Data Science Post Block Assignment 2: Task B Report

Datasets:

General comments on the dataset used.

* The dataset used is the Concrete Compressive Strength dataset from the UCI machine learning repository (<https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength>).
* Number of instances: 1030
* Number of Attributes: 9
* Attribute breakdown: 8 quantitative input variables, and 1 quantitative output variable.
* Missing Attribute Values: None
* The data is in raw from, not scaled
* All of the quantitative input variables are in the same unit (where applicable)
* kg in a m3 mixture
* Ager is measured in days
* The output Concrete compressive strength is measured in MegaPascal (MPa)

Possible pre-imputation cleaning steps

Scale data.

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

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

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

Many of the recent studies have focused on the imputation of categorical variables

* 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**