**Report for Nasa\_NEO Project:-**

CWID: - 20015840 , 20016333, 20012322

Name of Student: - Atharv Subhekar, Nikita Vanga, Pranali Vyas

This dataset provides information on a set of Near-Earth Objects (NEOs) which have been detected and monitored by NASA. Each of these objects has been watched by telescopes and detecting equipment and various data about its position, velocity, and makeup have been recorded.

A select few features were used to predict that the object is hazardous or not. Using label encoder we transformed the target values into binary values i.e. True = 1 and False = 0.

To keep the test data isolated we split the training data in to two parts where 80% of the data is for training and 20% of the data is for the testing data.

The data was plotted to visualize the features. From the visualized data, it is apparent that the range of the features is not standardized. Due to this, one feature can prove to be heavier or more decisive than another feature. This could affect the overall performance of the model. Hence, to avoid this we will standardize the features so that they will be in the same range.

Since, the problem is a binary classification problem, Logistic Regression and Support Vector Classification is used to classify.

The accuracy, precision, recall and F1 score of both models are listed below: -

**1)Logistic Regression:** The **accuracy** for this model is: 90.17% and **precision** is 51.87%. The **F1 score** and **Recall** was 13.391% and 7.68% respectively.

**2)Support Vector Classification:** The **accuracy** for this model is: 90.15% and **precision** is 81.81%. The **F1 score** and **Recall** was 0.99% and 0.50% respectively.

To keep the most important features and to increase variance between datapoints by using PCA. After training the models using logistic Logistic Regression the **accuracy** of the model has decreased to 89.57% and **precision** decreased to 25.12% and the **F1 score** and **Recall** also decreased and became 5.01% and 2.78% respectively. Similarly, the **accuracy** of the model decreased to 89.77% and **precision** decreased to 25% and the **F1 score** and **Recall** increased and became 3.23% and 1.72% respectively for Support Vector Classification.

So, from the above paragraph it can be concluded that using PCA will decrease the **accuracy** and **precision** for both the models. However, in contrast the precision, recall and F1-score is better.