## A Capstone Project Report on

NASA NEAREST OBJECT PREDICTION USING MACHINE

LEARNING

Submitted by

DHANISH S

## ABSTRACT

It is not uncommon for asteroids to hit Earth. Hundreds of meteorites reach the surface of our planet every year, most too small to be of any concern. But occasionally, large rocks can hit and cause damage. In 2013, the Chelyabinsk meteor exploded over Russia, injuring hundreds. Such effects can be shock waves, heat radiation, the formation of craters with associated earthquakes, and tsunamis if water bodies are hit. There is an infinite number of objects in the outer space. Some of them are closer than we think. Even though we might think that a distance of 70,000 Km cannot potentially harm us, but at an astronomical scale, this is a very small distance and can disrupt many natural phenomena. These asteroids can thus prove to be harmful. Hence, it is wise to know what is surrounding us and what can harm us amongst those. Thus, this dataset compiles the list of NASA certified asteroids that are classified as the nearest earth object.

Dataset link: <https://drive.google.com/file/d/1kDIVnD3DEtXlpnx4ipqztnbxdLaiTDq6/view?usp=sharing>

Source code link:

<https://github.com/Dhanishsekar/NASA-NEAREST-OBJECT-CAPSTONE-PROJECT>

## ACKNOWLEDGEMENTS

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my capstone project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I have fortunate to have Mr. PRASAD as my mentor. He has readily shared his immense knowledge in data science and guides me in a manner that the outcome resulted in enhancing my data skills.

I certify that the work done by me for conceptualizing and completing this project is original and authentic.

Date: July 10, 2022 Name: DHANISH S

## CERTIFICATE OF COMPLETION

I hereby certify that the project titled “Nasa Nearest Object Prediction Using Machine Learning” was undertaken and completed the project (10th July, 2022).

Mentor : Mr. Prasad

Date : 10th July,2022

Place : Karur

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | **2** |
|  | **ACKNOWLEDGEMENT** | **3** |
| **1** | **INTRODUCTION** | **7** |
| **2** | **DATA COLLECTION AND DATA PREPARATION** | **9** |
| 2.1 | DESCRPITION OF COLUMNS | **10** |
| 2.2 | DATA PREPROCESSING | **11** |
| 2.3 | NEED OF DATA PREPROCESSING | **11** |
| 2.4 | CHECKING DUPLICATES | **11** |
| 2.5 | HANDLING MISSING VALUES | **12** |
| **3** | **TRAIN VALIDATION SPLIT** | **13** |
| **4** | **FITTING MODELS TO DATA** | **14** |
| 4.1 | DECISION TREE CLASSIFIER | **14** |
| 4.2 | RANDOM FOREST CLASSIFIER | **15** |
| 4.3 | MODEL PERFORMANCE | **17** |
| **5** | **CONCLUSION** | **18** |
| **6** | **REFERENCES** | **19** |

**TABLE OF CONTENTS**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE NO.** | **TITLE** | **PAGE NO.** |
| 1 | ASTEROIDS | 7 |
| 2 | NEAREST OBJECT DETAILS DATASET | 9 |
| 2.4 | CHECKING DUPLICATES | 12 |
| 2.5 | HANDLING MISSING VALUES | 12 |
| 3 | TEST VALIDATION SPLIT | 13 |
| 4.1 | DECISION TREE | 14 |
| 4.1.1 | DECISION TREE CLASSIFIER MODEL | 15 |
| 4.2 | RANDOM FOREST | 16 |
| 4.2.1 | RANDOM FOREST CLASSIFIER MODEL | 16 |
| 4.3.1 | RANDOM FOREST ACCURACY | 17 |
| 4.3.2 | DECISION TREE ACCURACY | 17 |

**CHAPTER 1**

**INTRODUCTION**

Asteroids are small, rocky objects that orbit the Sun. Although asteroids orbit the Sun like planets, they are much smaller than planets. There are lots of asteroids in our solar system. Most of them are located in the main asteroid belt a region between the orbits of Mars and Jupiter. y smashing into the Earth, the asteroid impact released huge amounts of dust into the atmosphere. This dust blocked the sun's rays from reaching the surface of the Earth and this is bad news for plants.

