Asteroid Diameter Prediction

Project Report

[5/12/2021]

Spartificial

Prepared by: Pratham Grover (Delhi Technological University)
Mail Id: prathamgrover777@gmail.com

Introduction

Machine Learning is quickly becoming a popular method to analyze astronomical data. The amount of data collected by astronomers has grown beyond what humans can evaluate without assistance as they build larger observatories capable of spotting more things in the sky. Instead, scientists train computers to filter through data, uncovering crucial patterns and relationships that might otherwise go unnoticed.

In astrophysics and cosmology, vast, sophisticated, and multidimensional data sets must be analyzed. Data description and interpretation, pattern recognition, prediction, classification, compression, inference, and other tasks are common in such analysis. The employment of machine-learning technologies is one way to accomplish such jobs.

The purpose of supervised learning is to infer a function from labelled training data, which is a collection of training samples. Each example has known 'input' quantities whose values are to be used to forecast the 'outputs' values. As a result, the mapping from input to output is the function to be inferred. This mapping can be applied to datasets with unknown output values once it has been learned. Classification and regression are two common subtypes of supervised learning.

Scientific terms

Semi-major axis (au):

The semi-major axis is one of the most important orbital elements of an orbit, along with its orbital period. For solar system objects, the semi-major axis is related to the period of the orbit by Kepler's third law

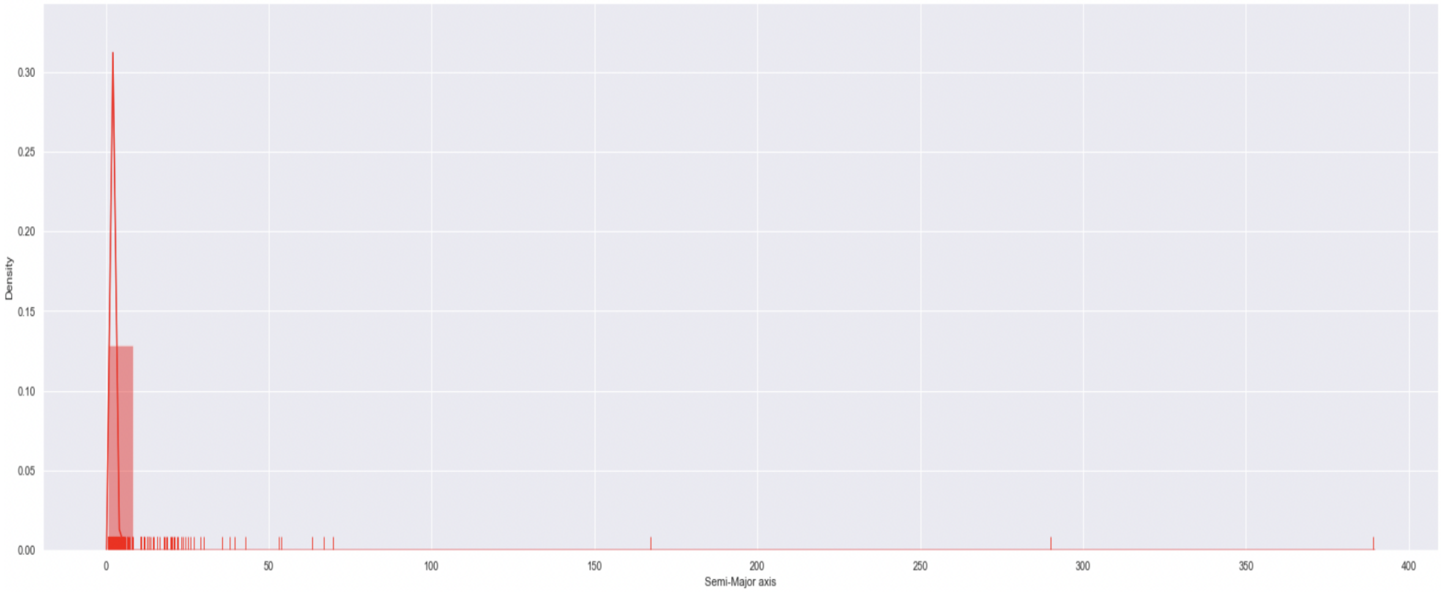
$$T^2 = a^3$$

where T is the period in years, and a is the semimajor axis in astronomical units. This form turns out to be a simplification of the general form for the two-body problem, as determined by Newton:

$$T^2 = \frac{4\pi^2}{G(M+m)} a^3$$

where G is the gravitational constant, and M is the mass of the central body, and m is the mass of the orbiting body. Typically, the central body's mass is so much greater than the orbiting bodies, that m may be ignored.

The semi-major axis used in astronomy is always the primary-to-secondary distance; thus, the orbital parameters of the planets are given in heliocentric terms.



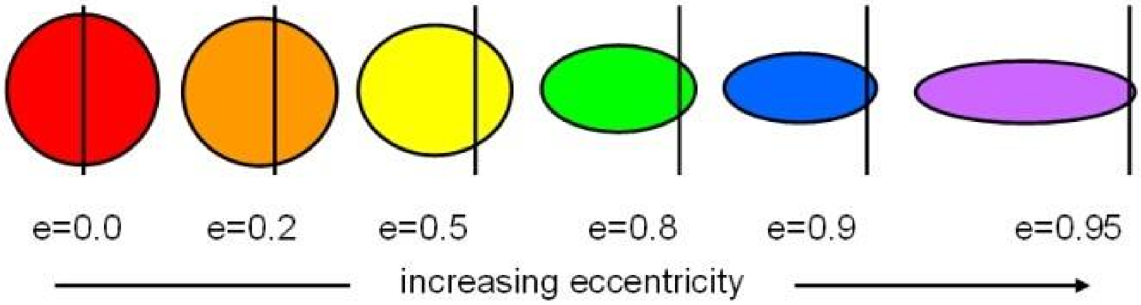
Eccentricity (e):

