BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Finding Habitable Exo Planets Using Boosting Algorithm

Authors
Md. Mashfiq Rahman
Naba Afrin

Supervisor
Dr. Mahbub Alam Majumdar
Professor
Department of CSE

A thesis submitted to the Department of CSE
in partial fulfillment of the requirements for the degree of
B.Sc. Engineering in CSE

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

Declaration

It is hereby declared that this thesis /project report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Md. Mashfiq Rahman
Student ID: 12301006

Naba Afrin
Student ID: 14301090

Supervisor:

Dr. Mahbub Alam Majumdar
Professor, Department of Computer Science and Engineering
BRAC University

December 2018

The thesis titled Finding Habitable Planets Using Boosting Algorithom
Submitted by:
Md. Mashfiq Rahman Student ID: 12301006
Naba Afrin Student ID: 14301090
of Academic Year ..... has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of ..... (An example is shown. It must be replaced by the appropriate Board of Examiners)

1.	Dr. Mahbub Alam Majumdar Professor, Department of Computer Science and Engineering, BRAC University	Chairman
2.	Name of Co-Supervisor Designation Address	Member
3.	Name of Head of the Dept. Designation Address	Member
4.	Name of Internal Member Designation Address	Member
5.	Name of Internal Member Designation Address	Member
6.	Name of External Member Designation Address	Member

Acknowledgements

The person we first like to show our most cordial gratitude to is our respected thesis supervisor DR. Mahbub Alam Majumdar, who helps us immensely to complete our thesis. Without his support we might not able to do our thesis. Secondly we want to mention the help from Dr. Iftekharul Mobin sir for guiding us through our process.Thirdly we want to acknowledge Moin Mostakim sir for his guidance. We also like to acknowledge our parents for their love and care. and we also want to be thankful to our almighty for giving us the ability to do our thesis.

Abstract

The first moment man extended their boundaries outside of our Earth, from that moment they were looking for another habitable planet, where they may live in future. Including NASA, many international space organization already sent a number of satellite on this mission. These mission have discovered thousands of new planetary candidates, many of which have been confirmed through follow up observations. A primary goal of the mission is to determine the occurrence rate of terrestrial-size planets within the Habitable Zone (HZ) of their host stars. Though many approaches have been taken to confirm their habitability, we tried a new approach by using boosting algorithms. We use the NASAs Extra Solar planets dataset of 3,577 planets and use their various characteristics like their Mass, Radius, Orbital Eccentricity, Temperature, Metallicity to determine the best set of alternatives of Earth. We classified the dataset based on these variables and used Extreme Gradient Boosting to compare the accuracy to find out our desirable results. We used different classifier to ensure the best accuracy. So we used Ada Boosting Classifier, KNeighbor's Nearest Classifier (KNN), Gradient Boosting Classifier, Decision Tree Classifier and Random Forest Classifier into our dataset.

Table of contents

List of figures

Chapter 1

Getting Started

1.1 Introduction

Finding a second Earth with the goal that one day our Future ages can live there is the long lasting dream of numerous Researcher. All universal space associations have sent numerous missions to locate an appropriate tenable exoplanet. The primary verification of an exoplanet was noted as ahead of schedule as 1917[16], yet was not licensed in that capacity. Nonetheless, the principal logical acknowledgment of an exoplanet was in 1988[16], despite the fact that it was not set up to be an exoplanet until some other time in 2012. The main affirmed location happen in 1992. Starting at 1 November 2018, there are 3,874 affirmed planets in 2,892 sun based systems[3][5], with 638 frameworks having more than one exoplanet. Already researcher utilized numerous ability to decide the livability of those exoplanets; in this paper we attempted the a standout amongst the most prevalent strategy for current occasions, Machine Figuring out how to get the best precision. It was likewise our worry that cosmology is a standout amongst the most immature fields of investigation of science and applying machine learning will open a ton of entryways in this field. If at any time the day comes when human need to live Earth and begin their Voyage for an options our strategy will be an extraordinary help to set their goal.

1.2 Motivation

From the very childhood, stars were a huge attraction for human kinds. If someone ever lying on a field and look into the night sky s/he will be ensured to mesmerized by the beauty of the stars. It also brought the questions in our mind that what if there are living things on other planets or if we can live on other planets also. So human are always interested on the huge

galaxy we have and what mystery is hidden there. Our main motivation is also to unlock the doors of cosmology to modern methods of Computer science so in future we can discover some of this hidden mystery. With the recent popularity and development in the areas of machine learning, we can see its potential to get more accurate data analyzing and also to great optimized performance. Also the progress rate of other field of studies already is very slower where cosmology is still expanding, that's why we choose Cosmology and applying machine learning to it.