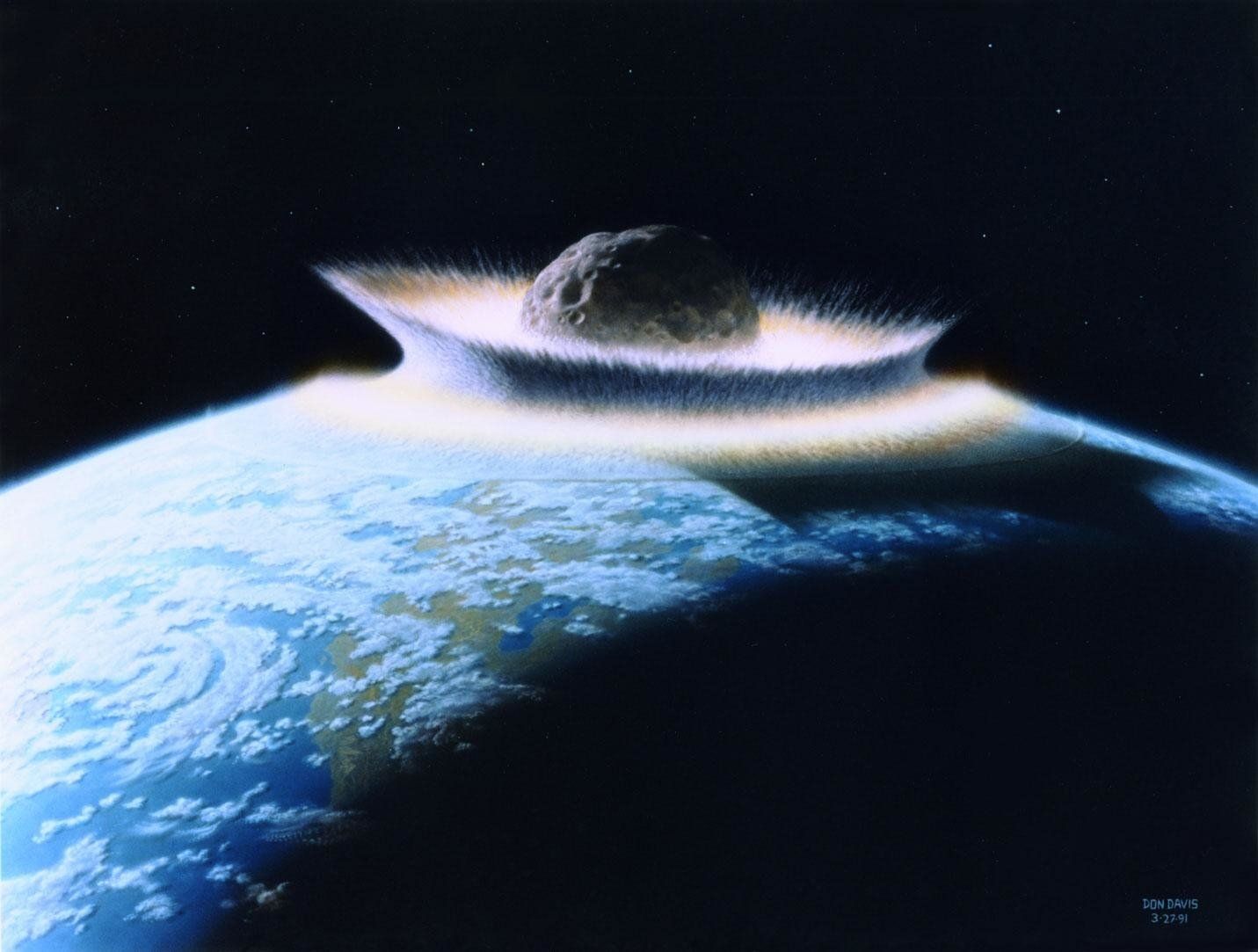


Figure 1: Asteroids

Asteroids are left over from the formation of our solar system. Our solar system began about 4.6 billion years ago when a big cloud of gas and dust collapsed. When this happened, most of the material fell to the center of the cloud and formed the sun. Some of the condensing dust in the cloud became planets. The objects in the asteroid belt never had the chance to be incorporated into planets. They are leftovers from that time long ago when planets formed.

