**Issue 01**

shuttle_train.data and shuttle_test.data files contain NASA space shuttle information divided into train and test data sets. These datasets contain 9 attributes, all of which are numerical. The last column is the class, which has been coded as follows:

1. Rad Flow
2. Fpv Close
3. Fpv Open
4. High
5. Bypass
6. Bpv Close
7. Bpv

   Open Approximately 80% of the data belongs to class 1. Use logistic regression, SVM, and MLP to create models. Check the accuracy of each model and visualize the predictions using confusion matrices.

**Answer:**

1. The MLP classification method delivered a remarkable accuracy of 99.9 percent.
2. SVM exhibited strong performance with an accuracy of 93.8 percent.
3. Logistic regression demonstrated a solid accuracy of 96.0 percent.

## 1. Introduction

NASA's space shuttle program has generated vast amounts of data, which is invaluable for research and analysis. In this study, we focus on a curated dataset that comprises two subsets: a training dataset and a test dataset. These datasets involve nine attributes, each of which is represented as a numerical value. The final column represents the class labels for the data points. To enhance the efficiency of analysis, these class labels have been encoded into categorical variables, as follows:

a. Rad Flow

b. Fpv Close
c. Fpv Open
d. High
e. Bypass
f. Bpv Close
g. Bpv Open

It is important to note that a significant portion of the data, approximately 80%, falls within the category labeled as "Rad Flow." This dataset presents an opportunity to explore and model complex relationships within the context of NASA's space shuttle operations, with a particular focus on the distribution of classes and their implications for predictive modeling and analysis. The dataset serves as a valuable resource for research aimed at enhancing the understanding of space shuttle operations and predictive modeling applications.

## 2. Tools

The analysis was conducted using Python in combination with several prominent machine learning frameworks and libraries, including numpy, pandas, matplotlib, seaborn, and scikit-learn. For an interactive and collaborative development environment, Anaconda's Jupyter Notebook served as the primary platform for this analysis.

## 3. Preliminary Data Analysis and Visualization

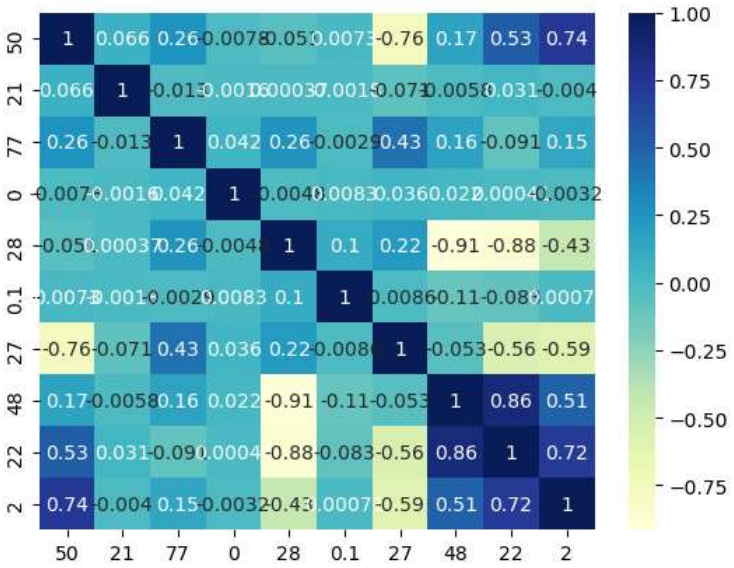
The dataset subjected to examination, commonly referred to as the training dataset, comprises a large 43,499 data rows, each featuring 10 columns. A corresponding test dataset, intended for evaluation purposes, includes 14,499 rows with the same 10 columns. Among these columns, 9

represent unique attributes, while the last column captures 7 distinct features, each thoughtfully encoded to enhance the analysis.

The central objective of this attempt was to construct classification-based machine learning models and then compare their performance against each other. Such a task is essential for understanding the dataset's essential patterns, predicting class labels effectively, and evaluating the predictive efficacy of different models.

### 3.1. Correlation analysis and feature reduction

The dataset exhibits a notable bias, with an vast majority, approximately 80%, concentrated within class 1. Therefore it is required resampling methods to reshape the data set. This bias needs the application of data smoothing techniques to balance the representation of different classes. Moreover, an analysis of correlations has revealed strong interrelationships between multiple features.




To address this, principal component analysis (PCA) is essential for identifying the most important features and enabling the elimination of redundant ones. The advantages of these data preprocessing decisions are various. They include a significant reduction in computational complexity, mitigating the risk of over fitting, and enhancing the ease of data visualization.

However, it is imperative to acknowledge the potential drawbacks, including the limited risk of losing valuable information and, as a consequence, a potential reduction in predictive accuracy. These trade-offs emphasize the importance of thoughtful data preprocessing to find an optimal balance between model simplicity and predictive performance.

## 4. Machine Learning techniques used and results

This study harnessed a range of powerful techniques, including the Support Vector Machine Classifier, Artificial Neural Network (specifically, the Multilayer Perceptron or MLP), and logistic regression.

