Data Mining – Project Checkpoint Questions:

1. Can you describe the specific goals and objectives of your project, and have they evolved or changed since the start of the project?

A: The goal stated initially in the debut document has not changed. The target of our project is to use machine learning to classify cosmic objects.

2. What tools or technologies are you using to perform data analysis and modeling in your project, and why did you choose them?

A: This project is completely being implemented using python for data analysis. Modeling will be done mainly using the SciKit Learn library in python. In discussion to currently include deep learning models using TensorFlow for classification.

3. How are you handling missing or incomplete data in your dataset, and what techniques have you applied for data imputation or cleansing?

A: The data currently in our possession is well-formatted public domain data. We have no missing values. Data was extracted from multiple sources and merged for performing the analysis. The cleansing mechanisms we mainly had to employ were to decide a standard format for the final merged data and performing operations on individual datasets for easy merging. 4 datasets were sourced with a total of approximately 300k rows and 17 features.

4. Can you provide insights into the feature engineering process you've undertaken? What features have you generated, and how have they impacted your modeling?

A: This data being used for this project has 17 features. Since this project is related to astronomy, a high level domain knowledge is required to perform feature engineering. Currently, we have checked the distributions of different features for each target class to filter out the filters with the help of box plots and violin plots. The above process combined with correlation analysis to determine the most significant features.

5. Are there any unique challenges or ethical considerations related to your dataset or project that you've encountered, and how have you addressed them?

A: The data being used is public domain data that is released for research and analysis purposes. Ethically there is no violation by using this data in our project. The analysis we have done and we will be doing in the next phases also have no ethical consequences.

6. How do you plan to evaluate the performance of your machine learning model(s)? What metrics are you using, and why?

A: As we are predicting multiple classes in the target variable, we plan on using the standard metrics namely; accuracy, precision, recall and f-1 scores for the models. Based on the results we obtain we plan to choose the top 2 models and perform hyperparameter tuning to check if the model can be improved.

7. Have you encountered any unexpected or interesting patterns or insights during the data exploration phase that have influenced your project's direction?

A: The current biggest concern we have is class imbalance. We observed that after merging data from multiple sources we have a very high class imbalance. Currently experimenting methods to remove this imbalance. The challenge is to find an approach that will remove the imbalance and also have no negative consequences on the modeling we are planning to perform.

8. What is your plan for model selection and tuning? How do you intend to optimize the performance of your models?

A: We plan on using multiple models like LogisticRegression, SVC, RandomForestClassifier, KNeighborsClassifier, GaussianNB, DecisionTreeClassifier, AdaBoostClassifier, GradientBoostingClassifier, ExtraTreesClassifier, BaggingClassifier, MLPClassifier, XGBClassifier. Also, we plan on implementing Deep Learning models of Artificial Neural Networks and Convolutional Neural Networks.

9. What are your timelines for completing different phases of the project, including data collection, analysis, modeling, and final presentation or report submission?

We have already completed data collection and preprocessing. 4 Weeks left till final submission. Considering this Week as week-1.

Week-1: Further Analysis of Data. Feature Engineering

Week-2: Testing methods for Class Imbalance

Week-3: Modeling and Results Inference

Week-4: Final modifications/ upgrades. Making Reports and Presentations.

10. How are you collaborating with your team members on this project, and what communication tools or strategies have you found most effective?

A: We are using Google Colab and GitHub for maintaining a record of our contributions. Communicating in-person whenever possible and making use of Zoom for any discussions.

11. Are there any additional resources or support you require to successfully complete your project, such as access to specific datasets, software, or hardware?

A: As of now, we have no additional requirements.

12. Have you faced any unexpected challenges or setbacks that have affected your project's progress, and how are you planning to mitigate or overcome them?

A: No Major Setbacks as of now. The amount of data we have now is challenging as too many samples will have an effect on the performance of the models if proper analysis is not taken into account.

13. What have you learned or gained from the project experience so far, and how will it contribute to your future data mining and machine learning endeavors?

A: Personally, locating different sources of data and deciding if they are applicable for the goals of the project has been a major learning.Collaborating with different people and taking into account different perspectives to approach this project has helped in widening my insights. These will be a lot of help in my career as a Data Scientist.