

K - Nearest Neighbor Project

Project Summary

The project is implementing a k-fold cross-validation for a One Nearest Neighbor (1NN) classifier, with some data preprocessing steps such as cleaning, transforming, and preparing raw data standardization and normalization. The code also includes some visualization functions using seaborn to visualize the data, as well as functions to calculate the accuracy of the classifier.

K-Nearest Neighbor (KNN) is an important algorithm in the field of machine learning and pattern recognition. It is a non-parametric algorithm used for both classification and regression tasks. The algorithm works by finding the K nearest data points to a given query point in a training dataset, and then making predictions based on the labels or values of those neighbors.

KNN is a simple and easy-to-understand algorithm. It does not make any assumptions about the underlying data distribution, and hence can be used in a wide range of applications. Unlike other machine learning algorithms, KNN does not require any training data to learn from. It only requires a labeled dataset to make predictions. KNN can be used for both classification and regression tasks. It is also applicable in both supervised and unsupervised learning settings. KNN is an interpretable algorithm, which means that it can provide insights into how it makes its predictions. It is easy to visualize the K nearest neighbors for any given query point, which can help in understanding the decision-making process. KNN is known for its high accuracy, especially when the dataset is small and the number of features is low. It can also handle noisy data and outliers relatively well.

Overall, the project aims to use the K-Nearest Neighbor Algorithm for predicting target attributes by creating training and testing datasets and implementing k-fold cross-validation and 1-NN algorithms using Python libraries and custom Python classes. Along with working with the KNN algorithm, the project will also focus on pre-processing data, debugging, model evaluation, normalization, standardization, and visualization to optimize a data mining algorithm for a real-world dataset.