**Report 1: Approach and Model Building**

**Introduction** The primary focus of this project is to carry out a comparative analysis of various machine learning techniques applied to classify stars into their respective spectral classes. The dataset is characterized by class imbalances and most of the examples are characterized by very small sizes.

**Importing of Libraries** The project makes use of various Python libraries to simplify data preprocessing, modeling, and evaluation. Pandas and NumPy are essential libraries for manipulating data, while Scikit-learn is used for machine learning tasks like splitting data and evaluating models, and TensorFlow is used to create neural networks. Moreover, SMOTE deals with class imbalance, while Matplotlib and Seaborn aid in data visualization. These tools together help with strong data analysis, training models, and evaluating performance for various algorithms such as Decision Trees, K-Means clustering, and CNNs.

**Data Visualisation** During the initial stages of our analysis, we imported the 'stars\_data.csv' dataset and conducted exploratory data analysis to examine its format and properties, verifying the absence of any null values and recognizing distinct records. Visual aids such as countplots and pie charts pointed out the uneven spread of the 'Spectral Class' target variable, specifically the lack of representation of class G, which only had one sample. This highlighted the importance of utilizing methods such as SMOTE to tackle class imbalance in subsequent phases of the evaluation.

**Data Preprocessing** The dataset comprised a total of 240 stars with details of the different features or properties. The spectral class of each star is to be predicted, which has seven classes. The target variable showed class imbalance thus classes with less than 2 samples were dropped, so as not to run into errors during the models training.

**Class Imbalance** The SMOTE (Synthetic Minority Over-sampling Technique) function was used in the package to handle the class imbalance. This form of over-sampling technique synthesizes samples for the minority class categories so as to avoid biasing the model to selecting the majority class.

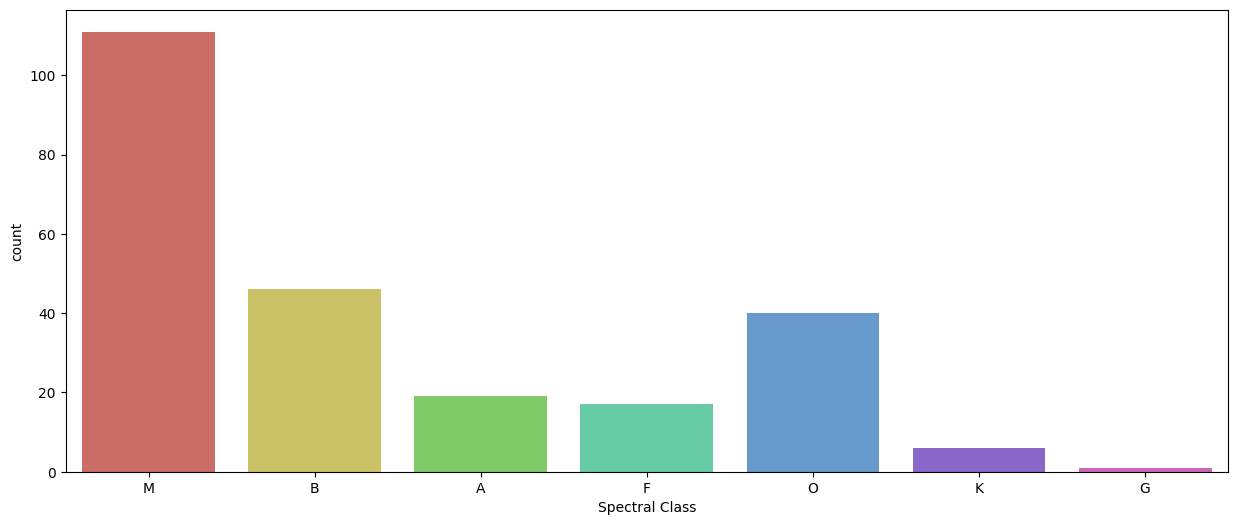
**Model Building** The target in this task was to build four models that can be used in this classification. The models were:

* **Decision Tree Classifier:** A supervised model chosen due to its simplicity and efficiency when handling classification tasks.
* **K-Means Clustering:** This is an unsupervised model used to identify patterns within the data minus prior knowledge of the labels.
* **Multi-Layer Perceptron (MLP) Neural Network:** This is a deep learning model with multiple layers of neurons aimed at learning the complex patterns in the data.
* **Convolutional Neural Network (CNN):** A deep learning model mostly used with image data but the power of its feature extraction mechanisms allows us to adopt this for use in this project.

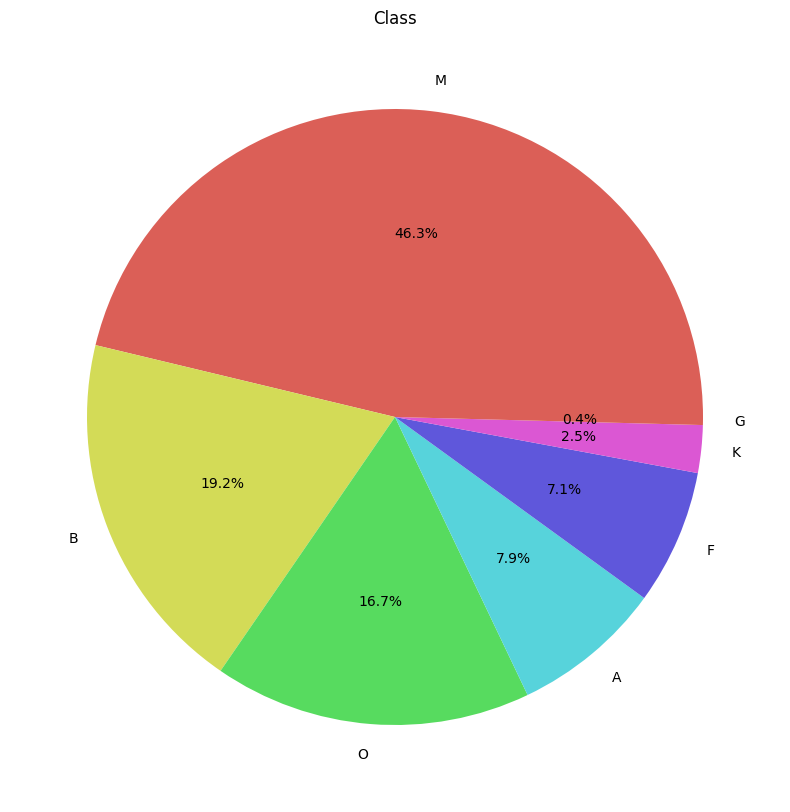
The construction of each model was done carefully in such a way that the data is well prepared and the models fine-tuned to perform tremendously.

**Appendix**

**Visualisation of the Target Class (count plot):**

****

**Pie Chart**

****