Machine Learning

Assignment 1

surname Name

May 9, 2024

***Stage 1: Data set and Pre-Processing***

The dataset is focused on asteroid diameter prediction and contains various features related to asteroids. It is officially maintained by the Jet Propulsion Laboratory of the California Institute of Technology, which is an organization under NASA.

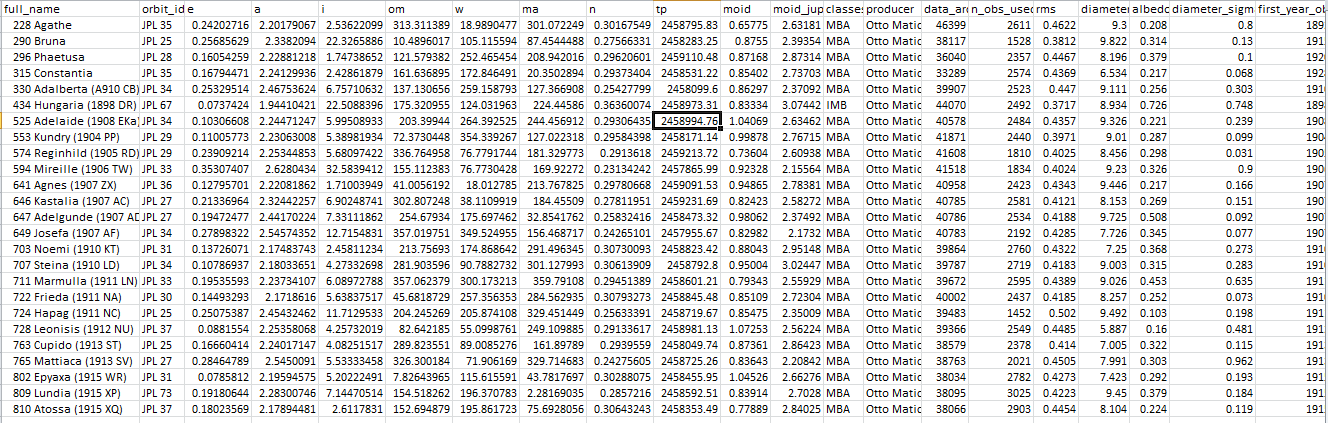
***Here are the column names and a brief description of each feature:***

1. full\_name: The full name of the asteroid.
2. orbit\_id: Unique identifier for the orbit of the asteroid.
3. e (eccentricity): Eccentricity of the asteroid's orbit.
4. a (semi-major axis): Semi-major axis of the asteroid's orbit in astronomical units (au).
5. I (Inclination): Inclination of the asteroid's orbit with respect to the x-y ecliptic plane in degrees.
6. Om (Longitude of the ascending node): Longitude of the ascending node of the asteroid's orbit.
7. W (Argument of perihelion): Argument of perihelion of the asteroid's orbit.
8. ma: Mean anomaly of the asteroid's orbit.
9. n: Mean motion of the asteroid's orbit.
10. tp: Time of perihelion passage of the asteroid's orbit.
11. moid (Earth Minimum orbit Intersection Distance): Minimum orbit intersection distance between Earth and the asteroid's orbit in astronomical units (au).
12. moid\_jup: Minimum orbit intersection distance between Jupiter and the asteroid's orbit in astronomical units (au).
13. classes: Class of the asteroid.
14. producer: The organization or entity that produced the dataset.
15. data\_arc: Data arc span in days.
16. n\_obs\_used: Number of observations used.
17. rms: Root mean square deviation of residuals.
18. diameter: Diameter of the asteroid in kilometers.
19. albedo: Geometric albedo of the asteroid.
20. diameter\_sigma: Uncertainty or error in the asteroid's diameter measurement.
21. first\_year\_obs: Year of the first observation.
22. first\_month\_obs: Month of the first observation.
23. last\_obs\_year: Year of the last observation.
24. last\_obs\_month: Month of the last observation***.***