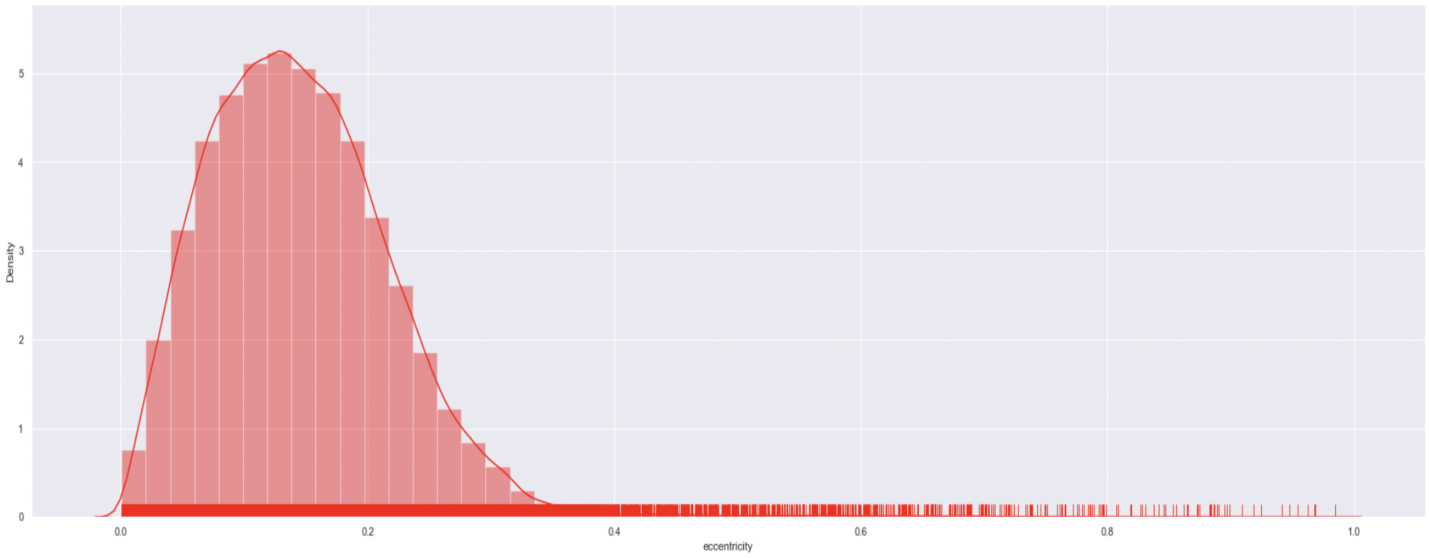
The orbital eccentricity (or eccentricity) is a measure of how much an elliptical orbit is ‘squashed’. It is one of the orbital elements that must be specified in order to completely define the shape and orientation of an elliptical orbit.

The equation of an ellipse in polar coordinates is:

$$r = \frac{a(1-e^2)}{1+e\cos\theta}$$

where a is the semi-major axis, r is the radius vector, is the true anomaly (measured anticlockwise) and e is the eccentricity. An ellipse has an eccentricity in the range 0 < e < 1, while a circle is the special case e=0.

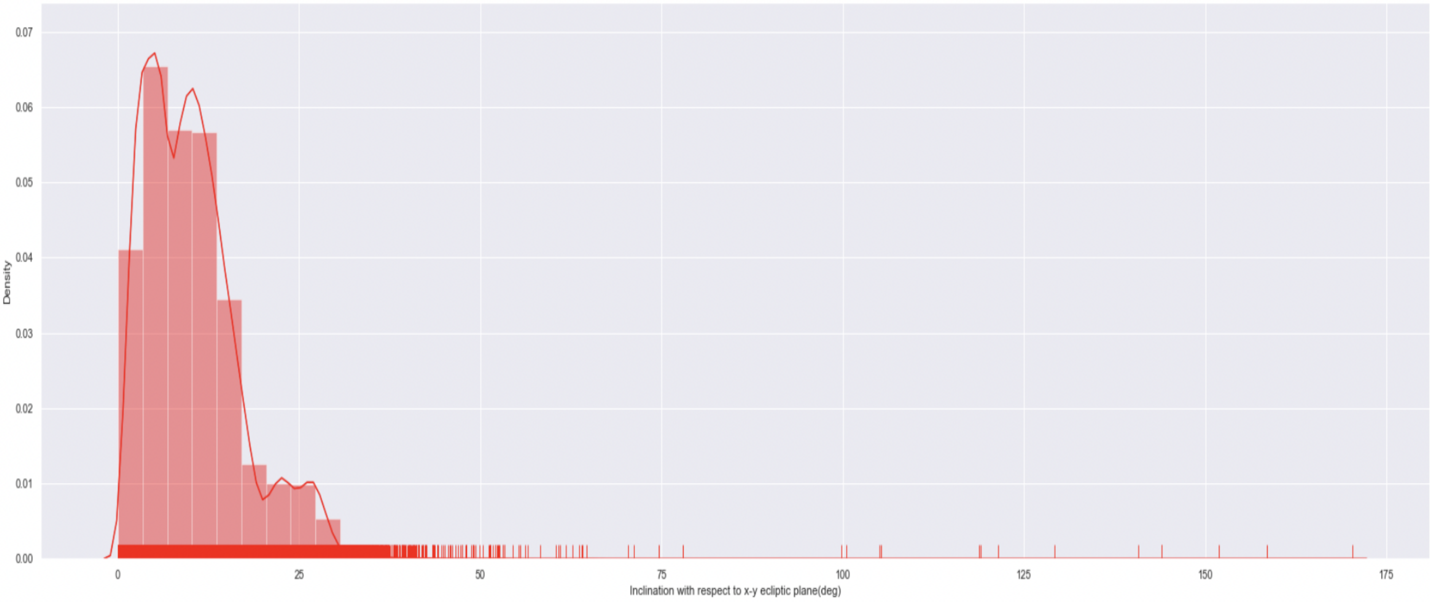




Inclination with respect to x-y ecliptic plane (i):

The inclination is one of the six orbital elements describing the shape and orientation of a celestial orbit. It is the angle between the orbital plane and the plane of reference, normally stated in degrees. Inclination can instead be measured with respect to another plane, such as the Sun's equator or the invariable plane

- An inclination of 30° could also be described using an angle of 150°. The convention is that the normal orbit is prograde, an orbit in the same direction as the planet rotates. Inclinations greater than 90° describe retrograde orbits. Thus:
- An inclination of 0° means the orbiting body has a prograde orbit in the planet's equatorial plane.
- An inclination greater than 0° and less than 90° also describes a prograde orbit.
- An inclination of 63.4° is often called a critical inclination, when describing artificial satellites orbiting the Earth, because they have zero apogee drift.
- An inclination of exactly 90° is a polar orbit, in which the spacecraft passes over the poles of the planet.
- An inclination greater than 90° and less than 180° is a retrograde orbit.
- An inclination of exactly 180° is a retrograde equatorial orbit.



