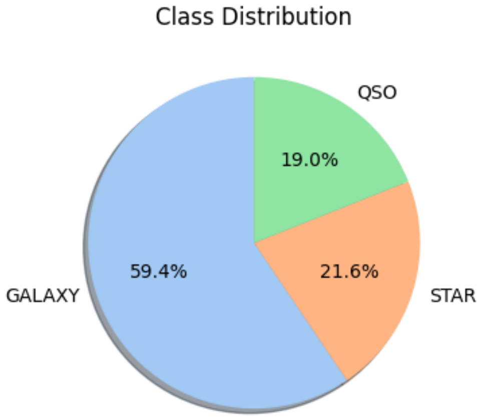
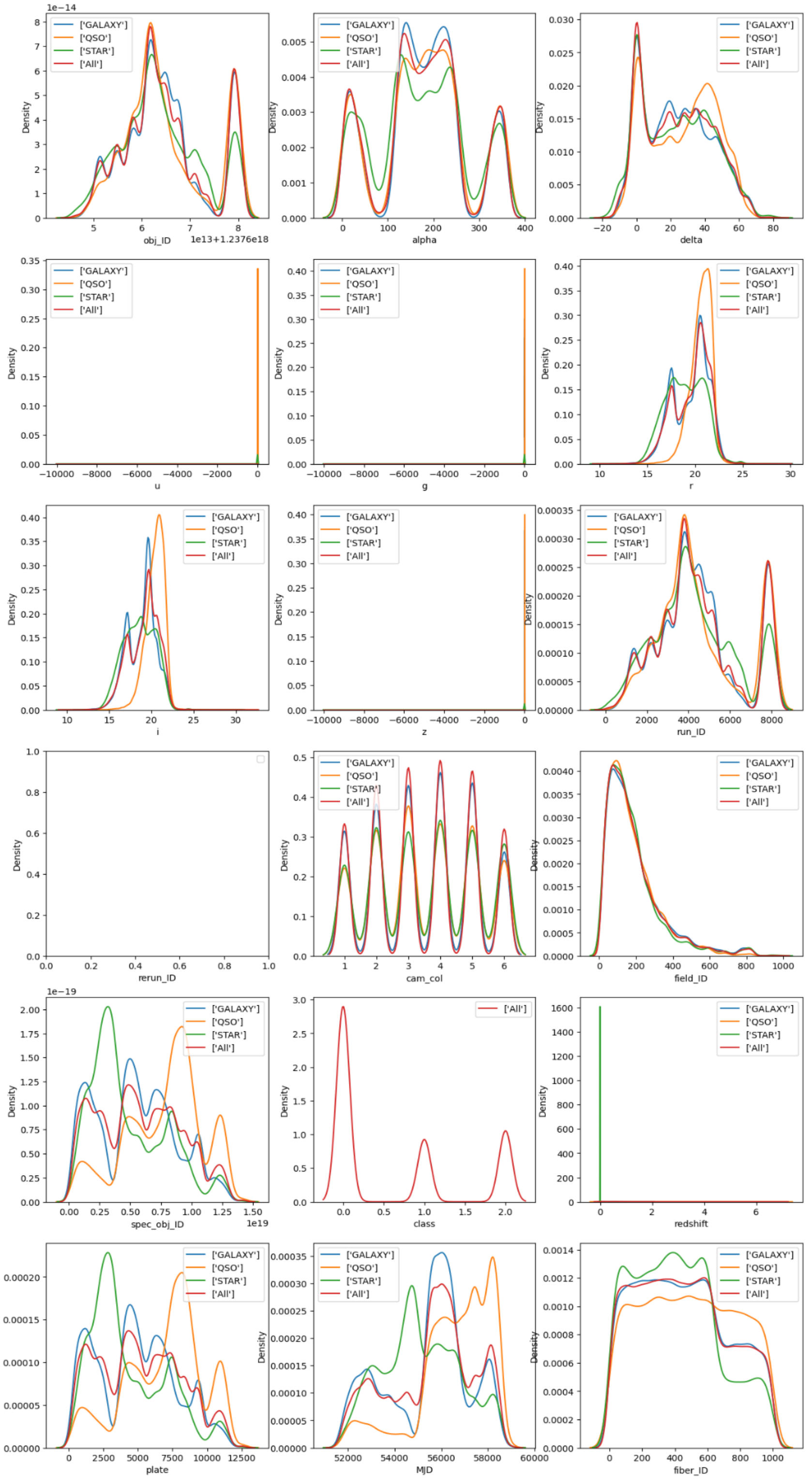
SATA6685 Project Progress Report

