## Method summary:

The K\_ Nearest Neighbours (K-NN) classifier was used as one of the models to train on and validate the performance of the imputation techniques. We used the K-NN classifier in the R studio and Orange to determine how K-NN responds to the various data sets namely, original training sets, original testing data set, mode data at 10, mode at 40, mode at 70, naïve bayes at 10, naïve bayes at 40, naïve bayes at 70 of the public nursery data set.

The implication of the K-NN is that the model is a simple model that works for both categorical and numerical data sets. This implementation, therefore, required that the data sets be handled in a way the k- NN algorithm could provide outputs into whether the algorithm could effectively learn the data set and predict outcomes through the performance metrics of the accuracy and the F1 score. The K-NN was therefore compared to the control and between the different imputed data of the same Nursery data set. The below steps were the coding methodology utilised in R studio to build the K-NN.

* Importing of the data sets,
* Converting the character data into a factor,
* Handling missing values if any are identified,
* Training the data and testing data sets,
* Determining the optimal value of k,
* Creating training and testing predications,
* Developing the K\_NN models from the trained, test and target vectors, and
* Calculating the performance metrices and the confusions matrix.

All the above steps were taken for each of the data sets mentioned above, however noteworthy without ease, particularly for the mode 40 data set, and the naïve bayes data sets at the 40, 70 percentages of imputation.

We then trained the model on the various datasets and tested against the non-imputed base table. For each dataset we recorded the accuracy and F1 score, which was then used to compare model performance across different datasets.

We trained on the base dataset (no missing values induced, or missing values imputed) and report the accuracy and F1 score for this model as the baseline values. These baseline values are then compared to the paired imputed datasets (Same % of missing values imputed using Mode value and Naïve Bayes) against the K-NN Accuracy and F1 score values.

LIMITATION

Using R Studio for this exercise posed several challenges, primarily due to encountering missing values in the target variable (class) within the original training dataset, as well as in other datasets, including other variables. These missing values were unexpected, considering the initial collation of the dataset was done without any such gaps. Addressing these missing values became necessary as K-NN generated error messages due to their presence, preventing the application of the algorithm.

The occurrence of missing values in R Studio could be attributed to various factors. Firstly, there might have been issues with data file encoding, leading to misinterpretation of certain characters or symbols. Secondly, incorrect interpretation of data types by R, such as numeric values being stored as character strings, could have contributed. Additionally, special characters or non-standard symbols might have caused misinterpretation. Lastly, file corruption or data file structure issues could have played a role in R misreading the data.

All the above issues were resolved using and it was found that the missing values resided in the target variable (“”class”) due to the misspell of the class “recommend” versus “recommend.” This perhaps indicated the necessity for clean data particularly when using the K-NN algorithm as it is certainly sensitive to missing values since it relies on measuring the distance between to classify them. Therefore, missing values will significantly impact the outcome of the algorithm, which was experienced where the k\_NN algorithm would not run until missing values were handled.

## Findings and observations:

In the report, it's crucial to summarize the significance of accuracy and F1 score evaluation metrics, particularly in the context of classification tasks, as they offer distinct insights into model performance. Accuracy in machine learning denotes the ratio of correctly classified instances to the total instances, serving as a measure of overall correctness. On the other hand, the F1 score represents the harmonic mean of precision and recall, both essential evaluation metrics. Ranging from 0 to 1, a higher F1 score indicates better model performance.

Upon examining the accuracy levels across different datasets with varying proportions of missing values and having resolved the missing values in the data sets, accuracy is at a 99% level for all the datasets. Which means that K-NN correctly classified the target variable of class. Similarly, the F1 score also indicated a value of 99% across the data set which indicates that the K-NN model performs better. In comparison to the Classification Tree model, it can be said that the K-NN performed better with regards to its effectiveness to predict the target variable of class for each of the variables.