The longitude of the ascending node (om) :

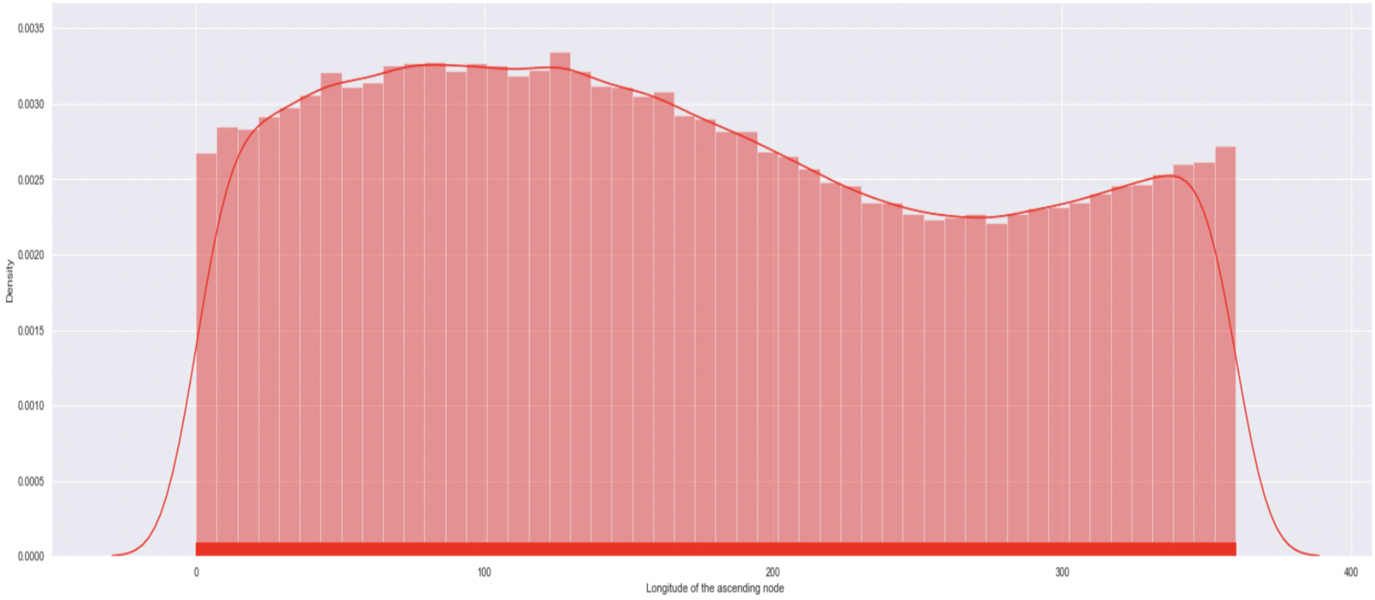
It is one of the orbital elements used to specify the orbit of an object in space. It is the angle from a specified reference direction, called the origin of longitude, to the direction of the ascending node, as measured in a specified reference plane.

The ascending node is the point where the orbit of the object passes through the plane of reference, as seen in the adjacent image. Commonly used reference planes and origins of longitude include:

- **For geocentric orbits,** Earth's equatorial plane as the reference plane, and the First Point of Aries as the origin of longitude. In this case, the longitude is also called the right ascension of the ascending node (RAAN). The angle is measured eastwards (or, as seen from the north, counterclockwise) from the First Point of Aries to the node.

- **For heliocentric orbits,** the ecliptic as the reference plane, and the First Point of Aries as the origin of longitude. The angle is measured counterclockwise from the First Point of Aries to the node.

- **For orbits outside the Solar System,** the plane tangent to the celestial sphere at the point of interest as the reference plane, and north as the origin of longitude. The angle is measured eastwards from north to the node.



Geometric albedo (albedo):

Albedo is a measurement of the amount of light reflected from the surface of a celestial object, such as a planet, satellite, comet or asteroid. The albedo is the ratio of the reflected light to the incident light:

$$A = \frac{\text{reflected light}}{\text{incident light}}$$

and has values between:

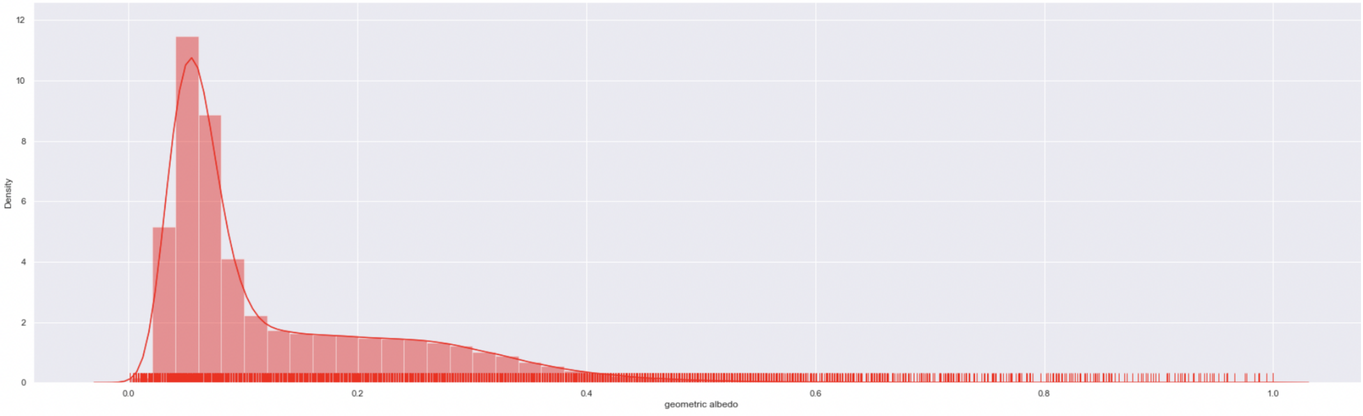
0: a black object that absorbs all light and reflects none; and

1: a white object that reflects all light and absorbs none.

