**INTRODUCTION TO STATISTICAL LEARNING-PROJECT REPORT**

**CLASSIFICATION OF HAZARDOUS ASTEROIDS**

**Siva Sameer Krishna. Yarlagadda**

[**Sy9bd@mail.umkc.edu**](mailto:Sy9bd@mail.umkc.edu)

**16271011**

**Project presentation ppt link:** [**google slides ISL project presentation**](https://docs.google.com/presentation/d/1g_951Jc_cou-TvvTCsa6h0drR2HUwaWThD3NdbaKC48/edit?usp=sharing)

**Motivation:**

Near-Earth Objects (NEOs) are comets and asteroids that have been pushed by the gravitational attraction of nearby planets into Earth’s orbit that allow them to enter the Earth’s neighborhood.

**Objective:**

To predict whether an asteroid is hazardous or not.

To find out the features that are mainly responsible for the hazardous nature of asteroids.

**Dataset:**

The data is about Asteroids extracted from NeoWs(Near Earth Object Web Service).

NeoWs (Near Earth Object Web Service) is a RESTful web service for near earth Asteroid information.

Contains 4688 records and 40 columns nearly.

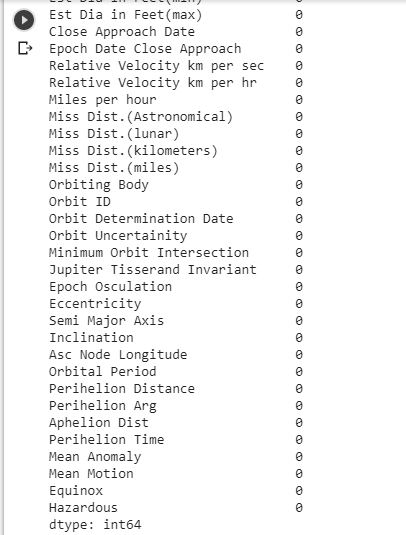
<https://www.kaggle.com/shrutimehta/nasa-asteroids-classification>

**Implementation:**

This is a **Classification problem** where we have to predict the label of Hazardous class which is known label. **Supervised learning** methods for classification such as Logistic Regression, Decision trees, Bayesian Classifier, K nearest neighbor , Support Vector Machine, etc. can be used to implement this.

**Data PreProcessing:**

There are no null values in dataset.



Hazardous attribute has categorical value, so we need to convert them into numeric type before feeding them into the model.

nasa['Hazardous'] = pd.to\_numeric(nasa['Hazardous']).astype(int)

**Correlation Factor:**

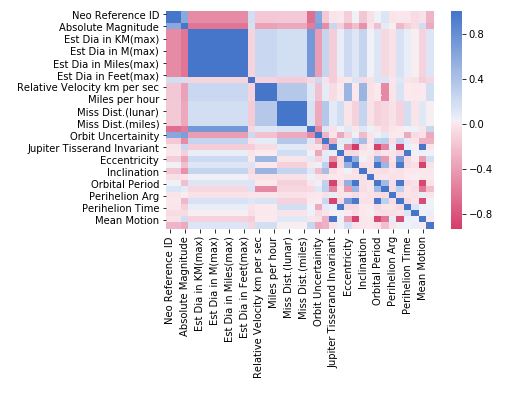
**We need to find which attributes are mainly contributing to tye model so we can get better accuracy. So correlation between columns is determined to find out those factors.**

correlations = df.corr()

