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**

The field of data science

* 1. **Objectives**

**Task A:** The objective was to analyse the job postings dataset scraped from LinkedIn, to derive insights into the data science job market.

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, text manipulation, standardizing text formats, ensuring data consistency and dimensionality reduction. The text data had to be transformed into a corpus and then a term document matrix, for it to be modelled.

**Modelling:** K-Means clustering is a used an unsupervised machine learning algorithm for partitioning a dataset into a predetermined number of clusters [2]. In the context of analysing job skills, K-Means was employed to group similar skills together. The Elbow Method is a heuristic that was used to determine the optimal number of clusters (k) in a dataset [3]. In this context the Elbow method was employed to obtain the optimal number of clusters. This proved to be an ineffective method of determining the optimal number of clusters. We then iteratively modelled the data using the range of k’s and evaluated the quality of the clusters using silhouette scores [4]. These silhouette scores were also used to evaluate the quality of the final clustering.

**Visualization:** We proceeded to utilise R Studio's visualization capabilities to visualisations to illustrate patterns, trends, and relationships within the data. The dimensionality of the dataset had to be reduced to effectively visualise the clustering of the data. Principle Component Analysis was used to facilitate this. The visualisation was done at the exploration, modelling and at the insight generation phases.

**Insight Generation:** The final step involved deriving insights from the visualized data. These insights provided valuable information in formulation the recruitment strategy.

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 data science job market provided valuable insights into skill requirements, emerging trends, and industry preferences. By leveraging these insights, recruitment agencies and organizations can optimize their recruitment strategies, improve talent acquisition processes, and cultivate a skilled workforce capable of driving innovation and success in the data-driven era.

1. **References**
2. P. M. &. A. Effendi, “Examining the Performance of K-Means,” International Journal of Research in Engineering, Science and Management, vol. 1, no. 3, pp. 1-2, 2018.
3. B. K. K. E. M. S. R. a. B. D. S. M A Syakur, “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster,” IOP Conference Series: Materials Science and Engineering, vol. 336, pp. 1-2, 2017.
4. K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096.