The expression for diameter d in km as a function of absolute magnitude H and geometric albedo is given by the following equation.

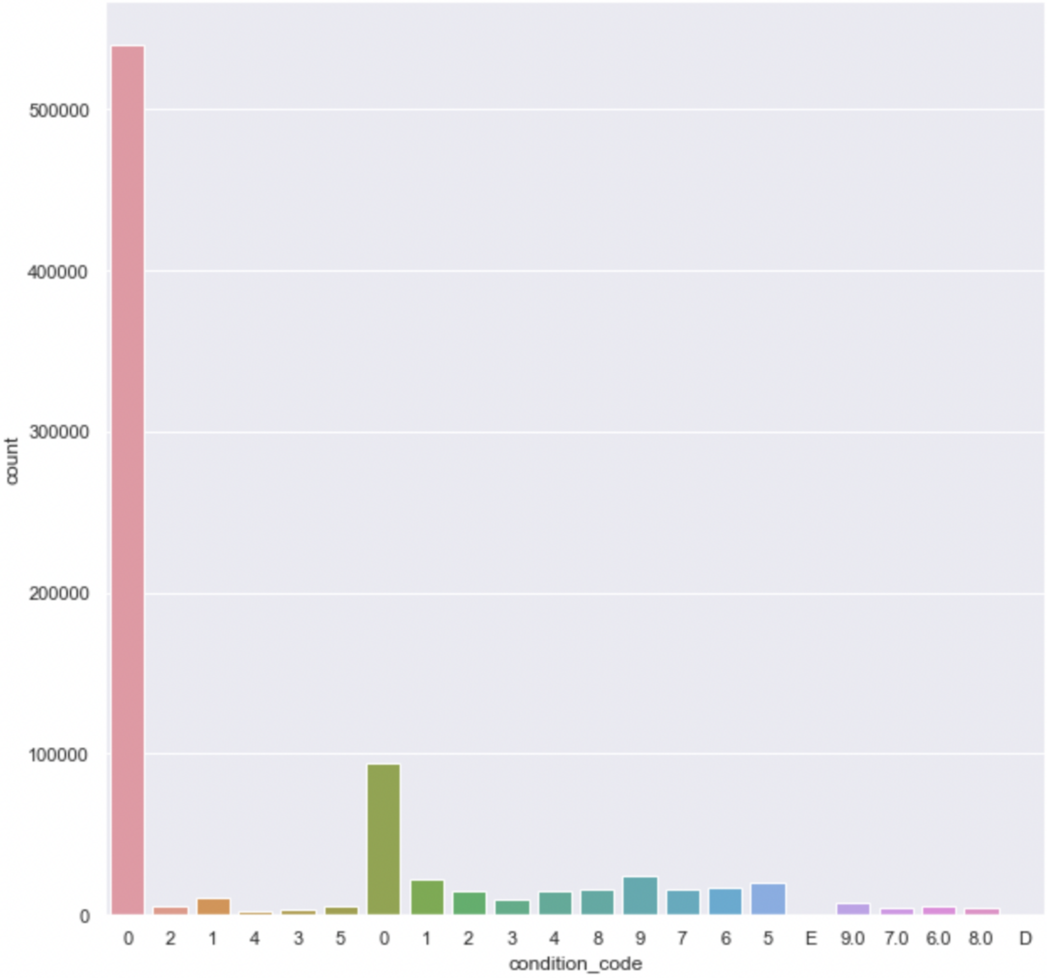
$$d = 10^{[\,3.1236\,-\,0.5\,\log_{10}(a)\,-\,0.2H\,]}$$

The above expression assumes a spherical object with a uniform surface (no albedo variation). When using this expression to estimate the size of an object, it is important to consider the uncertainty in H as well as the uncertainty in albedo (typically assumed based on some spectral class corresponding to an assumed composition of the object - e.g., S-class asteroid with an assumed albedo of 0.15).



Orbit Condition Code(condition_code):

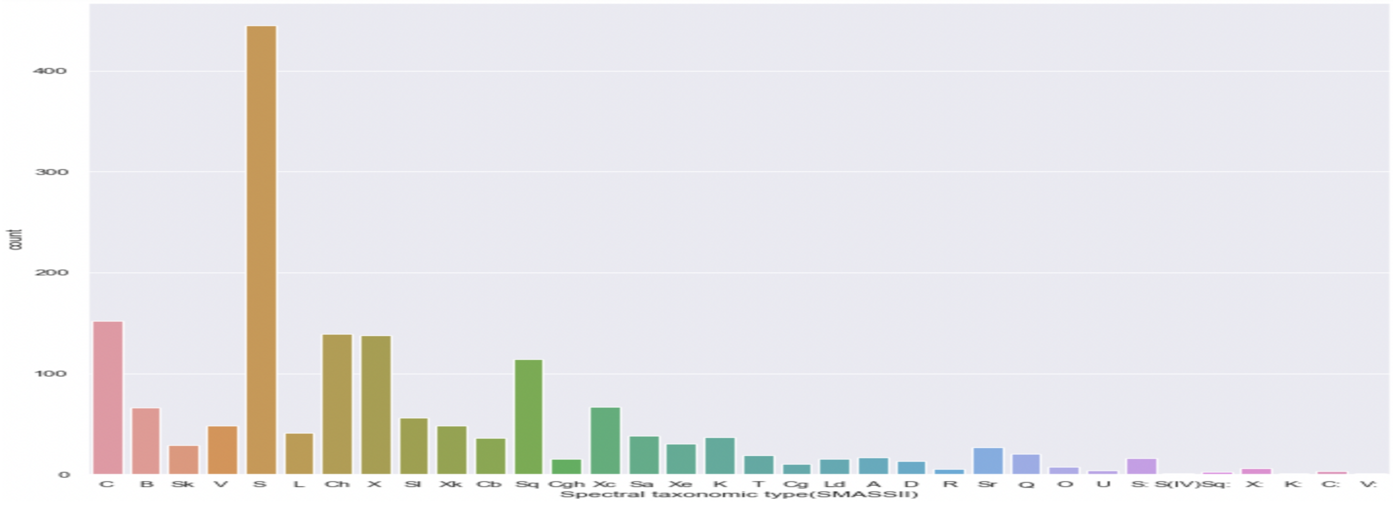
The condition code is a measurement of the uncertainty of the object's orbit. It ranges from a low of "0", meaning certainty about the orbit, to a high of "9", meaning high uncertainty. Additional measurements over time can be used to more accurately predict the object's orbit.



