**Stellar Classification**

**Milestone: Exploration of Candidate Data Mining Models, and Select the Final Model or Models**

Group 16

Student 1 Sanidhya Karnik

Student 2 Digvijay Raut

617-407-1206 (Tel of Student 1)

857-492-3195 (Tel of Student 2)

[karnik.san@northeastern.edu](mailto:karnik.san@northeastern.edu)

[raut.di@northeastern.edu](mailto:raut.di@northeastern.edu)

**Percentage of Effort Contributed by Student 1: 50%**

**Percentage of Effort Contributed by Student 2: 50%**

**Signature of Student 1: Sanidhya Karnik**

**Signature of Student 2: Digvijay Raut**

**Submission Date: 03/04/2024**

# **Model exploration**

For the classification task in our project, where the goal is to classify astronomical objects into three categories (stars, galaxies, and quasars) based on their spectral and photometric characteristics, several machine learning models can be considered. Below are some commonly used models for classification tasks along with their suitability for this specific problem:

**Logistic Regression:**

Despite its name, logistic regression is a linear model suitable for binary classification tasks. However, it can be extended to multi-class classification using techniques like one-vs-rest or SoftMax regression. For this problem, logistic regression might not be the best choice because it's a simple linear model and may not capture complex relationships between features and classes.

**Decision Trees:**

Decision trees are intuitive and can handle both numerical and categorical data. They can capture nonlinear relationships and interactions between features. However, decision trees tend to overfit the data, especially when the tree depth is not properly controlled.

**Random Forest:**

Random Forest is an ensemble learning method based on decision trees. It constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random Forest generally performs well in practice, handles high-dimensional data well, and is less prone to overfitting compared to individual decision trees.

**Support Vector Machines (SVM):**

SVM is a powerful supervised learning algorithm used for classification tasks. It works well in high-dimensional spaces and is effective in cases where the number of dimensions exceeds the number of samples. SVM aims to find the hyperplane that best separates the classes in the feature space. SVM can be adapted for multi-class classification using techniques like one-vs-one or one-vs-rest. However, SVM might not be the best choice for very large datasets due to its training time complexity.

**Gradient Boosting Machines (GBM):**

GBM is an ensemble learning technique that builds a strong learner by combining multiple weak learners (typically decision trees) sequentially. It builds the model in a stage-wise fashion and tries to correct the errors of the previous models. GBM is known for its high predictive accuracy and ability to handle complex relationships in the data.

**Neural Networks:**

Neural networks, particularly deep learning architectures, have shown remarkable success in various domains, including image recognition and natural language processing. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are commonly used for image and sequential data, respectively. For this problem, a deep learning approach, especially with architectures tailored for tabular data, could be explored.

**k-Nearest Neighbors (k-NN):**

k-NN is a simple and intuitive classification algorithm that can work well with reduced-dimensional data. It makes predictions based on the majority class among the k-nearest neighbors in the feature space.

**Gaussian Naive Bayes:**

Naive Bayes classifier assumes that features are conditionally independent given the class label. Although this assumption may not hold after PCA, Gaussian Naive Bayes can still be applied and can perform well, especially if the features are approximately normally distributed.

# **Model selection**

Given the characteristics of our problem (classification based on spectral and photometric characteristics) and the dataset size (100,000 observations), models like **Random Forest, Gradient Boosting Machines, k-NN, and Neural Networks** could be promising choices.

Experimentation and model evaluation on a validation set would ultimately determine which model performs the best for this specific task and dataset.