**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.

* 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 Python as the primary tool 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 numerical feature values. This dataset served as the cornerstone for assessing the effectiveness of the proposed imputation methods. For this study, we selected the ‘Adult’ dataset utilized in a study conducted by A. Farhangfara, L. Kurganb, and J. Dyc.

* 1. **Data Cleaning and pre-processing:**

The ‘Adult’ dataset was thoroughly evaluated for any possible data quality issues that may have an influence on model performance. The original dataset contained a mixture of numerical and categorical values. As this portion of the project is only concerned with numerical value imputation, only the numerical features were selected moving forward. The data was also in very different ranges; and because the distributions were non-Gaussian a min-max scaling approach was used to normalise them. After cleaning dataset was randomly divided into a 30:70 split forming the testing and training sets respectively. The testing set was kept constant throughout the experiment, serving as the testing set for all the training subsets. The training set was then copied six times, to form the basis to introduce the missing values.

* 1. **Introducing Missing Values:**

Missing data were randomly introduced into each of the copied datasets using the Missing Completely at Random (MCAR) mechanism. The missing values were added to instances across all features (including the target) at three different rates: 10%, 40%, and 70%. There were two copies of each missing value level.

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

A Decision Tree classification algorithm and the K- Nearest Neighbours were employed for this study. The original training set and then each of the imputed training sets were separately trained and then tested with these two models. The optimal tree depth and the optimal number for k was determined using cross validation, and choosing the parameter that delivered the best results. The models' performance metrics, including accuracy and F1-score, will be evaluated and compared.

1. **Results and Discussion**

Looking at the results left (Figure 1A&B) and comparing the effectiveness of mean (medium blue and purple bars) and k-NN (light purple and light blue bars) imputation at different levels by observing the performance of a Classification Tree(A) and that of the k-Nearest neighbour(B). Overall, the same pattern is repeated. At a low level of missing values (10%), the mean imputation and the k-NN imputation have a similar performance to each other and to the control (represented by the darkest purple and blue bars). What is very interesting is how drastically the performance of the k-NN imputation method decreases as the proportion of missing values increase. Even more interesting is how ‘reliable’ the basic mean imputation was throughout the different proportions of missing values. This could purely be due to the structure of the data. Another possibility could be that prior to imputation the numeric features of the dataset was normalised (using min-max scaling as the distributions were not gaussian). This normalisation may have had an influence on the calculated means. It is interesting to note that the pattern mentioned above is the same in both models used for evaluation.

***A***

***B***

According to Suh & Song, the type of dataset is an important factor when deciding on a imputation method [1]. They found that k-NN worked better for some datasets than others. Conversely Jadhav et.al. found that the type of dataset did not have an influence on imputation performance [2]. They did however find that the k-NN imputation generally outperformed the mean imputation. Comparing our results to those of Suh & Song (who also used the ‘Adult’ dataset in their analysis) we found that our results were different. In their study the k-NN performed better than the mean imputation [1]. The difference is probably due to the different performance evaluation methods used. Where we evaluated the effectiveness of the imputation by evaluating the performance of two classification models Suh & Song constructed their own performance metric (the imputation performance metric – IPM), specifically designed to evaluate the effectiveness of imputation. This metric can handle mixed datasets of both categorical and numeric features.

*Figure 1A&B:* Bar charts visualising the performance metrics (Accuracy and F1 scores) per level of imputation (10%, 40% and 70%) per imputation method (Control, Mean and k-NN) for the Classification Tree(A) and *k*-Nearest Neighbour(B) algorithms. Accuracy scores are indicated by the purple bars and the F1 scores are the blue bars.

1. **Flowchart Development**

Below is a compact version of the flowchart developed. A full sized version can be found attached to the end of this report.

1. **Conclusion**

In conclusion, the analysis of the analysis of the effectiveness of two different single imputation methods indicated that in some cases the simplest mean imputation may sometimes be the best. To get a clearer picture of this, it is recommended that the study is repeated using multiple datasets, and a variety of ways to calculate the effectiveness and performance of the imputation method. This study further confirmed what was seen in Task A: that the effectiveness of the imputation method depends on the dataset and purpose of the study. There is no single ‘best’ imputation method.

1. **References**
2. H. Suh and J. Song, “A comparison of imputation methods using machine learning models,” *Communications for Statistical Applications and Methods*, vol. 30, no. 3, pp. 331–341, May 2023. doi:10.29220/csam.2023.30.3.331
3. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of performance of data imputation methods for numeric dataset,” *Applied Artificial Intelligence*, vol. 33, no. 10, pp. 913–933, Jul. 2019. doi:10.1080/08839514.2019.1637138