Prepare note about KNNImputer

## **Introduction**

KNNImputer by scikit-learn is a widely used method to impute missing values. It is widely being observed as a replacement for traditional imputation techniques.

In today’s world, data is being collected from a number of sources and is used for analyzing, generating insights, validating theories, and whatnot. This data collected from different resources may often have some information missing. This may be due to a problem in the data collection or extraction process that could be a human error.

Dealing with these missing values, thus becomes an important step in data preprocessing. The choice of method of imputation is crucial since it can significantly impact one’s work.

Most statistical and machine learning algorithms work on complete observations of a dataset. As a result, it becomes essential to deal with missing information. A handful of literature in statistics deals with the source of missing values and ways to overcome the issue. The best way is to impute these missing observations with an estimated value.

In this article, we introduce a guide to impute missing values in a dataset using values of observations for neighboring data points. For this, we use the very popular KNNImputer by scikit-learn k-Nearest Neighbors Algorithm.

## **The Problem of Degrees of Freedom**

Missing values in a dataset can be a hornet’s nest for any data scientist. Variables with missing values can be a non-trivial problem as there is no easy way out to deal with them.

Generally, if the proportion of missing observations in data is small relative to the total number of observations, we can simply remove those observations. However, this is not the most often case. Deleting the rows containing missing values may lead to parting away with useful information or patterns.

In statistical terms, this leads to reduced degrees of freedom as the number of independent pieces of information goes down.

## **A shared sense of identity (Essence of kNN algorithm)**

Univariate methods used for missing value imputation are simplistic ways of estimating the value and may not provide an accurate picture always. For example, let us say we have variables related to the density of cars on road and levels of pollutants in the air and there are few observations that are missing for the level of pollutants, imputing the level of pollutants by mean/median level of pollutants may not necessarily be an appropriate strategy.

In such scenarios, algorithms like k-Nearest Neighbors (kNN) can help to impute the values of missing data. Sociologists and community researchers suggest that human beings live in a community because neighbors generate a feeling of security and safety, attachment to community, and relationships that bring out a community identity through participation in various activities.

A similar imputation methodology that works on data is k-Nearest Neighbours (kNN) that identifies the neighboring points through a measure of distance and the missing values can be estimated using completed values of neighboring observations.

#### Example-

Suppose, you run out of stock of necessary food items in your house, and due to the lockdown none of the nearby stores are open. Therefore, you ask your neighbors for help and you will end up cooking whatever they supply to you. This is an example of imputation from a 1-nearest neighbor (taking the help of your closest neighbor).

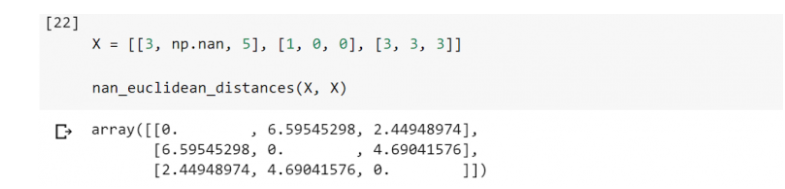
Instead, if you identify 3 neighbors from whom you ask for help and choose to combine the items supplied by 3 of your nearest neighbors, that is an example of imputation from 3-nearest neighbors. Similarly, missing values in datasets can be imputed with the help of values of observations from the k-Nearest Neighbours in your dataset. Neighboring points for a dataset are identified by certain distance metrics, generally euclidean distance.

* [Introduction to k-Nearest Neighbor](https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/?utm_source=blog&utm_medium=kNNImputer)

The idea in kNN methods is to identify ‘k’ samples in the dataset that are similar or close in the space. Then we use these ‘k’ samples to estimate the value of the missing data points. Each sample’s missing values are imputed using the mean value of the ‘k’-neighbors found in the dataset.

## **Imputation Approach with KNNImputer**

We will use the**KNNImputer**function from the impute module of the sklearn. KNNImputer helps to impute missing values present in the observations by finding the nearest neighbors with the Euclidean distance matrix.

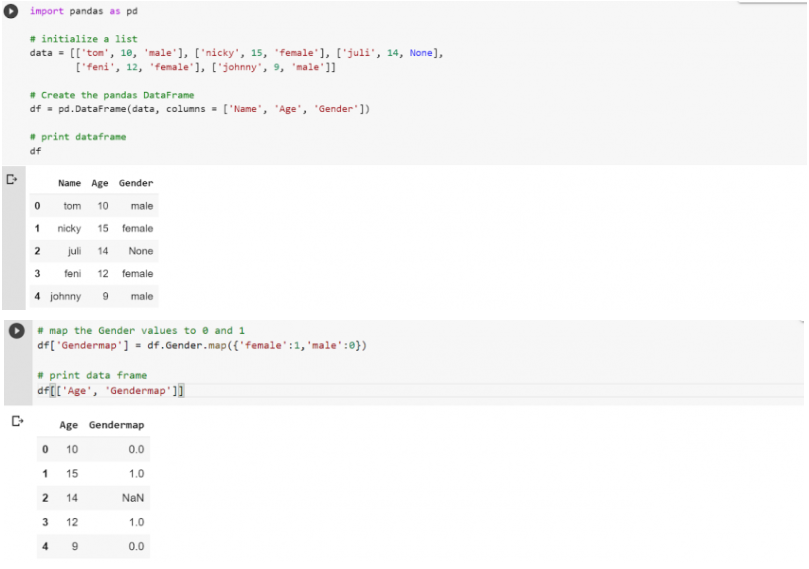


In this case, the code above shows that observation 1 (3, NA, 5) and observation 3 (3, 3, 3) are closest in terms of distances (~2.45). Therefore, imputing the missing value in observation 1 (3, NA, 5) with a 1-nearest neighbor will give an estimate of 3 which is the same as that of the second dimension of observation 3 (3, 3, 3).

Further, imputing the missing value in observation 1 (3, NA, 5) with a 2-nearest neighbor will give an estimate of 1.5 which is the same as the mean value of the second dimension of observations 2 and 3, i.e. (1, 0, 0) and (3, 3, 3), respectively.



Till now, we discussed missing value treatment using kNNImputer for continuous variables. Below, we create a data frame with missing values in categorical variables. For imputing missing values in categorical variables, we have to encode the categorical values into numeric values as kNNImputer works only for numeric variables. We can perform this using a mapping of categories to numeric variables.





## End Notes

In this article, we learned about the missing value, its reasons, patterns, and how you can use KNNImputer to impute missing values. Additionally, you may go through these resources to understand the concept of KNN better-

To summarize, the choice of k to impute the missing values using the kNN algorithm can be a bone of contention. Furthermore, research suggests that it is imperative to test the model using cross-validation after performing imputation with different values of k. Although the imputation of missing values is a continuously evolving field of study, kNN act as a simple and effective strategy.