Several NASA space missions have also flown by and observed asteroids. The NEAR Shoemaker spacecraft landed on Eros, an asteroid near Earth, in 2001. Then, the Dawn spacecraft travelled to the asteroid belt in 2011. It orbited and studied the giant asteroid Vesta and the dwarf planet Ceres.

Small objects frequently collide with Earth. There is an inverse relationship between the size of the object and the frequency of such events. The lunar cratering record shows that the frequency of impacts decreases as approximately the cube of the resulting crater's diameter, which is on average proportional to the diameter of the impactor. Asteroids with a 1 km (0.62 mi) diameter strike Earth every 500,000 years on average. Large collisions – with 5 km (3 mi) objects – happen approximately once every twenty million years. The last known impact of an object of 10 km (6 mi) or more in diameter was at the [Cretaceous Paleogene extinction event](https://en.wikipedia.org/wiki/Cretaceous%E2%80%93Paleogene_extinction_event) 66 million years ago.

## CHAPTER 2

**DATA COLLECTION AND DATA PREPARATION**

Dataset for NASA Nearest Object is found from various sources such as from Kaggle, web and browser searches. The Datasets contains 90836 rows and 10 columns. The columns are as id: Unique Identifier for each Asteroid, name : Name given by NASA,est\_diameter\_min : Minimum Estimated Diameter in Kilometres,est\_diameter\_max : Maximum Estimated Diameter in Kilometres,relative\_velocity : Velocity Relative to Earth in Kmph,miss\_distance : Distance in Kilometres missed,orbiting\_body : Planet that the asteroid orbits,sentry\_object : Included in sentry - automated collision monitoring system,absolute\_magnitude : Describes intrinsic luminosity.

# 

Figure 2: Nearest Object Details Dataset

**2.1 DESCRIPTION OF COLUMNS**

* **id:** Unique Identifier for each Asteroid
* **name:** Name given by NASA
* **est\_diameter\_min**: Minimum Estimated Diameter in Kilometres
* **est\_diameter\_max**: Maximum Estimated Diameter in Kilometres
* **relative\_velocity**: Velocity Relative to Earth in Kmph
* **miss\_distance:** Distance in Kilometres missed
* **orbiting\_body**: Planet that the asteroid orbits
* **sentry\_object:** Included in sentry - automated collision monitoring system
* **absolute magnitude**: Describes intrinsic luminosity

# ****2.2 DATA PREPROCESSING****

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

**2.3 NEED OF DATA PREPROCESSING:**

* For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.
* Another aspect is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and best out of them is chosen.

This project performs lot of pre-processing steps. The dataset has been taken from Kaggle where I performed the data pre-processing such as checking duplicates, handling missing values and so on.

**2.4 CHECKING DUPLICATES**

An important part of Data Analysis is analysing duplicate values and removing them. In this model I have checked whether the duplicate values are present are not. Therefore, by checking this I found that my model dataset doesn’t contains any duplicate values.

