Evaluation Portfolio Submission

JASTI CHENNA KESAVA

Data analyst intern

### SIGNIFICANCE:

Big data is a stream known to evaluate,manage,maintain and observe data as well as patterns in data,in order for this analysis of data a huge amounts of raw data is needed and is a must to maintain (or) process data to make predictions in order to manage the company’s historical data, this task is done by us data analysts to analyse the data based on its importance,one on the most important requirement in this process is a dataset.

### PROBLEM EVALUATION:

We know the significance of a “DATASET” in a analysis process but a small error in a dataset causes the whole algorithm to collapse,the dataset can be of many types of data:

1)Structured data

2)Semistructured data

3)Unstructured data

Considering the most famous,Structured datasets one of the main cause off errors in the dataset is the existence of Missing values There are multiple ways to deal with this situation a few Include as below:

### SOLUTION:

Coming to the solutions for this problem here are a few:

->DELETION OF THE ROWS

->MEAN/MEDIAN/MODE IMPUTATION

->MACHINE LEARNING MODELS

->LINEAR REGRESSION

#### **Deletion of rows:**

If the dataset is in a tabular or comma separated format we can simply remove the row that contains the missing value,this can be done as a quick solution but is not recommended as the missing values can be not random and the rows following the missing row may have missed a few values

But if the no of rows with missing values is significantly high in a dataset it is hard to reach an more precise analysis/results

#### **Mean/median/mode Imputation:**

Now coming to the next method we can replace the missing values with a mean/median/mode of the column of the dataset this ensures that the dataset has no loss and will help in the analysis of the dataset significantly,however it is still a dispute between what to replace the missing value with out of the three

#### **Machine learning models:**

However there are many algorithms to choose from which will replace the missing value with a suitable replacement some of these methods are:

**Random forest:**

Random Forest is an ensemble learning method that builds multiple decision trees during training. Each tree "votes" on the imputation of missing values, and the final imputation is based on the aggregated results.

**KNN:**

KNN imputation involves finding the k-nearest neighbours for a data point with missing values and imputing those missing values based on the values of its neighbours. The imputation is done by averaging or using a weighted average of the neighbours' values.

Linear regression:

Linear regression models the relationship between a dependent variable (the variable with missing values) and one or more independent variables (other variables without missing values).