Reintroduction:
In astronomy, stellar classification is the classification of stars based on their spectral characteristics. The classification scheme of galaxies, quasars, and stars is one of the most fundamental in astronomy. Our project aims to classificate stars, galaxies, and quasars based on their spectral characteristics.

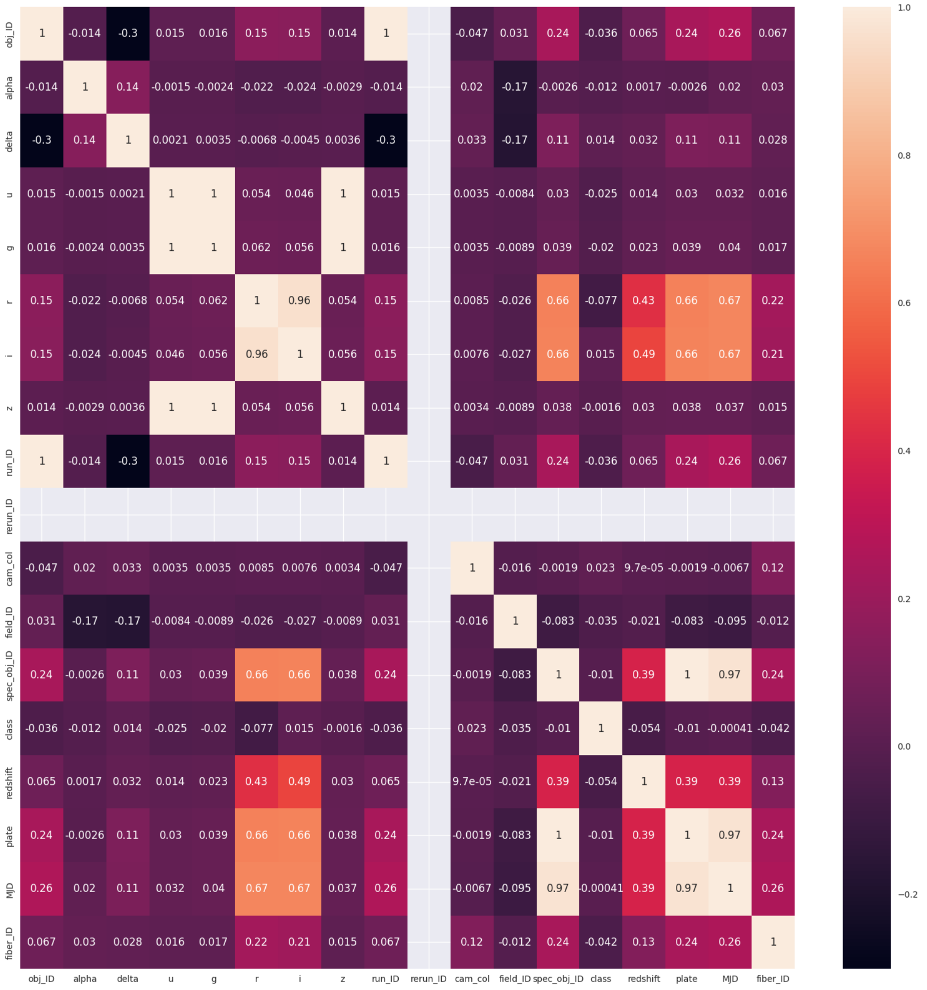
1.What progress has been made on your project?
Initially, we downloaded the dataset and conducted thorough data analysis and preprocessing. Two functions, namely Plot and plot_log, were developed to visualize the distributions of specific columns in the dataset. These functions facilitate the comparison of distributions across different classes and for the entire dataset. We also calculated correlation coefficients among variables. The dataset underwent some preprocessing, including handling imbalanced classes through SMOTE and feature scaling using StandardScaler. Class distribution is shown below:



We also have KDE graphs to help us determine useful features.