Both the training and test datasets were accurately leveraged as different units, development a comprehensive and difficult assessment of each machine learning (ML) and deep learning (DL) method employed.

The main goal behind the application of these methodologies is their collective operation in addressing the research objectives.

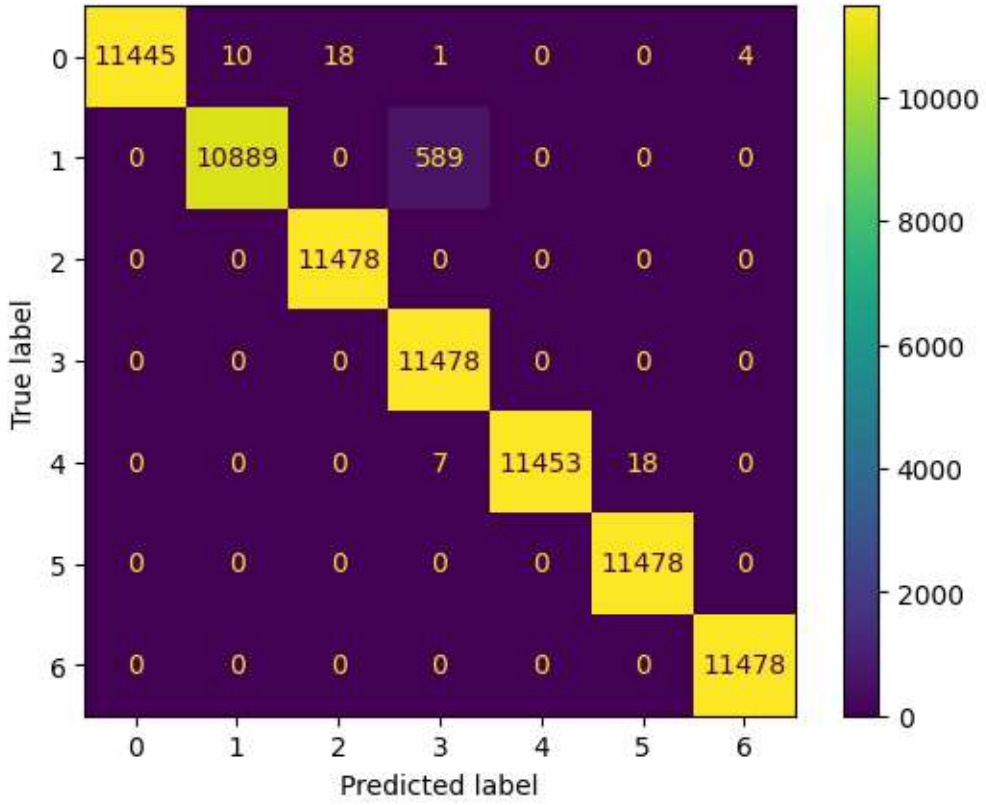
### 4.1. Artificial Neural Network

For the classification of the dataset, the Multilayer Perceptron (MLP) method was employed. This MLP architecture consists of six hidden layers and was subjected to a maximum of 2000 iterations to ensure convergence.

The results of this classification endeavor are noteworthy. The accuracy achieved by the MLP classifier method is an impressive 0.9995560320997169, underscoring its effectiveness in accurately categorizing the data.

Both the training accuracy (0.999572785605388) and the test accuracy (0.9919473278072337) reveal the model's robust performance in maintaining high levels of accuracy across different datasets.

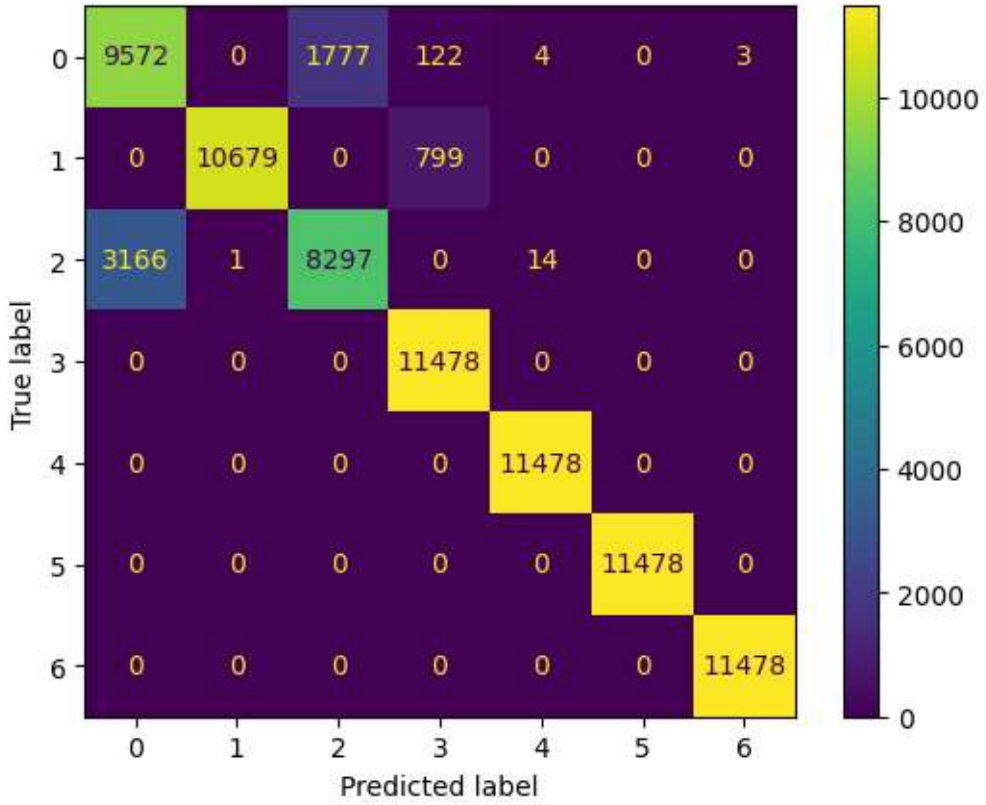
The confusion matrix, which provides a detailed breakdown of the model's classification performance, is presented below for comprehensive evaluation.




