combined using an ensemble technique. They showed that the hybrid KNN algorithm outperformed both the traditional KNN algorithm and the random subspace method on several benchmark datasets.

In conclusion, the literature suggests that KNN can be enhanced by adapting the value of K based on the density of the data points, assigning different weights to the K nearest neighbors based on their distance, and combining KNN with other algorithms using ensemble techniques. These techniques can improve the accuracy and robustness of the KNN classifier for various classification problems.

# Theoretical background and parameters

The k-nearest-neighbor classification algorithm was first used in an unpublished US Air Force School of Aviation Proceedings of the International Conference on Intelligent Computing and Control Systems to execute characteristic analysis when clear parametric approximations of probability densities were unknown or difficult to determine. In the KNN classifier algorithm, there are several parameters that plays important role in algorithm designs, and those parameters are discussed below,

## Distance calculations

The KNN algorithm works by finding the k-nearest neighbors to a new input sample in the training dataset based on a distance metric. The distance metric can be Euclidean distance, Manhattan distance, or any other appropriate distance measure. Once the k-nearest neighbors are identified, the algorithm assigns the new sample to the class that is most common among its k-nearest neighbors. The most common distance calculations techniques used in kNN are described below:

### Euclidean Distance: Euclidean distance is the straight-line distance between two points in a Euclidean space. In other words, it is the distance between two points in a 2D or 3D space. The Euclidean distance between two points (a1, b1) and (a2, b2) can be calculated as follows:

distance = sqrt((a2 - a1)^2 + (b2 - b1)^2)

2. Manhattan Distance: Manhattan distance is the distance between two points measured along the axes at right angles. It is named after the grid-like layout of the streets in Manhattan. The Manhattan distance between two points (a1, b1) and (a2, b2) can be calculated as follows:

distance = |a2 - a1| + |b2 - b1|

## K value selection:

The value of K is a hyperparameter that needs to be selected before making predictions. The value of K controls the number of neighbors that are used to make predictions. A larger value of K results in a smoother decision boundary but may result in more misclassifications of the training data. A smaller value of K results in a more complex decision boundary but may result in overfitting the training data. The choice of K is typically made using a validation set or cross-validation. The algorithm is trained on a portion of the data, and the remaining data is used to validate the performance of the algorithm for different values of K. The value of K that results in the highest validation accuracy is selected as the final value of K.

## Voting principle:

In the KNN algorithm, the prediction for a new observation is made based on the class labels of its K nearest neighbors. The class with the most number of occurrences among the K nearest neighbors is selected as the predicted class for the new observation. This is known as the majority voting principle.

The majority voting principle is simple yet effective and has been shown to perform well in many applications. However, it is important to note that the choice of K can affect the voting outcome, and a larger value of K may result in a less confident prediction. Additionally, in the case of imbalanced data, where one class has significantly more instances than the other, the majority voting principle may result in biased predictions. In such cases, techniques such as weighted voting or distance-weighted voting can be used to balance the influence of each class.

Apart from these parameters Feature selection and Preprocessing techniques such as normalization, standardization, and scaling can be considered while designing the model and increasing its accuracy of the model.

# Theoretical background and parameters

In this section, the design and working methodology of the KNN classifier is explained in detail with the help of examples.

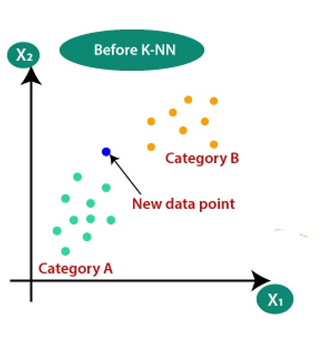


Figure. 1: Plotted KNN Data for prediction

As described in Figure. 1 the categorical data is plotted on X and Y axis and data has two primary categories i.e. Category A and Category B. The goal is to predict the new data points given in the blue column which do not belong to any category. This is a classic example of a classification-based problem and In order to predict the outcome we have to follow the below-mentioned steps,

Step-1: Select the number K of the neighbors for computation and it is recommended that select the odd value of K for better predictions. In this case, let's choose k=5. so the model will select the 5 nearest neighbors.

Step-2: In the second step, the algorithm will calculate the Euclidean distance as shown in Figure 2 from all the 5 nearest neighbors from the new data points and store the distance in the distance table.

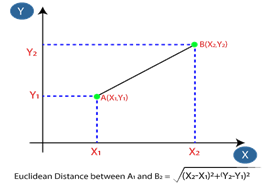


Figure. 2: Euclidean Distance Calculation

Step-3: The distance table will have all the values along with their known category label.

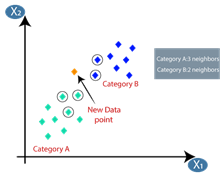


Figure. 3: Euclidean Distance Calculation

Step 4: In step 4, As Figure 3 shows the voting principle will calculate the vote for each category and then the highest number of vote categories is assigned to new data points. For example, the Category of the New Data Point is A and the final outcome is given in Figure.4 below,

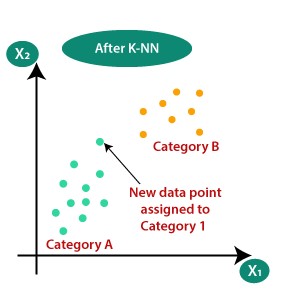


Figure. 4: Final KNN Classifier Outcome

As discussed in the above example the KNN Classifier is an effective machine-learning algorithm for classification problems and works well with all kinds of categorical dataset.

# Model Construction and Classification Process

## Base model prototype with hard-coded sequnce

## The enchace version of KNN Classifer

1. Inout and training phase

TP AND Multisequnce learning things

1. Desigin mdel

Clases