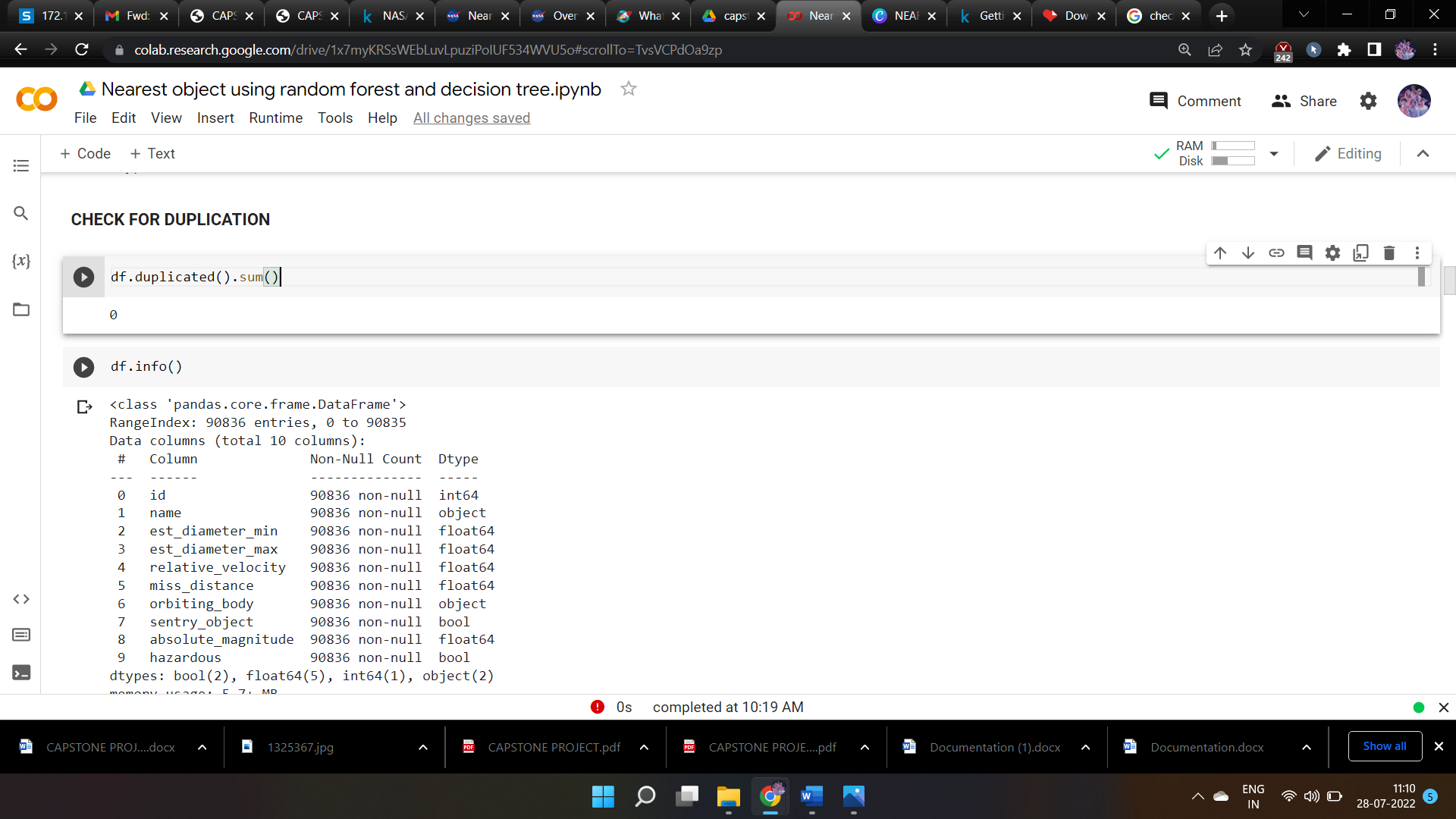


Figure 2.4 Checking Duplicates

**2.5** **HANDLING MISSING VALUES**

In order to check missing values in Pandas Data Frame, we use isnull() and notnull() function. Both Function help in checking whether a value is NaN or not. These Functions can also be used in Pandas seies in order to find null values in a series. In this dataset I have checked whether there exits missing values are not by using isnull() function. It shows that there are no missing values in this dataset.