### 4.2. Support Vector Classifier

To classify the dataset, Support Vector Machines (SVM) were employed. This approach resulted in a training accuracy of 0.9378863777245389 and a test accuracy of 0.9267418415353595, demonstrating the effectiveness of SVM in accurately categorizing the data.

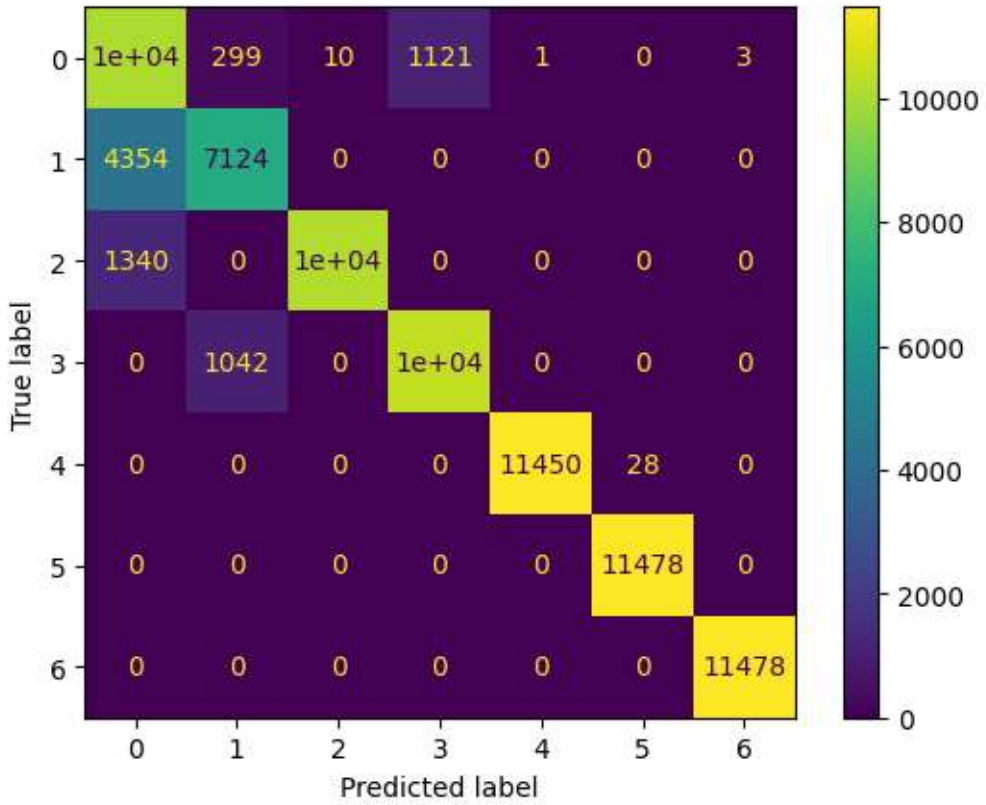
Further insights into the model's performance are provided by the accompanying confusion matrix, which offers a detailed breakdown of the classification results. This matrix serves as a valuable resource for assessing the model's performance and identifying areas for potential improvement.




**4.3 Logistic Regression**

The training phase of this classification model yielded 0.9443699844192397. During the evaluation on the test dataset, it achieved a accuracy of 0.8979662957707913.

To gain a more comprehensive understanding of the model's performance, a confusion matrix is presented below. This matrix provides a detailed breakdown of the model's classification outcomes, enabling a thorough assessment of its strengths and areas for further consideration.




**5. Results and Conclusions**

Impressive accuracy levels were achieved across various classification methods.

Specifically:

4. The MLP classification method delivered a remarkable accuracy of 99.9 percent.
5. SVM exhibited strong performance with an accuracy of 93.8 percent.
6. Logistic regression demonstrated a solid accuracy of 96.0 percent.

These results underscore the capabilities of these methods in achieving high levels of accuracy in the classification task.