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

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

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

Missing values are a common occurrence in datasets. It's not unusual to encounter datasets where a significant portion, sometimes up to half, of the entries are missing. Such datasets pose challenges for data analysis methods that require complete data. To address this issue, imputation methods are often employed to fill in the missing values. This paper investigates how different imputation methods impact the performance of classifiers when applied to imputed data. The study examines the effect of missing data imputation using two imputation methods, mean and K-NN imputation. The evaluation employed two popular classifiers, K-nearest-neighbour, and decision tree classifiers.

Note: perhaps mention something specific regarding the imputation of numerical data

* 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.

**Task B:** Building upon the groundwork of Task A, this task aims to use the K Nearest neighbours’ classifier to address the missing numerical value. Additionally, the effectiveness of this approach will be compared against a baseline numerical value imputation method using two distinct machine learning models.

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:

* 1. **Dataset Selection**

The first step entails selecting a publicly available dataset with documented research on handling missing feature values. This dataset served as the cornerstone for assessing the effectiveness of the proposed imputation methods. For this study, we selected the nursery dataset utilized in a study conducted by A. Farhangfara, L. Kurganb, and J. Dyc. [No, for the MEng portion we used the ‘Adult’ dataset.]

* 1. **Introducing Missing Values:**

Missing data were randomly introduced into each of the datasets using the Missing Completely at Random (MCAR) mechanism. The missing values were added to all attributes across all datasets at three different rates: 10%, 40%, and 70%.

* 1. **Imputation Approach**
* **Numeric value Imputation:** Numeric value imputation approach was implemented to establish a reference for comparison. [Mean imputation?]
* **K Nearest Neighbours Imputation:** The K Nearest Neighbour was employed to impute missing numerical feature values.
  1. **Model Comparison:**

The decision tree algorithm and the K- Nearest Neighbours were employed for this study, with each model trained using both imputation approaches. The models' performance metrics, including accuracy, precision, recall, and F1-score, will be evaluated and compared.

1. **Results and Discussion**

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

According to [*A comparison of* imputation methods using machine learning models ] the type of dataset is an important factor when deciding on a imputation method. They found that k-NN worked better for some datasets than others. Conversely [Comparison of Performance of Data Imputation Methods for Numeric Dataset] found that the type of dataset did not have an influence on imputation performance. They did however find that the k-NN imputation generally outperformed the mean imputation. These and other studies [ref numbers], have found that the proportion of missing data also does not influence the imputation performance. This is in contrast to what we found. This could be because…

1. **Flowchart Development**
2. **Conclusion**

In conclusion, the analysis of the

1. **References**