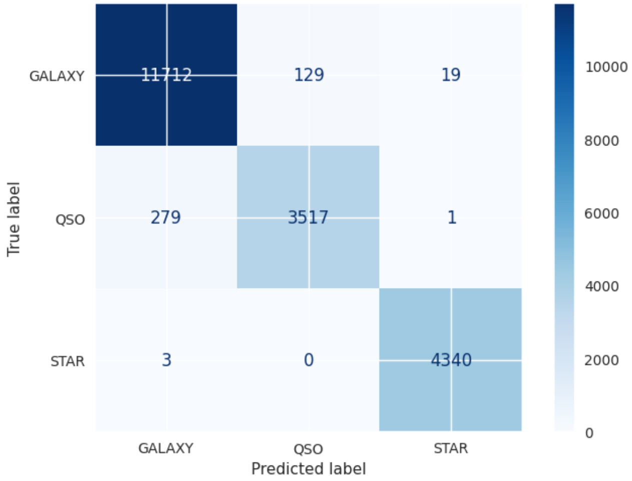
Correlation coefficients between variables



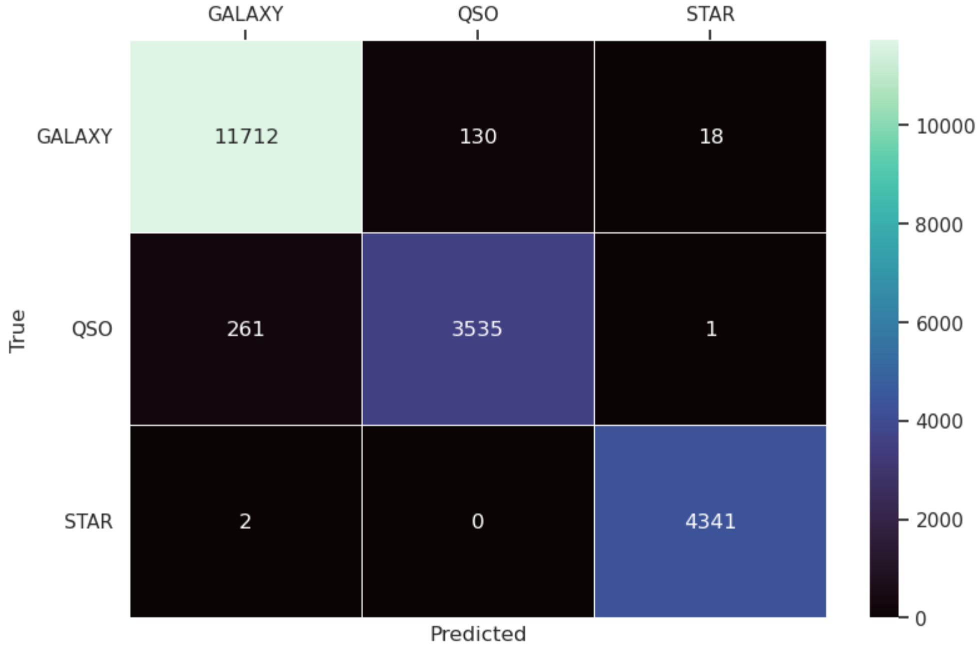
	obj_ID	alpha	delta	u	g	r	i	z	run_ID	rerun_ID	cam_col	field_ID	spec_obj_ID	class	redshift	plate	MJD	fiber_ID
obj_ID	1	-0.014	-0.3	0.015	0.016	0.15	0.15	0.014	1		-0.047	0.031	0.24	-0.036	0.065	0.24	0.26	0.067
alpha	-0.014	1	0.14	-0.0015	-0.0024	-0.022	-0.024	-0.0029	-0.014		0.02	-0.17	-0.0026	-0.012	0.0017	-0.0026	0.02	0.03
delta	-0.3	0.14	1	0.0021	0.0035	-0.0068	-0.0045	0.0036	-0.3		0.033	-0.17	0.11	0.014	0.032	0.11	0.11	0.028
u	0.015	-0.0015	0.0021	1	1	0.054	0.046	1	0.015		0.0035	-0.0084	0.03	-0.025	0.014	0.03	0.032	0.016
g	0.016	-0.0024	0.0035	1	1	0.062	0.056	1	0.016		0.0035	-0.0089	0.039	-0.02	0.023	0.039	0.04	0.017
r	0.15	-0.022	-0.0068	0.054	0.062	1	0.96	0.054	0.15		0.0085	-0.026	0.66	-0.077	0.43	0.66	0.67	0.22
i	0.15	-0.024	-0.0045	0.046	0.056	0.96	1	0.056	0.15		0.0076	-0.027	0.66	0.015	0.49	0.66	0.67	0.21
z	0.014	-0.0029	0.0036	1	1	0.054	0.056	1	0.014		0.0034	-0.0089	0.038	-0.0016	0.03	0.038	0.037	0.015
run_ID	1	-0.014	-0.3	0.015	0.016	0.15	0.15	0.014	1		-0.047	0.031	0.24	-0.036	0.065	0.24	0.26	0.067
rerun_ID																		
