# INDIVIDUAL WORK

The team meeting was carried to start preparing the project's tasks. At the conclusion of the meeting, the project's targeted scope and development phases were decided.

The team used to join remotely for feedback meetings two or three times each week. There was continuous and noticeable communication within the group. Several Google Meet screen-sharing conversations were conducted during the implementation stage.

A section of the documentation was divided, and each group member started to construct the content in accordance with responsibility. The finalized documentation are created by integrating together all group members' individual work.

Ultimately, we may say that these project targeted goals have been successfully achieved through effective collaboration.

**RISHIKESH ANIL WAGHELA** – 33% [Feature Selection, Model Development and Evaluation, Data Preparation]

**Reflection of Learning:-**

1. Data Preparation:- An researcher spends 80% of their time preparing data. To streamline this procedure, specialist data preparation tools are crucial. The entire process of handling categorical attributes, numerical attributes, and missing values in the dataset is considered as data preparation. Although the vast majority of models only work with numeric values, many important real-world characteristics are categorical rather than numeric. Whenever they are employed as categorical characteristics, they acquire levels or numbers. However, we require all characteristics to be numerical for our logistic model. The categorical feature, which has 4690 labels, is the most significant obstacle. Given that there are numerous transformation techniques, one that is applied to this attribute is the frequency count method. This is simple to put into practice, not a feature area expansion and able to perform well with algorithms which are tree-based. There are some demerit of this method which one can face is As a result of replacing two labels with the same number of observations when there are two unique labels with the same number of observations, we may lose important information.

2. Feature Selection:- Developing an effective set of features for the training set is a vital factor in the achievement of a machine learning projects and is defined as feature engineering. There are many strategies for choosing the best feature, including filtering, wrapping, and other methods. One of which we have used correlation coefficient(Pearson coefficient). We can anticipate one variable from another if the two are correlated. Since the second feature does not provide any new information, the model only requires one correlated feature. Here, we'll make use of the Pearson Correlation. Created a function that eliminates the feature which is highly correlated with other features in order to increase efficiency and eliminate the need for manual effort.
3. Model Development:- Since, we have deployed three models , in which we got different accuracy. To improve the accuracy and biased our dataset we implemented

methods like Oversampling and UnderSampling. In UnderSampling method it Keeps all the records in the minority class and reducing the size of the majority class. By tunning the hyperparameters of logistic regression we got the accuracy of 100%. As can be seen, all models are working fairly well using only baseline modelling and with further hyperparameter tuning. For a better comprehension of the dataset, perhaps we could instead find a method to fill the diameter, albedo, and diameter sigma NaN values.