**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 poses a common challenge in the data exploration and cleaning phases, warranting careful attention to its origins [1]. While integration errors often introduce missing values, issues during data generation or collection further complicate the matter.

Addressing missing values involves several strategies. Simply removing instances or features containing missing data risks losing valuable information and introducing bias during inference. Transforming missing values into a new feature can also lead to inference challenges. Imputation, however, offers a promising solution, particularly when a significant portion of the data is affected [2].

This study delves into the impact of various imputation methods on classifier performance when applied to imputed data. Specifically, it examines the effects of mode imputation and Naïve Bayes classifier imputation. Through the evaluation employing widely used classifiers such as K-nearest-neighbours and decision tree classifiers, the study aims to assess the effectiveness of these imputation techniques.

* 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 utilized R Studio and Python as our primary tools for imputation, modelling, and evaluation. The methodology encompassed several essential steps:

**Data Selection:** The initial phase involved selecting a publicly available dataset with documented research on handling missing feature values. Our choice rested on the nursery dataset utilized in a study by A. Farhangfara, L. Kurganb, and J. Dyc, providing a solid foundation for assessing the effectiveness of our proposed imputation methods.

**Data Exploration:** Exploring the nursery dataset to gain a profound understanding of its structure, variables, and content. Given our experiment's primary focus on evaluating two distinct imputation methods, it was imperative to ensure the data's cleanliness to avoid obscuring potential insights.

**Data Pre-processing**: Pre-processing entailed splitting the data into a 30:70 test: train split, with the testing dataset reserved for subsequent analysis. Additionally, three copies of the training subset were generated and induced with missing values at varying proportions (10%, 40%, and 70%). These, alongside the pristine training control, formed the basis for our modelling endeavours.

**Imputation Approach:** The study employed mode imputation for numeric value imputation, serving as a reference for comparison. Additionally, we leveraged the Naïve Bayes imputation technique algorithm to impute missing numerical feature values.

**Model Comparison**: To evaluate the efficacy of our imputation approaches, we employed both decision tree and K-Nearest Neighbours models. Each model underwent training using both imputation methods, and their performance metrics—accuracy, precision, recall, and F1-score—were meticulously assessed and compared.

**Visualization**: Line plots were crafted to vividly illustrate the performance disparity between the two models when trained with data imputed using distinct methods.

1. **Results and Discussion**

After pre-processing, the optimal number of clusters were to be selected for the K-means clustering.

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.