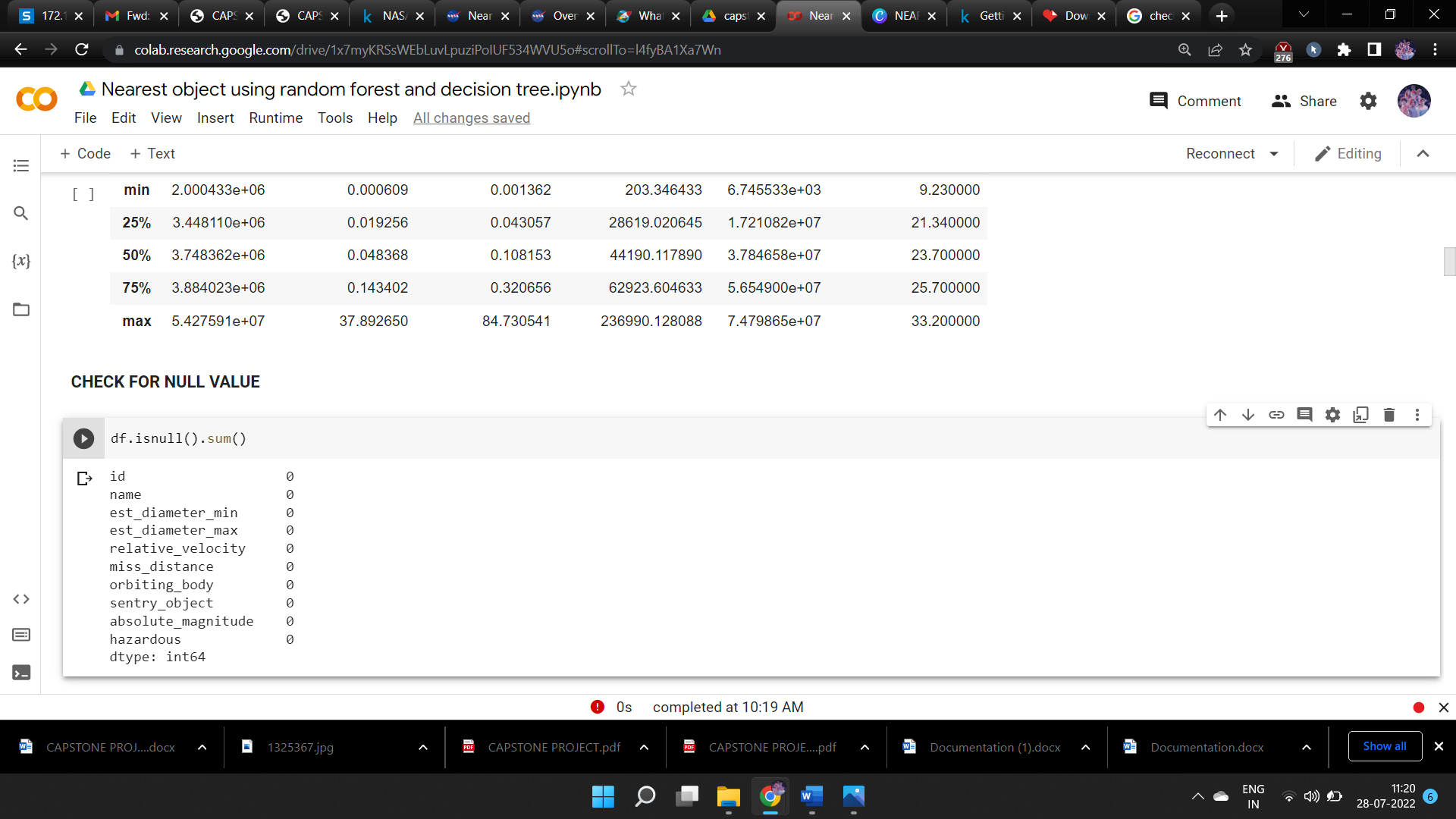


Figure 2.5 Handling Missing Values

# CHAPTER 3

**TRAIN VALIDATION SPLIT**

The dataset was divided into two parts for training and testing purposes. Here we are taking 20% of the dataset for testing and the balance 80% of the dataset for training the model.

