Introduction

The strength of any machine learning algorithm often lies in the size and quality of the input data. When the dataset provided is small, it exaggerates the problem, as small datasets usually overfit and create poor generalization in models. The aim of this project was to estimate and compare the performance of different machine learning algorithms under these constraints with precision. This will be done through a keen approach while preprocessing the data and developing the models so that each model is tuned and optimized for this current dataset.

Approach

2.1 Data Preprocessing

1. **Standardization**: Numeric features in the dataset are standardized with StandardScaler to a mean of 0 and standard deviation of 1. This was necessary because a few of the algorithms are sensitive to the scale of features.
2. **Encoding of Categorical Variables**: The categorical features 'Star colour' and 'Spectral Class' have been label-encoded to convert them into numeric format. Because of this, they could be used with machine learning algorithms from hereon.
3. **Sampling data**: Due to the data being highly unbalanced, I discarded the Spectral Class ‘G’ which had only 1 datapoint which did not help with over sampling. SMOTE (Synthetic Minority Over-sampling Technique) helped to balance the class distribution in the training data. This helps to mitigate the issue of imbalanced classes, which can adversely affect model performances.
4. **Dimensionality Reduction**: Principal Component Analysis (PCA) has been administered to the dataset to reduce dimensionality while capturing 85% of the variance. So, the PCA would be useful to simplify the dataset, reduce noise, speed up model training for the clustering algorithms.

2.2 Model Construction

Supervised Learning Models

Logistic Regression, SVC, K-Nearest Neighbours, Decision Tree, Random Forest: All these models were taken up to the stage of implementation using the Sklearn library. Hyperparameter tuning was done using GridSearchCV to find the best configurations of each model.

Unsupervised Learning Models

K-Means, Agglomerative Clustering, DBSCAN, Spectral Clustering, Gaussian Mixture Models: Clustering model attributes have been put into work in order to identify the intrinsic structure in data. Parameter optimization such as the number of clusters for K-Means and linkage criteria for Agglomerative Clustering is done in order to have a superior effect on clustering performance. Silhouette scores and visualizations were used for evaluation.

Neural Networks

Multi-Layer Perceptron (MLP): The following is the model created for the MLP, where there are three hidden layers with 128, 256, and 128 nodes, respectively. Besides, l2 regularizer and an exponential decay schedule was also used to provide a closer rate with the learning rate. An Adam optimizer was created and the loss is sparse categorical cross-entropy.

Convolutional Neural Network (CNN): The architecture chosen for the CNN has 2 Conv1D layers each followed by Max Pooling layers and joined by a 64-node dense layer. This network architecture is thus helpful to capture spatial hierarchies in the data. The learning rate used out of the inverse time decay, and the model was trained on the Adam optimizer with a sparse categorical cross-entropy loss.