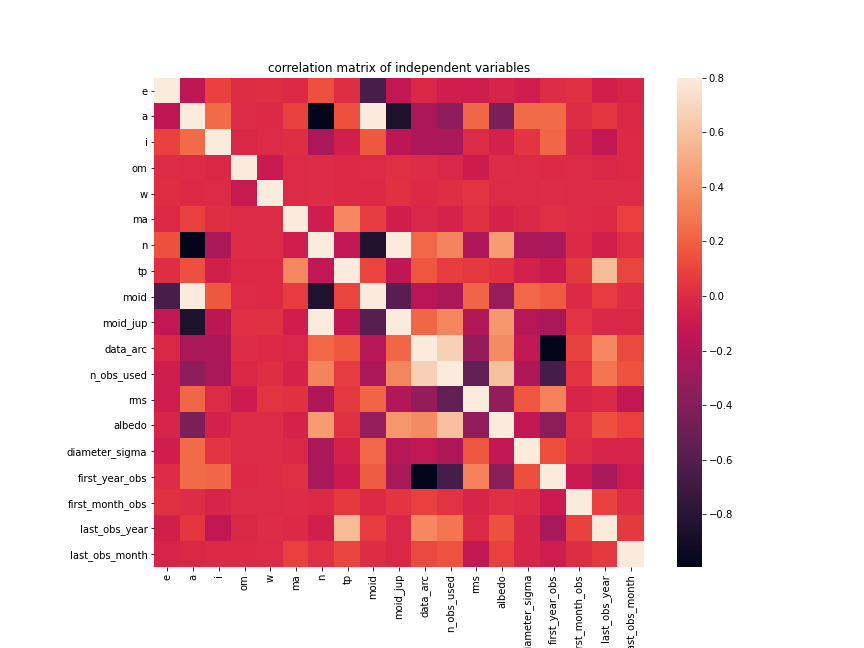
These features provide information about the orbital characteristics, physical properties, and observational details of the asteroids in the dataset.

A.I. algorithms find its way to the solution in case of astronomy?. Yes, it could. But the most challenging thing which I found, in this case, was the availability of sufficient data.  
Although the struggle to find appropriate data to work with was not much longer. I finally came across this dataset which is officially maintained by **Jet Propulsion Laboratory of California Institute of Technology** which is an organization under **NASA**. With this, I also found my way to solve the most interesting problem, which is to predict asteroid diameter with A.I. algorithm.  
  
I researched this topic and found that the prediction of asteroid diameter is not an easy job. There had been various methods to solve the estimation of asteroid diameter, every method trying to surpass its previous one.  
  
I tried solving this problem with various ML algorithms like Linear Regression, Decision Tree regression, Random Forest Regressor and finally concluded that Random Forest Regressor could solve this problem with the least error possible and higher accuracy.

***First 25 Rows:***

******

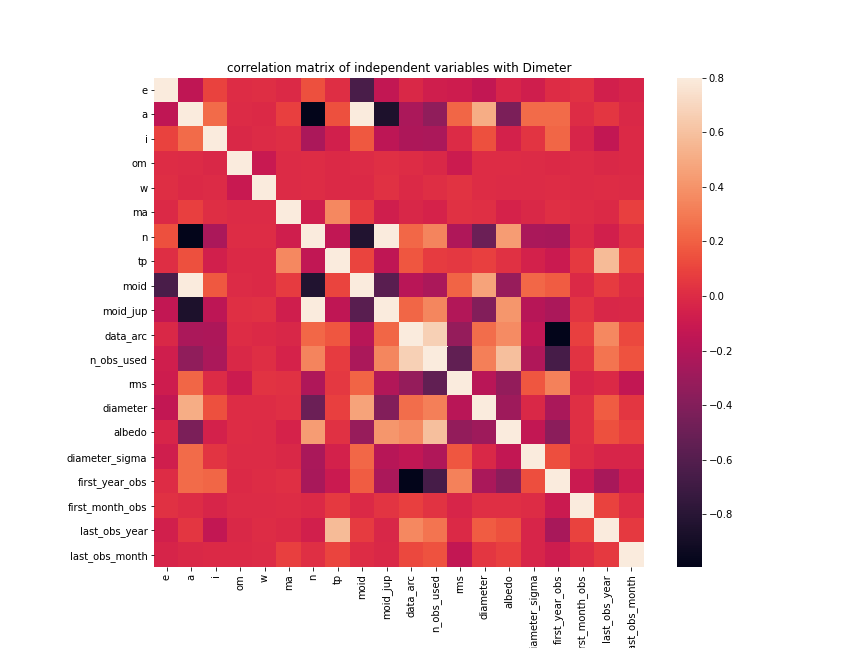
***Correlation Metrics With Heatmap:***

******

In the above correlation heatmap plot, we can observe that some variables exhibit a high correlation with each other, resulting in data redundancy. As a result, we need to drop the features that show a low correlation with our target variable.