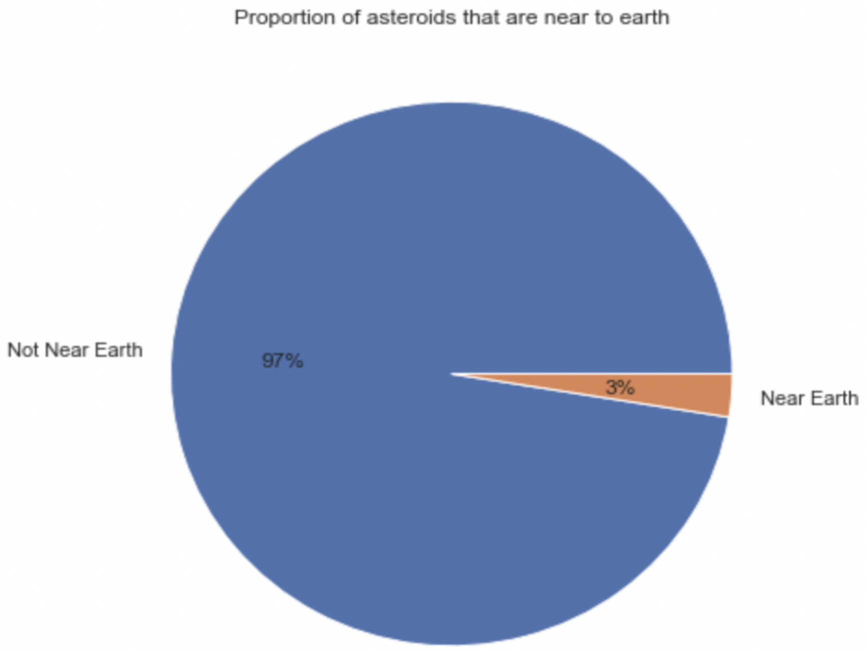
Spectral taxonomic type(SMASSII):

Asteroids are assigned a type based on spectral shape, color, and sometimes albedo. This scheme includes 14 types with the majority of asteroids falling into one of three broad categories, and several smaller types. They are, with their largest exemplars:

- C-group dark carbonaceous objects, including several sub-types:
    - B-type (2 Pallas)
    - F-type (704 Interamnia)
    - G-type (1 Ceres)
    - C-type (10 Hygiea) the remaining majority of 'standard' C-type asteroids. This group contains about 75% of asteroids in general.
- S-type (15 Eunomia, 3 Juno) silicaceous (i.e. stony) objects. This class contains about 17% of asteroids in general.
- X-group
    - M-type (16 Psyche) metallic objects, the third most populous group.
    - E-type (44 Nysa, 55 Pandora) differ from M-type mostly by high albedo
    - P-type (259 Aletheia, 190 Ismene; CP: 324 Bamberga) differ from M-type mostly by low albedo and the small classes:
- A-type (446 Aeternitas)
- D-type (624 Hektor)
- T-type (96 Aegle)
- Q-type (1862 Apollo)
- R-type (349 Dembowska)
- V-type (4 Vesta)

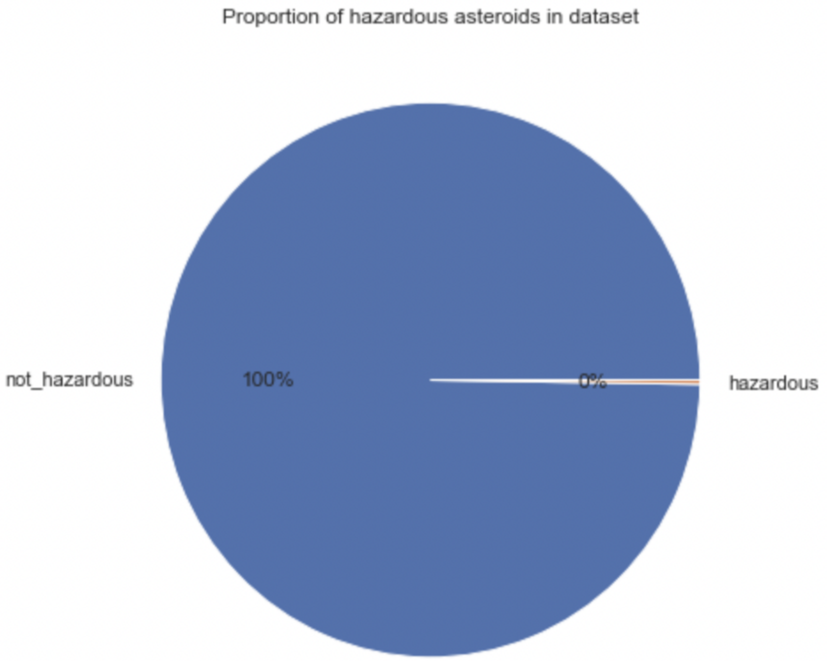


Near Earth Objects (neo):