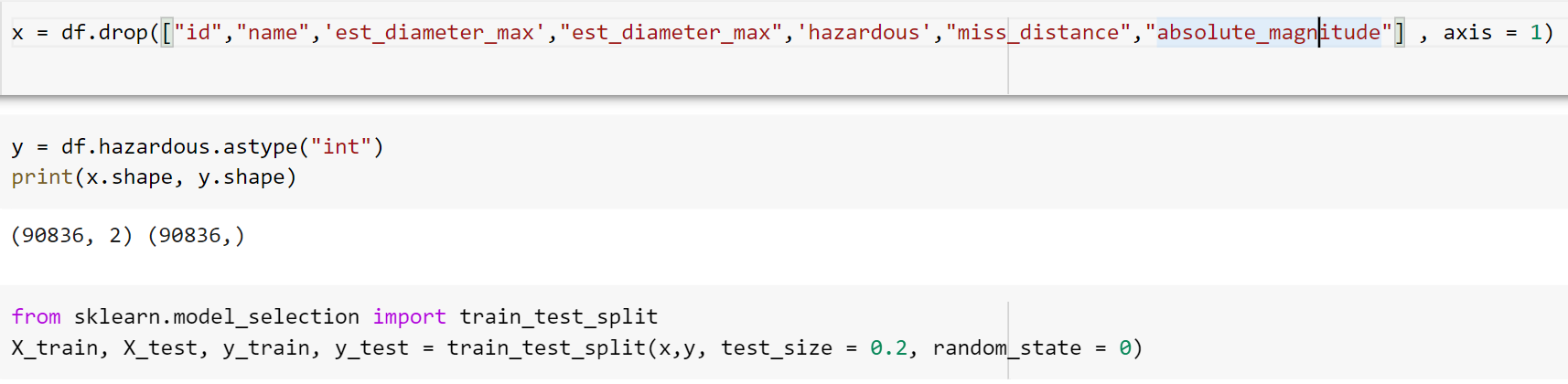


Figure 3: Test Validation Split

## CHAPTER 4

**FITTING MODELS TO DATA**

# 4.1 DECISION TREE ****CLASSIFIER****

Decision tree classifier observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.



Figure 4.1: Decision Tree

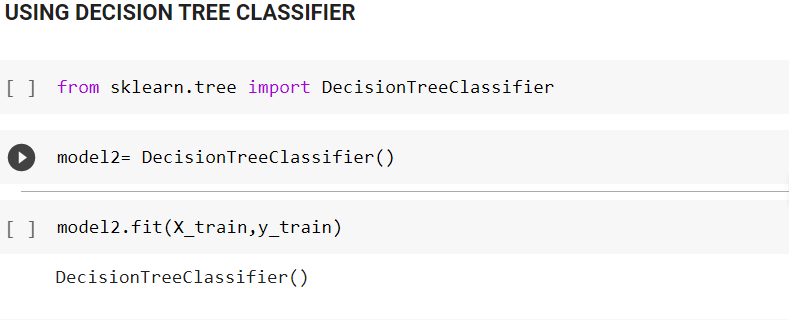


Figure 4.1.1 Decision Tree classifier model

# ****4.2 RANDOM FOREST CLASSIFIER****

# Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn’t depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called **Aggregation**.

# Random Forests in Machine Learning | Random Forests for Data Science

# Figure 4.2 Random Forest

# 

# Figure 4.2.1 Random Forest Classifier Model

# 

# ****4.3 MODEL PERFORMANCE****

# By analyzing the Decision Tree Classifier and Random Forest Classifier performance the Random Forest algorithm give more prediction accuracy than Decision Tree. So, we can choose Random Forest Classifier for our model.

# Random Forest Classifier Accuracy

# 

# 

# Figure 4.3.1 Random Forest Accuracy

# Decision Tree Classifier Accuracy

# 

# Figure 4.3.2 Decision Tree Accuracy

## CHAPTER 5 CONCLUSION

I have evaluated two kinds of machine learning algorithm for NASA nearest object using Decision Tree and Random Forest. In this project, I have collected the datasets from Kaggle. I have performed some important processes like handling missing value, data type conversion, and EDA process. Then I have splitted the dataset into two parts 80% for training and 20% for testing. At last, I fit the dataset to the model and evaluated the performance of the model. Finally, the Random Forest Classifier gives the more accuracy while compare to the Decision Tree Classifier. Therefore, Random Forest classifier is the best classification model for the dataset which I have uploaded for Nasa Nearest object. It has the accuracy score of 0.89 for the test sets.

## CHAPTER 6 REFERENCE

[1] Sameerchand Pudaruth “Nasa nearest object prediction”, International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764

[2] Mariana Listiani, 2009. “K means clustering Analysis for Nasa nearest object prediction”. Master Thesis. Hamburg University of Technology.

[3] Kanwal Noor and Sadaqat Jan, “prediction of asteroids using Machine Learning Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 167 – No.9, June 2017.s