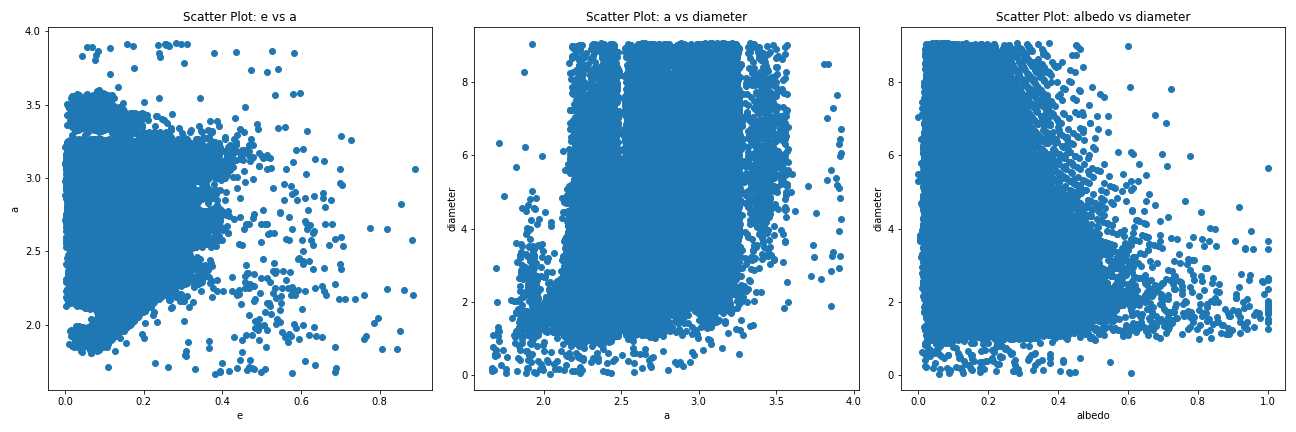
We decide to drop the "first\_year\_obs"(highly correlated with “data\_arc” and have low correlation with “Diameter”) and "moid\_jup"(highly correlated with “a” and have low correlation with diameter) finally we decide to drop these features.

***Correlation Metrics with target variable:***



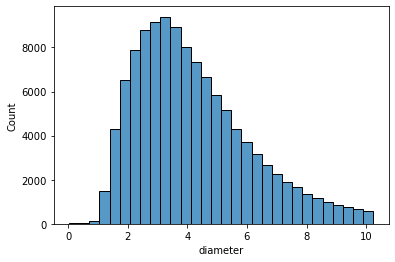
From the above plot, we can observe that the features "O" and "W" have a very weak correlation with the target variable "Diameter". As there is no significant relationship between these features and the target variable, it is advisable to drop them from the dataset.

***Three plot of two variable at a time:***

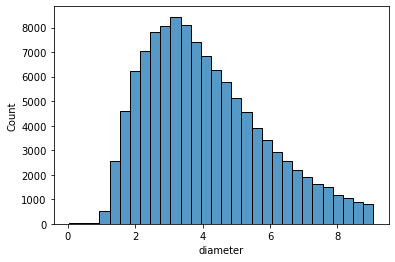
******

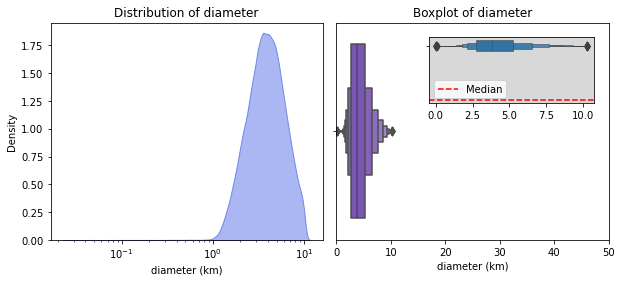
* The first plot (a vs. e) shows a moderate correlation between the semi-major axis (a) and eccentricity (e) of asteroids. The scatter points form a pattern whereas the semi-major axis increases, the eccentricity tends to increase as well..
* The second plot (a vs diameter) shows the relationship between the semi-major axis (a) and the diameter of asteroids. It suggests that as the semi-major axis increases, the diameter also tends to increase. However, most of the diameter values fall within the range of 25 and 30 for the semi-major axis.
* The third plot (albedo vs diameter) displays the relationship between the albedo (reflectivity) and diameter of asteroids. It indicates a negative correlation, suggesting that as the albedo increases, the diameter tends to decrease. This implies that asteroids with higher albedo tend to have smallerdiameters.