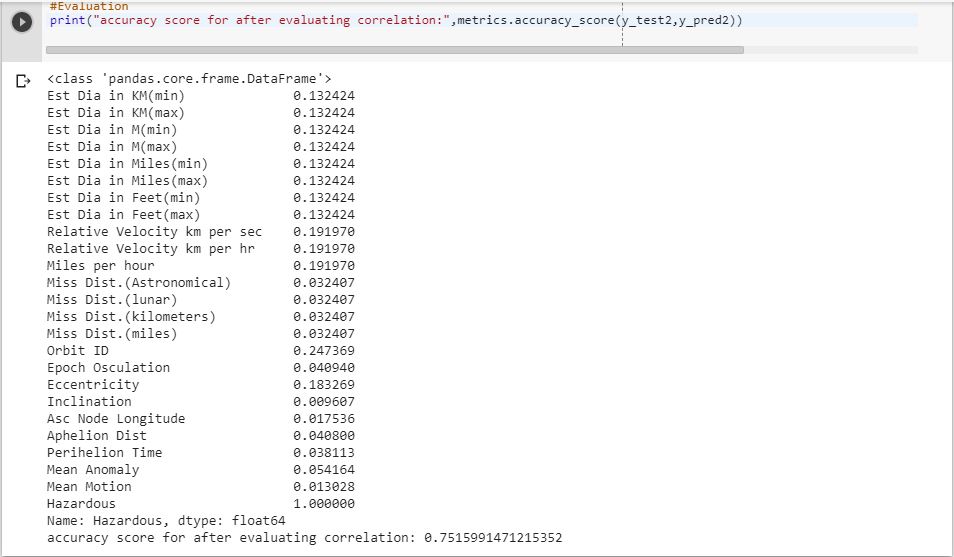
sns.heatmap(correlations, cmap=cmap)

plt.show()

**Correlation heat map:**



We considered only the columns which have correlation above 0, i.e., positively correlated.

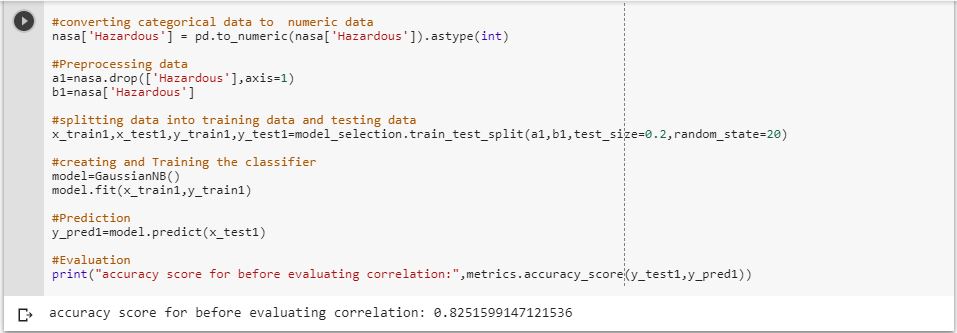


**Naïve Bayes classifier:**

Hazardous is target attribute and remaining attributes are passed as features to the model.

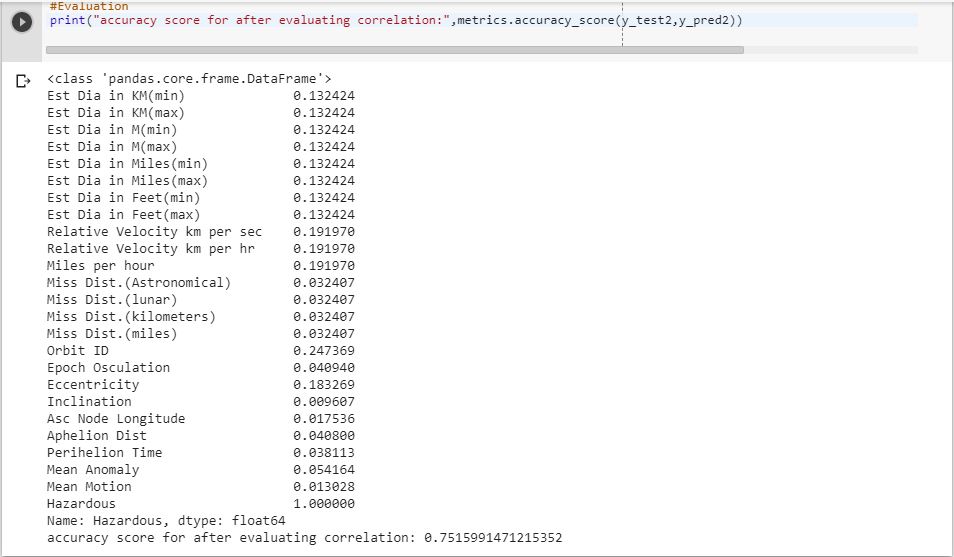
Test size is taken as 20% and training size 80%.

Before removing negative correlated columns:



After removing negative correlated attributes and evaluating the model.



