Hazardous Asteroids Prediction

A20466182 - Niveditha Mangala Venkatesha
A20293727 - Saipooja Kandala
A20496724 - Kajol Tanesh Shah
A20482282 - Sravani Bollisetti
A20473893 - Surya Thatee

1) Project Proposal

i. A formal description of the project with a stated research goal:

- Asteroids are large chunks of rock that float around in space. Normally, these cosmic monoliths reside in a vast belt between Mars and Jupiter, but some are occasionally knocked loose. When they reach close enough to Earth, they are frequently trapped in our gravity and broken up by the atmosphere, resulting in pieces of fiery rocks. Fortunately, most of the asteroids burn up in the atmosphere and fall to Earth in the form of ash or small rocks. Larger space rocks, on the other hand, pose a genuine threat. There are 2,203 known potentially dangerous asteroids (approximately 8% of the total near-Earth population) as of August 2021, with 159 predicted to be bigger than one kilometre in diameter. And our objective is to build a model that can detect whether asteroids are hazardous or not.

ii. A specific question or set of questions that the project seeks to address:

- What effect do the dataset's other attributes have on determining whether an asteroid is hazardous or not?
- Determining whether an asteroid is dangerous or not?
- Classifying asteroids according to orbit classes and determining which orbit classes contain hazardous asteroids?

iii. A proposed methodology/approach to the analysis that will be performed:

- Data Gathering.
- Preparing the data include cleaning and dealing with missing values in the dataset.
- Using R visualization tools to visualize the data.
- Choosing significant and useful model features.
- Dividing the dataset into a training and a testing set.
- Selection of a machine learning model and training the model with the train dataset.
- Putting the model to the test with a test dataset.

- Checking the model's correctness by comparing the anticipated and actual output values for the test data.

iv. A metric or set of metrics which will measure analysis results:

- The Root mean square error metric is commonly used to measure the discrepancies between values predicted by a model and values observed.
- The mean absolute error metric will be used to calculate the difference between two continuous variables.
- To determine the performance feature, precision, recall, and F1 measures will be used.

2) Project Outline

i. Literature review and related work

- https://www.researchgate.net/publication/338489667_Identifying_Earth-impacting_asteroids_using_an_artificial_neural_network
- https://www.nasa.gov/sites/default/files/atoms/files/nasem_report_finding_hazardous_asteroids.pdf
- https://www.spacesafetymagazine.com/space-hazards/asteroid-hitting-earth/identifying-potentially-dangerous-asteroids/

ii. All data sources and reference data with descriptions

- https://www.kaggle.com/sakhawat18/asteroid-dataset

Data Description:

1) SPK-ID: Object primary SPK-ID
2) Object ID: Object internal database ID
3) Object full name: Object full name/designation
4) pdes: Object primary designation
5) name: Object IAU name
6) NEO: Near-Earth Object (NEO) flag
7) H: Absolute magnitude parameter
8) Diameter: object diameter (from equivalent sphere) km Unit
9) Albedo: Geometric albedo
10) Diameter_sigma: 1-sigma uncertainty in object diameter km Unit
11) Orbit_id: Orbit solution ID
12) Epoch: Epoch of osculation in modified Julian day form
13) Equinox: Equinox of reference frame
14) e: Eccentricity

15) a: Semi-major axis au Unit
16) q: perihelion distance au Unit
17) i: inclination; angle with respect to x-y ecliptic plane
18) tp: Time of perihelion passage TDB Unit
19) moid_ld: Earth Minimum Orbit Intersection Distance au Unit

iii. Data processing and pipeline - cleaning, imputing, transformation, outlier detection, etc.

- Data import – Importing the data from kaggle data source.
- Data Cleaning - removing characters from keyword like null values, plot columns and duplicates.
- Data preview and Analysis based on Data Description after replacing the missing values with suitable value.

iv. Data stylized facts - distributional analysis, clustering, dimensionality reduction, etc.

- Data Visualization - Bar charts, Line graph, Histograms Boxplots and Correlation Plot**.**

v. Model selection

- As this is a classification problem, we will use the Random forest method and logistic regression because they perform the best in this situation, according to the research. We will also use Artificial Neural Network algorithm to determine which of these two methods performs better.

vi. Software packages, applications, libraries, and associated tools, etc.

- We'll be using the R programming language, and the packages listed below will be used in R studios.
  - randomForest
  - stats
  - caret
  - dplyr
  - neuralnet
  - math
  - ggplot2