***Distribution of Diameter before removing outlier:***

******

***Distribution of Diameter after removing outlier:***

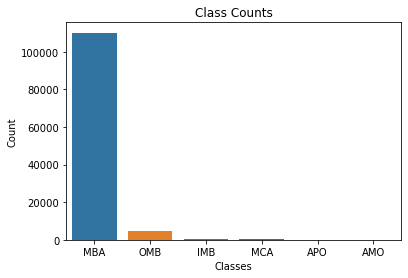
******

******

Alright, we see that the dataset contains values from 0.0025 to 10.0+, with the most of the values being in a range from 1 to 10. Even though some asteroids seem to be huge and have really large diameters of 10 km or more, the median diameter for the asteroids from this dataset is around 3.8 km.

Note, it's wrong to think that the greater the diameter of the asteroid, the more dangerous it is for Earth. The asteroid is considered potentially hazardous only if it is a Near-Earth object, and generally largest asteroids are less common among NEAs, because they have a greater gravitational attraction to the Sun, which causes them to be more stable in their orbits farther from Earth. Additionally, larger asteroids are more rare overall, as they represent a smaller fraction of the total population of asteroids in our solar system.

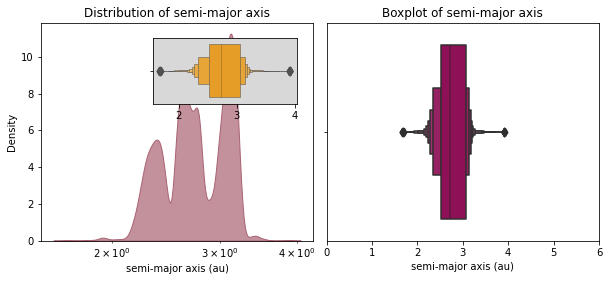
***Class Distribution:***

******

We see that MBA is certainly leading with around 92% of all asteroids being of that class. Our dataset contains some asteroids from the Outer Main-belt, and much less of other types. Not to mention, that we have only 80 asteroids orbiting outside AMO, all other asteroids usually cross the plane of orbit of certain planets in the solar system.

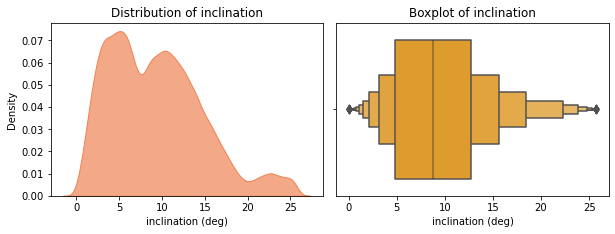
This actually might make it difficult for our model to generalize when it comes down to predicting diameters in the long run. We would want to make the model predict the diameters for far asteroids correctly, but as of now we won't focus on it, since most of the asteroids we have are MBAs.

***Distribution of semi major axis (au):***

******

So we see that the most popular value for semi-major axis is around 2.5 and 3, and, overall, it ranges between 1.5 au and 4 au, depending on the class of the asteroid.

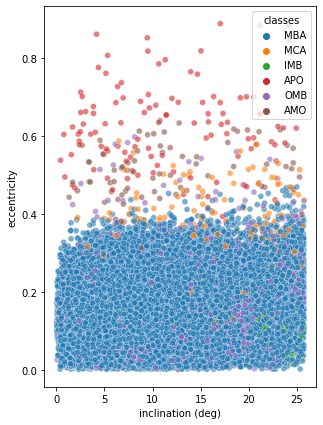
***Distribution of inclination:***

******

So we see that most of the eccentricities are from 0 deg to 15 deg. This may be a hint telling us that most of the asteroids in our dataset are MBAs (low eccentricity due to stable orbit). Not to mention that for these values of inclination we have the range for semi-major axis from 1.5 au to 4 au. Such semi-major axes commonly have MBAs, OMBs, IMBs, APOs, MCAs, APUs.

Therefore, the relationship between an asteroid's inclination and eccentricity can provide important information about the class asteroid belongs to, and thus, its overall characteristics.