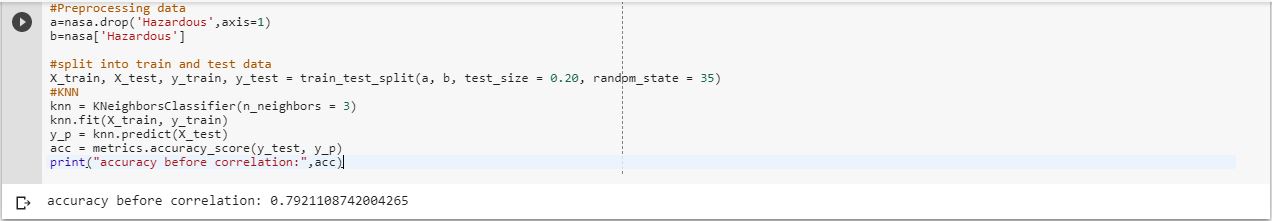


There should be increase in accuracy after removing the negative correlating columns, but to the contrary accuracy decreases.

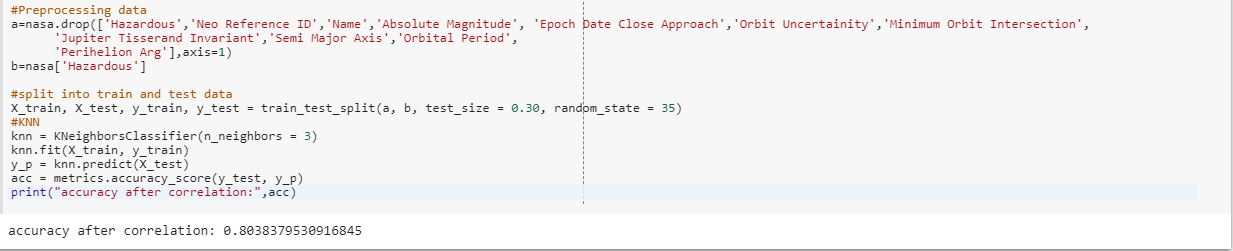
**K Nearest Neighbor classifier:**

KNN model is created and evaluated for accuracy.

Before removing negative correlation columns:



After removing negative correlation columns:

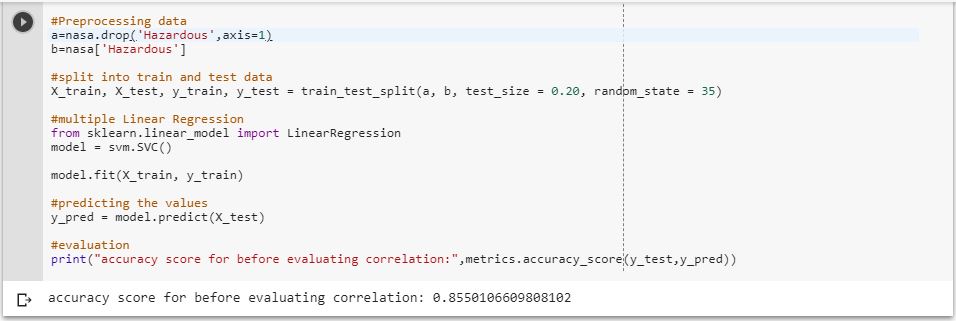


We can see the minute increase in accuracy.

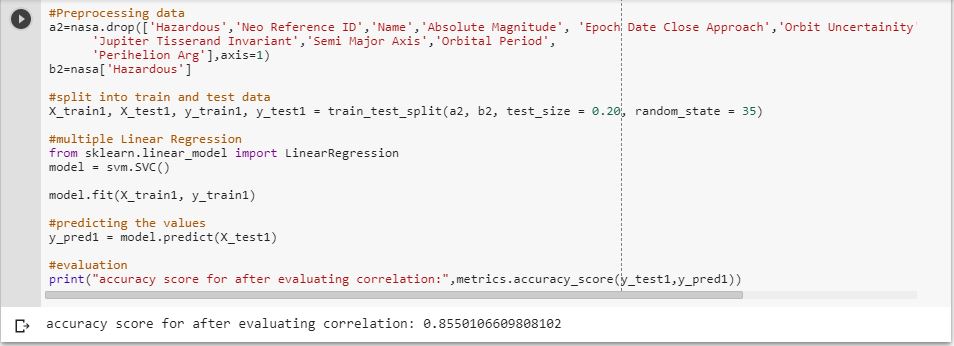
**SVM:**

Support vector machine linear classifier is built and evaluated.

