**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

The simplest way to deal with missing values, is to simply remove instances or features that contain them. If the proportion of instances containing the missing values are small enough, then removing them is acceptable. The issue arises however if the dataset is not that big to begin with

Some experience is also required to know when to impute the missing data, when to remove a feature or instance. A general rule could be to leave it as is if the missing values are <=10%, impute if the missing values are 30-60% and remove the feature/variable if the missing values are >60%[reference].

There are various methods to impute missing data. These include:

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

*Model Imputation:* This involves using a predictive model to predict what the missing value should be. These include Kaive Bayes and k-NN.

*Ensemble Imputation:* xxx

Much research has been done of the different methods of data imputation[reference]. Studies have found thet 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’

* 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.

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 two provided datasets to gain a comprehensive understanding of their structure, variables, and content. This exploration aimed to identify any potential data inconsistencies, missing values, or outliers that could affect the quality of the analysis. This was done primarily by calculating the count, cardinality and % missing values of the dataset.

**Data Pre-processing:** Following the data exploration phase, the datasets underwent thorough cleaning and pre-processing. This step involved handling missing values,.

**Modelling:** K-Nearest Neighbours

**Visualization:** We

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