***Distribution of eccentricity with respect to classes:***



From the plots this can be concluded:

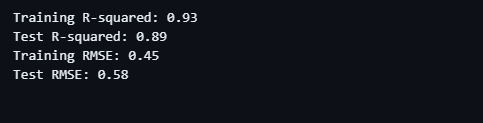
* MBAs typically have low eccentricities and inclinations, which means their orbits are relatively stable.
* MCAs usually have low-inclination but high-eccentricity orbits, because they may have been perturbed by the gravity of Mars.
* APO have orbits that bring them close to or cross the orbit of Earth. As a result, they tend to have higher eccentricities and inclinations than MBAs.Certain types of NEAs such as ATEs tend to have higher inclinations than other.

***Stage 3: Model Implementation:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **Performance or Evaluation Metrics**  **Cross Validation** **Root-Mean-Square Error (RMSE)** | **Evaluation Metrics of Test RMSE** | **Evaluation Metrics of R\_Squared** | **Evaluation Comment (e.g., good, bad or best)** |
| Linear Regression | **0.8537865929855865** | **0.8548016521438068** | **0.77** | The linear regression model has the highest test RMSE (0.8548016521438068) and the lowest R-squared value (0.77) among the three models. This suggests that the linear regression model may not be the best fit for the data, as it has a weaker predictive performance and explains less of the target variable's variance compared to the other models. |
| Decision Tree | **0.8163839096996377** | **0.8111603731138083** | **0.79** | The decision tree model has a slightly higher test RMSE (0.8111603731138083) and a lower R-squared value (0.79) compared to the random forest model. However, it still performs reasonably well, capturing a substantial portion of the target variable's variance. |
| Random Forest | **0.5542037591223947** | **0.5557782625857277** | **0.90** | The random forest model performs the best among the three models evaluated, as it has the lowest test RMSE (0.5557782625857277) and the highest R-squared value (0.90). This indicates that the random forest model has a good predictive performance and explains a significant amount of the variance in the target variable. |

After running HyperParameter Optimization for the Decision Tree Regressor, Random Forest Regressor, the best model actually turned out to be the Random Forest Regressor. The R² value is highest for both the training and validation sets as well as the lowest error values across all models. It should be noted that the Random Forest model used optimal parameters from a 334 minute RandomizedSearchCV. (nearly 5.5 hours )

When comparing the optimized Random Forest Regressor to the test set, I got the following results:

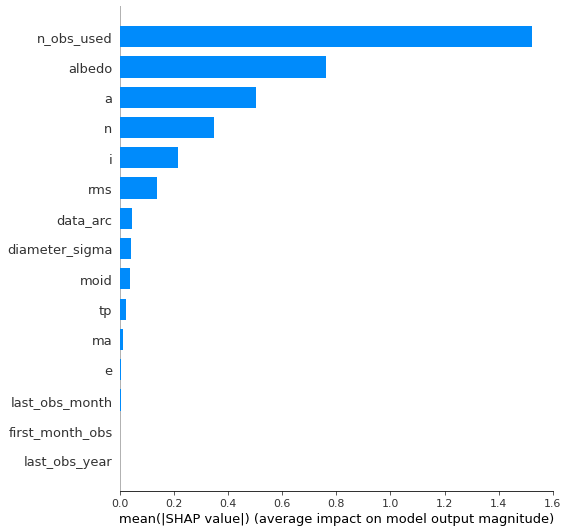


The results from the test set show very favorable results compared to the validation sets. The Testing MAE and MSE were both lower than the errors of the validation set and the Testing R² actually improved as well. This helps prove the Random Forest Regression model was successful in accurately predicting an asteroid’s diameter.

# *****Visualizations & Insight*****

**Shapley Plots**

I created Shapley plots to help describe the impact important features had in either increasing or decreasing the predicted diameter in km. Below is an example of the plot that was generated.



These plots are extremely helpful and useful in understanding how the Random Forest Regressor model predicts a value. The first three bars show the features that increased the predicted diameter whereas the last four bar show the variables that brought down the prediction.

# *Conclusion*

The importance of predicting asteroids can be the difference between life or death. Granted, there is a hint of exaggeration to that statement but given a large enough object, accelerating to surface of Earth at the force of gravity, (9.8 [m/s²]) there can be some serious damage. The field of predicting asteroids based off of observational characteristics is still in its infantile stages and only growing. With the ever-advancing methods for collecting data combined with powerful innovation in ML models, predicting asteroid diameters can only improve. The impact machine learning can make is truly boundless and this project is confirmation of the utility machine learning models have in predicting asteroids’ diameters. The overall project can be viewed as a success as it was able to predict accurate and powerful results using a machine learning model.