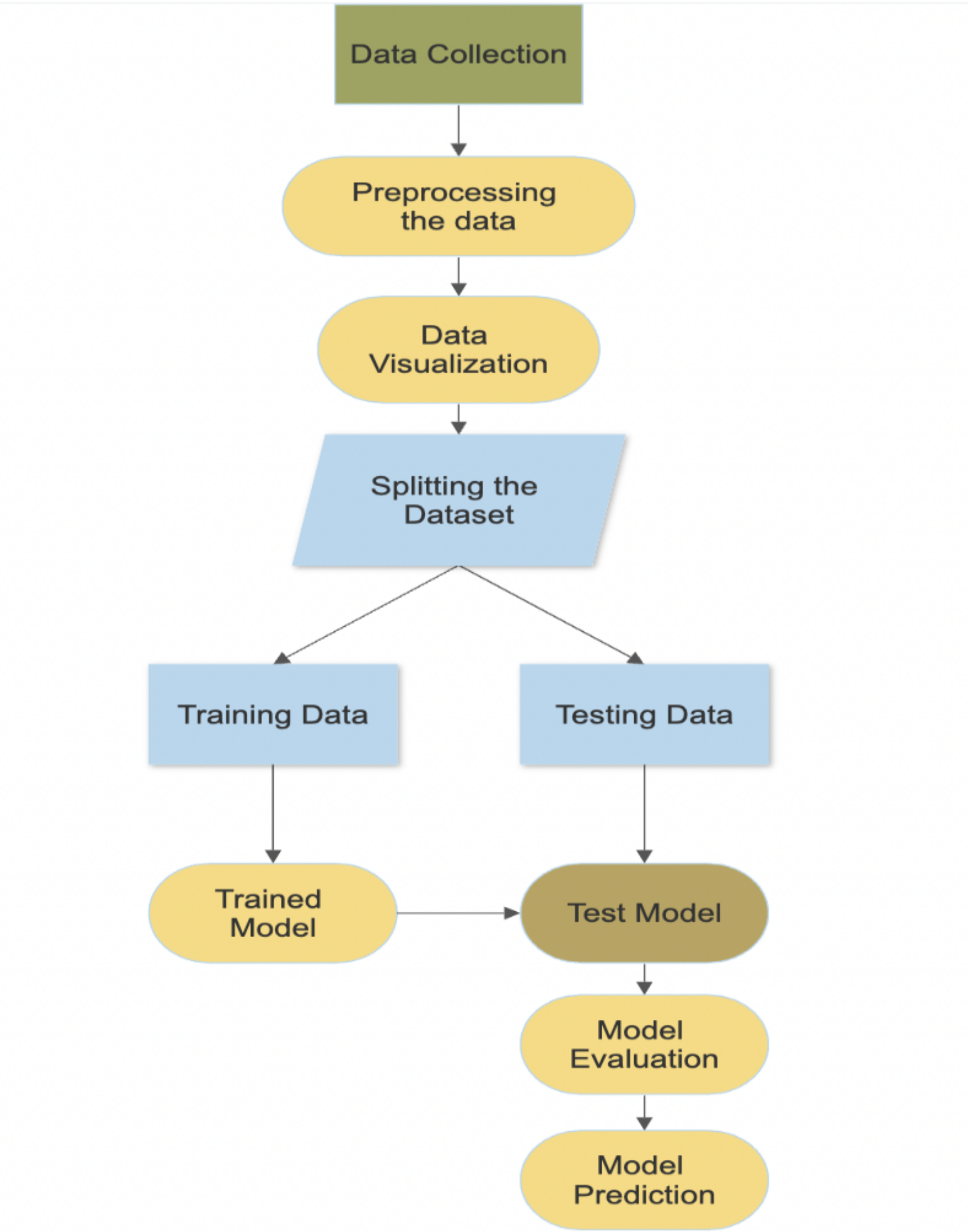
```
There are 21389  potential asteroids that are near to earth and 818341 that are not .
```

Physically Hazardous Asteroids(Pha):



```
There are 2014  potential asteroids that can collide and 820800 that might miss the collision.
```

Flowchart


Data Collection
Preprocessing the data
Data Visualization
Splitting the Dataset
Training Data
Testing Data
Trained Model
Test Model
Model Evaluation
Model Prediction

Libraries Used

Standard Libraries:

- Pandas: Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

- Numpy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

- Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

- Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- Math: The Python Math Library provides us access to some common math functions and constants in Python, which we can use throughout our code for more complex mathematical computations.

Sci-kit

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Metrics functions : The sklearn.metrics module implements functions assessing prediction error for specific purposes.

Model selection: Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

Preprocessing: This package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

Tree: Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Ensemble: The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability / robustness over a single estimator.
Two families of ensemble methods are usually distinguished:

- In **averaging methods**, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimators because its variance is reduced.
  Examples: Bagging methods, Forests of randomized trees etc

- In **boosting methods**, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.
  Examples: AdaBoost, Gradient Tree Boosting etc

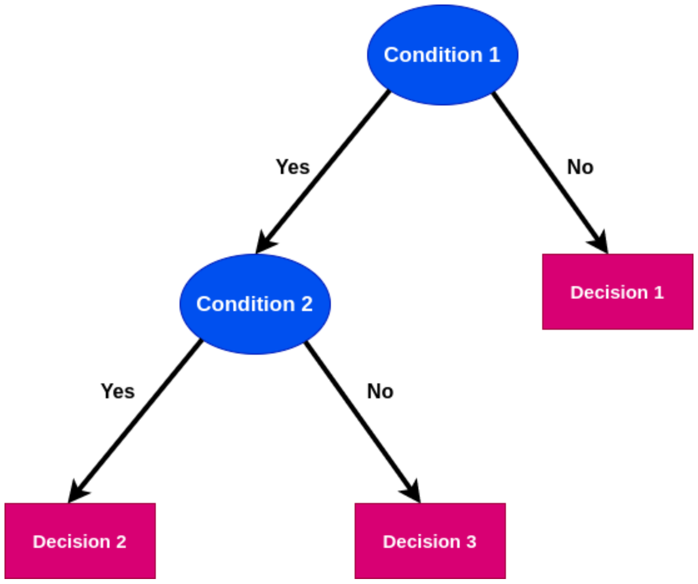
Neural network: A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropagation for training the network.

Data Modeling

Decision Tree:

A decision tree is a hierarchical data structure implementing the divide and conquer strategy. It is an efficient non-parametric method, which can be used for both classification and regression. This includes learning algorithms that build the tree from a given labeled training sample, as well as how the tree can be converted to a set of simple rules that are easy to understand.
In this hierarchical model for supervised learning whereby the local region is identified in a sequence of recursive splits in a smaller number of steps. Decision tree is composed of internal decision nodes and terminal leaves. Each decision node m implements a test function with discrete outcomes labeling the branches.
Given an input, at each node, a test is applied and one of the branches is taken depending on the outcome. This process starts at the root and is repeated recursively until a leaf node is hit,at which point the value written in the leaf constitutes the output.