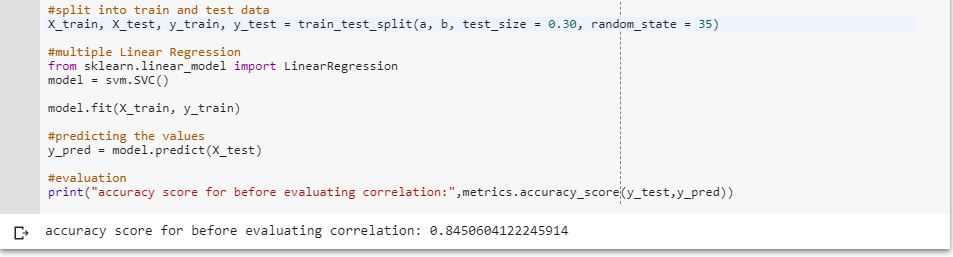
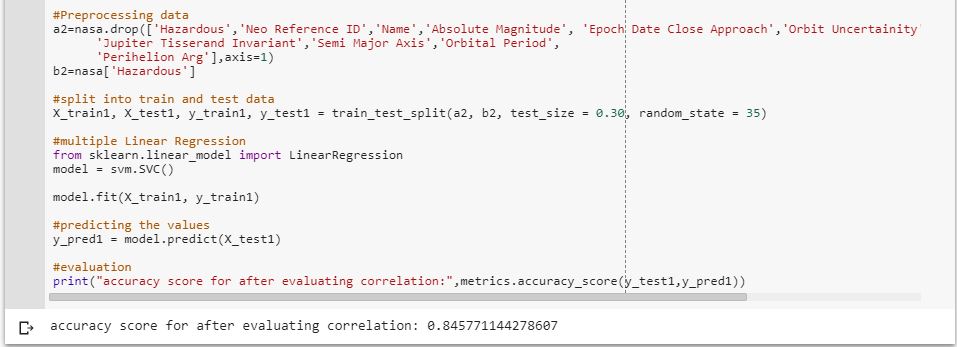
Before removing negative correlated columns:



After removing negative correlated columns:



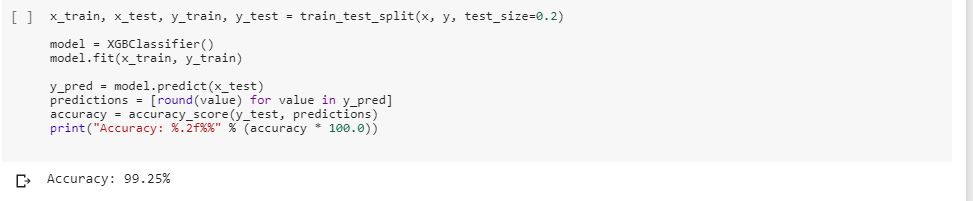
Accuracy is not changed before and after correlation data processing. So we changed the test size to 30% instead of 20%.

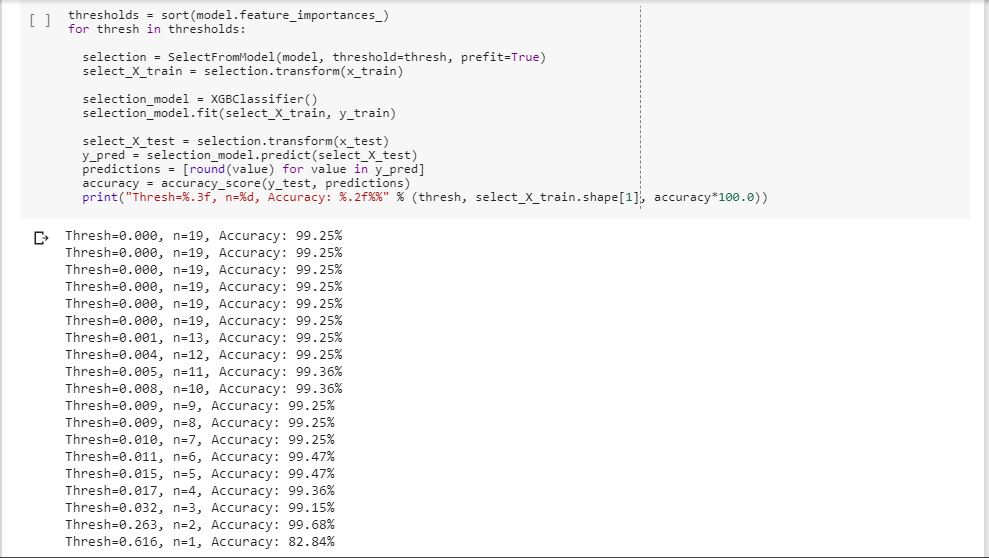
Minute change in accuracy is observed after evaluating correlation.

