**Data Science Post Block Assignment 3: Task A**

Sarel Vermaak [190980235], Kebaabetswe Tlhoaele [28816749], Lize Mostert [23537140],

Difedile Rasenyalo [28294882] & Isabel de Waal [20805055]

1. **Introduction** 
   1. **Background**

Missing data is a common data quality issue encountered during the data exploration and cleaning phases. If features have hissing values, it is important to determine why these values are missing. Often missing values can creep in during data integration, in these cases, the integration errors can be fixed to resolve the missing value issues [1]. On the other hand, missing values can also be introduced during the data generation or collection phases – these are more difficult to deal with.

There are various ways to deal with missing values. One is to simply remove instances or features that contain them. This is not the best approach as it could lead to the loss of valuable information and may lead to bias during inference [2]. Another method is to convert the missing value into a new feature, but this has shown to lead to serious inference problems [3 according to 2]. Finally, one can impute the missing values. Imputation is generally a good idea if a significant portion of the data contains missing values for a few features. Generally imputation above 60% of missing values are not recommended [1].

There are different imputation methodologies, of which we will be focusing on the following two:

*Baseline imputation (data driven):* Imputing the variable simply based on the mode(categorical) or mean/median(numerical).

*Model Imputation:* a model is used to perform parameter estimation.

Studies have found that the ‘best’ imputation method depends on the type of data in the dataset, the proportion of missing values and even the distribution of the features have an influence on the type of imputation that is considered ‘best’ [2].

* 1. **Objectives**

**Task A:** The objective was to evaluate the effectiveness of a baseline imputation versus that of a Naïve Bayes imputation on the performance of classification models. For this study we also decided to look at the influence of the proportion of missing values on the effectiveness of the imputation method. This will be done by comparing the performance of two classification models who have been trained on the control data and the imputed data and then tested with unchanged data.

1. **Methodology**

In this study, we leveraged R Studio and Python as the primary tools for imputation, modelling, and evaluation. The methodology involved several key steps:

**Data Exploration:** The first phase of the analysis involved exploring the Nursery dataset to gain a comprehensive understanding of their structure, variables, and content. As the main aim of this experiment is to evaluate the effectiveness of two different imputation methods, it was essential that the data be as clean as possible, as not to confuse any possible insights.

**Data Pre-processing:** For this experiment, the pre-processing consisted of splitting the data into a 30:70 test: train split. The testing dataset was then reserved for later use. Three copies of the training data-subset were made and were induced with missing values in varying proportions (10%, 40% and 70%). These together with the training control was then used during the modelling.

**Modelling:** K-Nearest Neighbours Classifier and a Classification Tree was used to evaluate the effectiveness of the data Imputation.

k-Nearest Neighbour: [insert chosen k and method here]

Classification tree: [insert parameter settings here]

**Visualization:** Bar plots were constructed in Microsoft Excel 365 to clearly visualise the difference in performance of the two models when trained with data that was imputed in two different ways.

1. **Results and Discussion**

Generally, the results were quite interesting. Figure 1 below shows the performance metrics of each of the classification tree for each imputation method per proportion of missing values. The first thing to notice is that the Mode and NB imputation methods resulted in comparable model performances, these performances in turn was very comparable to the control, indicating that the imputation techniques was ‘sucsessful’.

*Figure 1 A&B:* Bar charts visualising the Accuracy (A) and F1 Scores(B) for the classification tree model.

1. **Conclusion**

In conclusion, the analysis of the

1. **References**
2. AML textbook, page 64
3. Impact of imputation of missing values on classification error for discrete data - article
4. J.L. Shafer, Analysis of Incomplete Multivariate Data, Chapman and Hall, London, 1997.