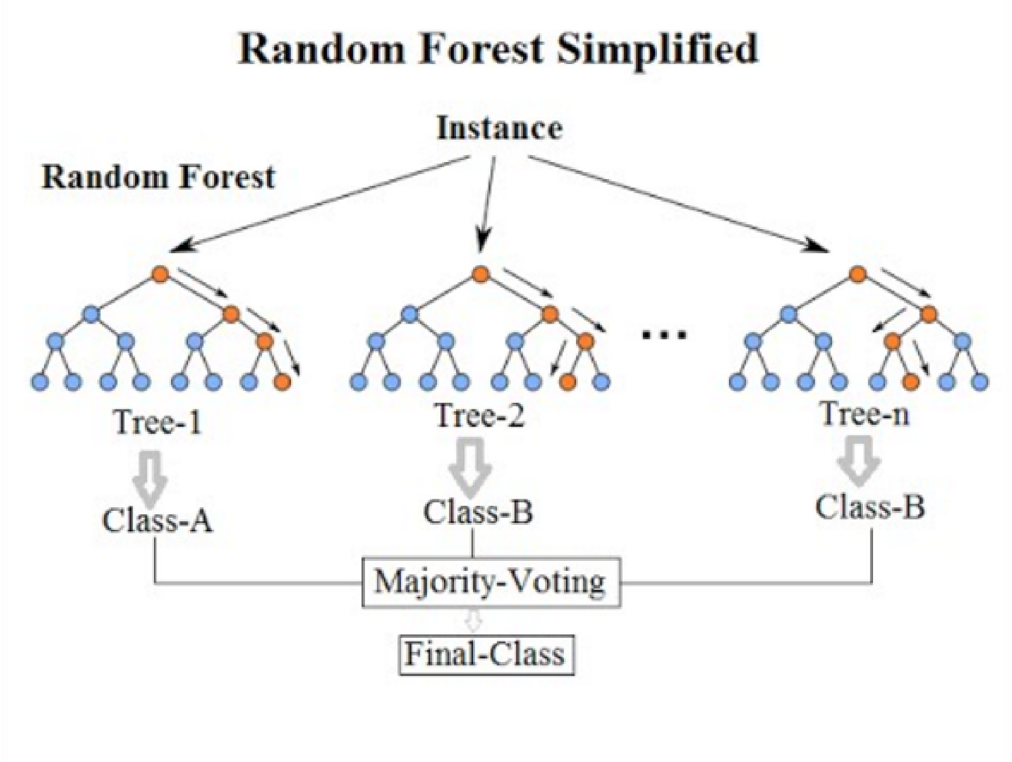
Regression Tree: A regression tree is constructed in almost the same manner as a classification tree, except the impurity measure that is appropriate for classification is replaced by a measure appropriate for regression.



Random forest:

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or “the random subspace method”, generates a random subset of features, which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is made up of a bootstrap sample, which is a data sample obtained from a training set with replacement. One-third of the training sample is set aside as test data, referred to as the out-of-bag (oob) sample. Using feature bagging, another instance of randomness is injected into the dataset, increasing the dataset's variety and decreasing the correlation between decision trees. The prediction will be determined differently depending on the type of difficulty. Individual decision trees will be averaged for a regression job. Finally, the oob sample is used for cross-validation, bringing the prediction to a conclusion.

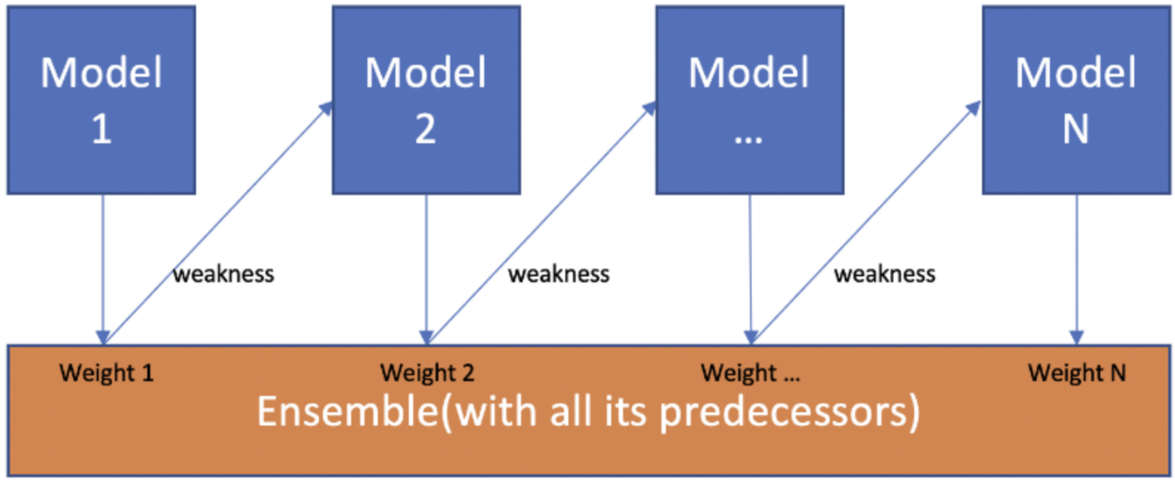


AdaBoost:

The AdaBoost algorithm, short for Adaptive Boosting, is a Boosting approach used in Machine Learning as an Ensemble Method. The weights are re-allocated to each instance, with higher weights applied to improperly identified instances. This is termed Adaptive Boosting. In supervised learning, boost is used to reduce bias and variation. It is based on the notion of successive learning. Each subsequent student, with the exception of the first, is grown from previously grown learners. In other words, weak students are transformed into strong students.

During the data training period, it creates a certain number of decision trees. The improperly categorised record in the first model is given priority as the first decision tree/model is constructed. Only these records are sent to the second model as input. The procedure continues until we have decided on a number of base learners to develop.

