CMP3749M Big Data Assessment Item 1

# Task 1 – PySpark Analysis of Nuclear Plants dataset

## Part 1

**Firstly, the required libraries and modules were imported, and a Spark session was created. A function named ‘clean\_data’ was created, which takes in a data frame for cleaning. It checks for duplicate rows or missing/null values and drops the rows on which they occur. The dataset was read into a Spark data frame and then cleaned using the aforementioned function. It was found that there were no duplicate or missing/null values in the dataset, with the number of samples being 996, before and after cleaning.**

**Incomplete datasets can have negative effects on the reliability of ML predictions. Dropping rows with missing/null or duplicated values is only one way of handling this task. Other methods, such as imputing data or data prediction using machine learning (ML), could have been used in the analysis of the dataset. Imputation refers to filling missing values with substituted data (Li et al., 2015). Simple strategies include replacing the values with the mean, median, or mode of that column/row. However, while this method is easy to implement, it can introduce bias into the applications of the data (Jäger et al., 2021).**

**More sophisticated ML methods, such as k-nearest neighbours, can be used for imputation. These methods can predict qualitative data, such as the ‘Status’ column in the nuclear plant's dataset, and quantitative data by changing the parameter of the distance metric. This method can be slow, as it must go through the whole dataset (Batista and Monard, 2003). This can be a huge drawback when dealing with billions of rows and columns within big datasets, due to the computational cost of processing the entire dataset.**

## Part 2

## **Two functions were created to complete the task. A function named ‘filter\_df’ takes in a data frame and filters it based on the ‘Status’ column within the data frame. Another function, ‘summary\_stats’, takes in a data frame and uses the previously mentioned function to filter the ‘normal’ and ‘abnormal’ statuses, calculating the summary statistics using a built-in PySpark function. Finally, a function named ‘boxplot\_df’ takes in the data frame and filters it using the ‘filter\_df’ function, creating two boxplots, one for each group: ‘normal’ and ‘abnormal’. Both boxplots give a good visual analysis of the data, showing outliers and clear differences in the values for ‘vibration\_sensor\_3 for each group, “Normal” and “abnormal”, possibly being the reason for the status changing. The outliers were not removed from the dataset as it was not required for the tasks but there are various methods to do so. Trimming, which removes the row where the outlier appears and winsorization which sets extreme values of data to a designated percentile within the dataset (Kwak and Kim, 2017).**

## Part 3

A correlation matrix was created using pandas library and visualised using seaborn library. Two features, ‘pressure\_sensor\_4’ and ‘power\_range\_sensor\_4’ have a Pearson coefficient of 0.82. This indicates a high correlation between the two features and there may be redundancy in using both features for analysis as they could carry similar information about the target class, this could result in increased storage requirements and reduced computational performance.

# Task 2 – MapReduce for Margie Travel dataset

## Part 1

## Part 2

## Part 3

# Task 3 – Big Data Tools and Technology Appraisal

# References

Batista, G.E.A.P.A. and Monard, M.C. (2003) An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence,* 17(5-6) 519-533. Available from <https://doi.org/10.1080/713827181> [accessed 06/01/2024].

Jäger, S., Allhorn, A. and Bießmann, F. (2021) *A Benchmark for Data Imputation Methods.* Frontiers Media SA. Available from <https://doi.org/10.3389/fdata.2021.693674> [accessed 06/01/2024].

Kwak, S.K. and Kim, J.H. (2017) Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology,* 70(4) 407-411. Available from <https://doi.org/10.4097%2Fkjae.2017.70.4.407> [accessed 07/01/2024].

Li, P., Stuart, E.A. and Allison, D.B. (2015) Multiple Imputation: A Flexible Tool for Handling Missing Data. *JAMA,* 314(18) 1966-1967. Available from <https://doi.org/10.1001/jama.2015.15281> [accessed 06/01/2024].