**Assignment - I**

**Question1 : Steps needed to perform imputation on a dataset.**

There are 3 types of ways, normally imputation is used

1. Imputation using Statistical methods(Mean, Median and Model)
2. Imputation using Machine Learning Algorithms
3. Imputation using Datawig Package
4. **Imputation using Statistical methods**

Imputation is one of the most common practices usually followed for treating the missing values in the dataset. For using statistical method for imputation distribution of the data need to be checked before replacing the missing values. Data could be sometimes skewed to left or right, or it could be symmetrically distributed. Also, for numerical and categorical variables we use different approaches. Mode is usually used for categorical variables, mean and median are considered for treating missing values in numerical variables. If the data is normally distributed mean is used replace the missing values. If the data is asymmetrical using median makes more sense.

1. **Imputation using Machine Learning Machine Learning Algorithms**

Machine Learning algorithms are strong tools for solving any complex computational tasks, likewise it can also be used in treating outliers in data analysis. The first step of imputation is to create subset of the original dataset. After that choose a Model generating algorithm like Simple Imputer for example. Train a machine learning model then using trained model we will be predicting missing values of the original dataset.

1. **Imputation using Datawig Package**

The first step of the Datawig, install the package from the Datawig official documentation page, Secondly, data is loaded. After that imputer needs to be initialized. Fitting the model is the next step. Once the models are fitted, data will be trained using the model. Like other machine learning algorithms, trained model, testing will be carried on the rest of the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl No** | **Student Name** | **English** | **Irish** | **Maths** | **Results** |
| 1 | Robert | 96 | 66 | NA | Pass |
| 2 | Vanda | 68 | 70 | 0 | Fail |
| 3 | Ken | NA | 45 | 56 | Pass |
| 4 | Robin | 90 | NA | 48 | Pass |
| 5 | Mrunali | 56 | 50 | 51 | Pass |
| 6 | Kate | 98 | 74 | 29 | Fail |
| 7 | Lidija | 0 | NA | NA | Fail |

Suppose we are working with the above dataset, First, the data need to be described and the target variable to be established. As per the dataset we have 7 records and 5 features. Please note that this dataset is only for the example, practically we take n\*100 or n\*1000 records for empirical analysis where n is the number of features. Now we know that ‘results’ is categorical missing values in this feature needs to be treated with mode in statistical imputation method. English, Irish, Maths are continuous numerical variable in which the missing values will be treated with mean or median

we will make a distplot or barplot from the data and see wat kind of distribution is the data is having.