cam_col	-0.047	0.02	0.033	0.0035	0.0035	0.0085	0.0076	0.0034	-0.047		1	-0.016	-0.0019	0.023	9.7e-05	-0.0019	-0.0067	0.12
field_ID	0.031	-0.17	-0.17	-0.0084	-0.0089	-0.026	-0.027	-0.0089	0.031		-0.016	1	-0.083	-0.035	-0.021	-0.083	-0.095	-0.012
spec_obj_ID	0.24	-0.0026	0.11	0.03	0.039	0.66	0.66	0.038	0.24		-0.0019	-0.083	1	-0.01	0.39	1	0.97	0.24
class	-0.036	-0.012	0.014	-0.025	-0.02	-0.077	0.015	-0.0016	-0.036		0.023	-0.035	-0.01	1	-0.054	-0.01	-0.00041	-0.042
redshift	0.065	0.0017	0.032	0.014	0.023	0.43	0.49	0.03	0.065		9.7e-05	-0.021	0.39	-0.054	1	0.39	0.39	0.13
plate	0.24	-0.0026	0.11	0.03	0.039	0.66	0.66	0.038	0.24		-0.0019	-0.083	1	-0.01	0.39	1	0.97	0.24
MJD	0.26	0.02	0.11	0.032	0.04	0.67	0.67	0.037	0.26		-0.0067	-0.095	0.97	-0.00041	0.39	0.97	1	0.26
fiber_ID	0.067	0.03	0.028	0.016	0.017	0.22	0.21	0.015	0.067		0.12	-0.012	0.24	-0.042	0.13	0.24	0.26	1

Presently, we have successfully implemented Gradient Boosting, Logistic Regression, and Random Forest as baselines. Results for these approaches are shown below

Random Forest Accuracy: 97.79%
Confusion Matrix:



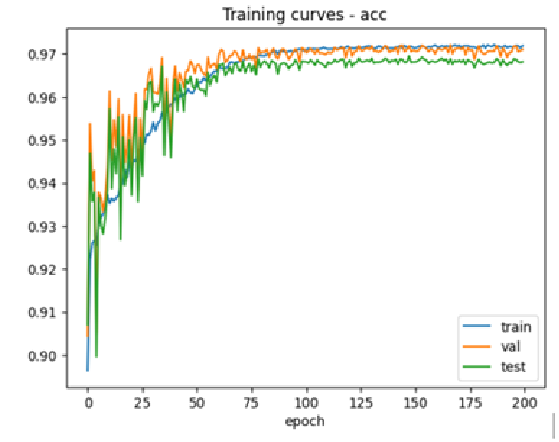
Gradient Boosting Accuracy: 97.94%
Confusion Matrix:

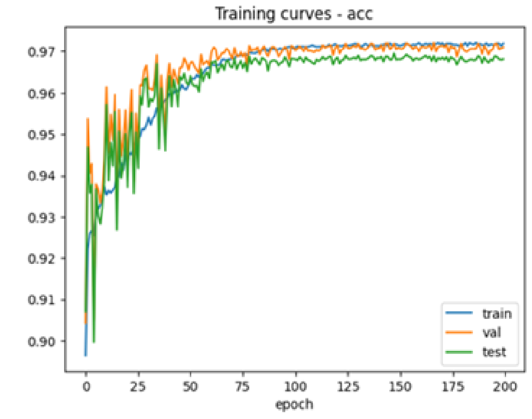


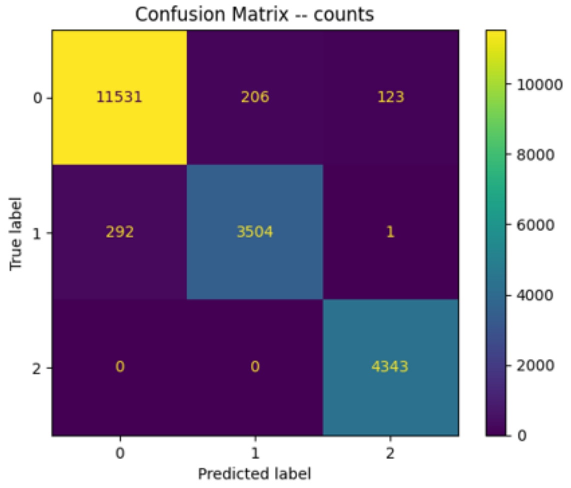
Logistic Regression Accuracy: 94.36%

Logistic Regression provides a nice baseline for the minimum that our neural network should be able to achieve (as we can implement logistic regression easily with a neural net).

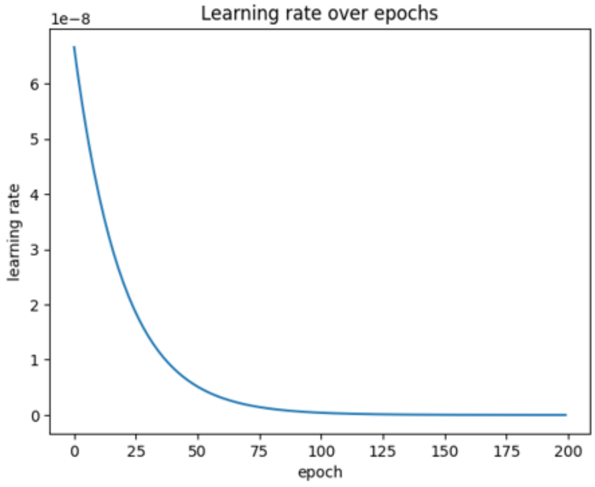
Current Multilayers Perceptrons Accuracy: 97.21%
Training curves and confusion matrix:






Confusion Matrix -- counts

True label \ Predicted label	0	1	2
0	11531	206	123
1	292	3504	1
2	0	0	4343



2.Have you run into any problems with your proposed project? If so, how do you plan on overcoming them?
One challenge encountered in our project is overfitting. To address this issue, we have implemented regularization within our network.

3.What are your plans for completing the project?
At this point, we want to try out multiple HPO strategies for our neural network. We will likely start with grid search and random search. Raytune and Bayesian optimization are our other main HPO considerations right now. We may further refine the baseline models, and see if we can outperform them with our updated neural network.

4.How has each team member contributed to the progress of your project?
Collaboratively, we completed the initial data processing phase. Zac took charge of implementing Gradient Boosting and Logistic Regression for stellar classification. Yingying successfully handled Random Forest and Multilayer Deep Neural Network components. Moving forward, our team will unite to finalize the project report and deliver a cohesive presentation in the upcoming weeks.