**XGB Classifier:**

**We got better accuracy for XGB classifier.**

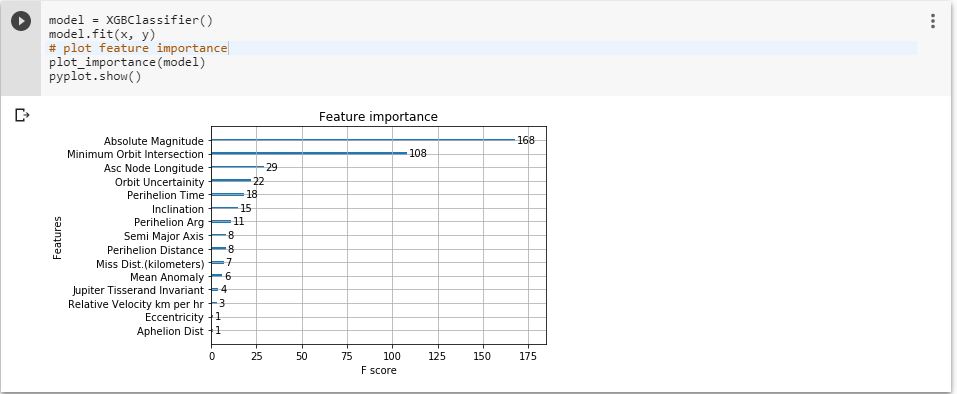


Since XGB evaluates the model in several steps and and each step learns from its previous steps, accuracy at each step is determined.



We can see that the final accuracy is almost achieved with initial 2 steps.

To find out the factors responsible for hazardous nature, default method in XGB classifier is used.

****

We can see that **Absolute magnitude** and **Minimum Orbit intersection** are important.

These 2 factors are only contributing to the accuracy of the model which we observed in the above output(accuracy at each step).

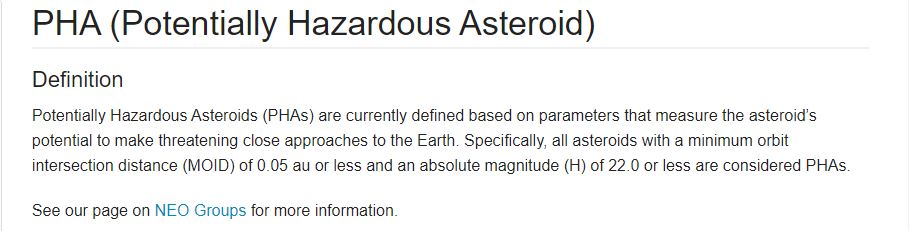
So we took only these 2 factors into consideration while building the model since only these are contributing to hazardous nature.



**Conclusion:**

XGB classifier gives better accuracy than all other classifiers.

**Absolute Magnitude and Minimum Orbit Intersection** are the major factors affecting the hazardous nature of asteroid.



We can see the definition given by NASA for potential asteroids and it is mentioned there as Absolute magnitude and Minimum orbit intersection with specific magnitudes are considered hazardous which we found from the model XGB classifier.

**References:**

1. <https://www.kaggle.com/shrutimehta/nasa-asteroids-classification>
2. <https://api.nasa.gov/api.html#apod>
3. <https://cneos.jpl.nasa.gov/glossary/PHA.html>