Model 1,2,..., N are individual models (e.g. decision tree)



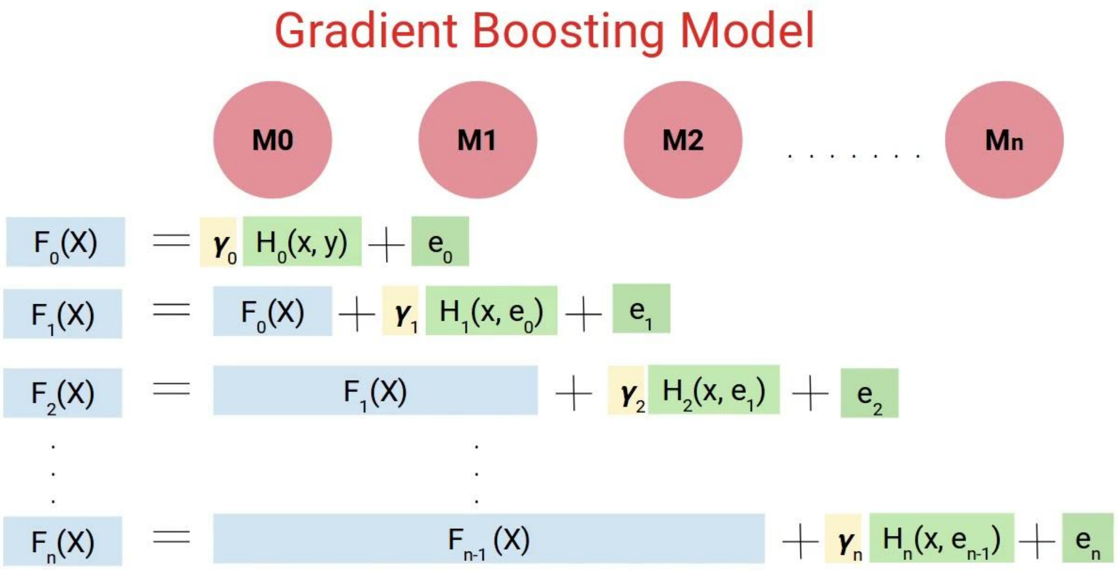
Source: Google

Gradient Boosting:

One of the most powerful algorithms in the field of machine learning is the gradient boosting technique. As we all know, machine learning algorithm faults can be divided into two categories: bias error and variance error. Gradient boosting is one of the boosting strategies that is used to reduce the model's bias error.

The base estimator of the gradient boosting technique, unlike the Adaboosting algorithm, cannot be mentioned. The Gradient Boost algorithm's base estimator is fixed, i.e. Decision Stump. The gradient boosting algorithm's n estimator can be tuned, just like AdaBoost. The default value of n estimator for this algorithm is 100 if the value of n estimator is not specified.

Gradient boosting methods can be used to forecast not just continuous but also categorical target variables (as a Regressor) (as a Classifier). Mean Square Error (MSE) is the cost function when used as a regressor, and Log loss is the cost function when used as a classifier.

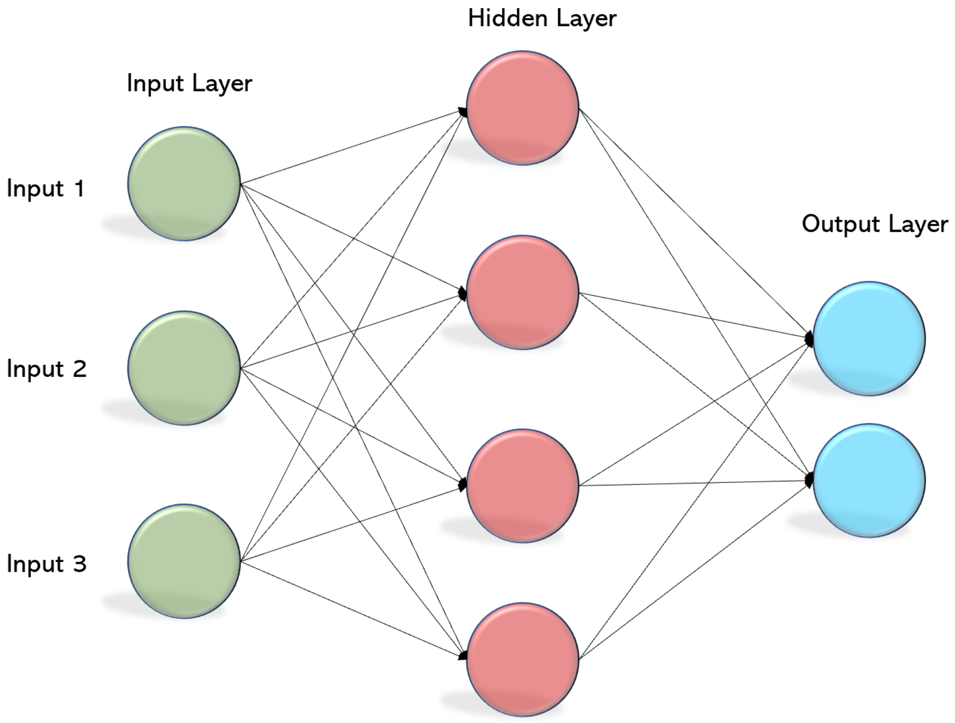

Gradient Boosting Model
M0
M1
M2
Mn
$F_0(X) = \gamma_0\ H_0(x, y) + e_0$
$F_1(X) = F_0(X) + \gamma_1\ H_1(x, e_0) + e_1$
$F_2(X) = F_1(X) + \gamma_2\ H_2(x, e_1) + e_2$
$F_n(X) = F_{n-1}(X) + \gamma_n\ H_n(x, e_{n-1}) + e_n$

Multi-Layer Perceptron:

It's a neural network with a non-linear mapping between inputs and outputs. Input and output layers, as well as one or more hidden layers with numerous neurons layered together, make up a Multilayer Perceptron. While neurons in a Perceptron must have an activation function that enforces a threshold, such as ReLU or sigmoid, neurons in a Multilayer Perceptron can have any activation function they like.

Because inputs are integrated with initial weights in a weighted sum and applied to the activation function, the Multilayer Perceptron falls under the category of feedforward algorithms. Each linear combination, on the other hand, gets propagated to the next layer.

Each layer feeds the output of its computation, or internal representation of the data, to the next. This applies to all hidden layers as well as the output layer.



Outcome and Future:

Artificial Intelligence's (AI) key goal is for computers to not only be evolved enough to execute activities that humans do, but also to do them in a fashion that humans do, involving 'thinking' and action generated from intelligence.

This is my first machine learning internship, and I found all of the data and did all of the analysis myself. I am quite pleased with my results and invite anyone to strive to improve on or find anything better than my model.

The goal of machine learning in data-driven astronomy is to make use of the ever-increasing volumes of data collected by modifying and analysing it without relying heavily on human input.