1.3 Objective

Our main objective is to use boosting algorithm to find out the highest accuracy in search of a habitable extra solar planets. Our main goal is using different classifier to figure out which gives the highest accuracy rate and also their performance on these specific data. We also want to show that there are a lot of opportunities still waiting for us in the field of cosmology specifically using Machine Learning (ML). There are also some very intriguing topics like dark matter, string theory etc. where Machine Learning can be the next big step.

Chapter 2

Literature Review

2.1 Background Studies

Here we have given both the studies we have done for our cosmology part and also machine learning part.

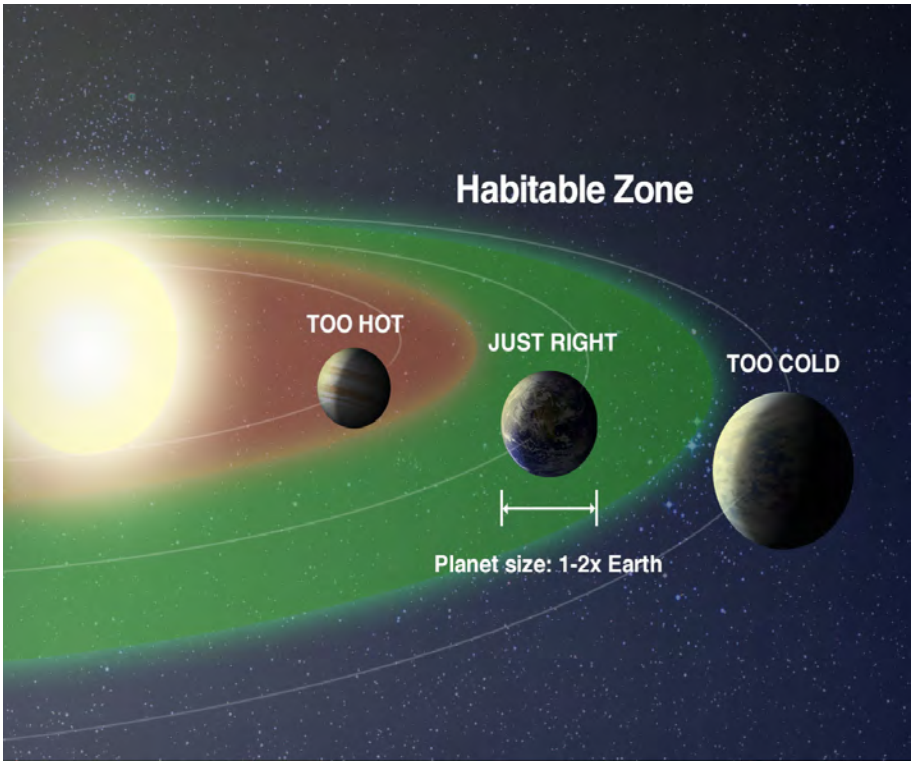
Searching Extrasolar Planets

In the past examinations specialist worked with different parameters first to discover distinctive exoplanets and their attributes. Photometric alignment and excellent arrangement is one of numerous procedures utilized by them. Photometric adjustment and outstanding order approach utilized by the Excellent Grouping Task to collect the Kepler Information Inventory (KIC). The KIC is a list obliging photometric and physical information for historical underpinnings in the Kepler mission field of view; it is utilized by the mission to choose most positive targets. Four of the noticeable light (g, r, I, z) sizes utilized in the KIC are fixing to Sloan Advanced Sky Study extents; the fifth (D51) is an Abdominal muscle size estimated to be steady with Castelli and Kurucz (CK) show air transitions [3]. We found that The Kepler Mission depends on express differential photometry to distinguish the 80 sections for every million (ppm) motion from an Earth-Sun comparable travel. Such constancy requires wonderful instrument security on timescales up to 2 days and efficient mistake evacuation to superior to 20 ppm. To this end, the shuttle and photometer experienced 67 days of approving, which incorporated a few informational collections taken to describe the photometer execution [4].

By then the most basic parameter is The Habitable Zone (HZ) of a star. The Habitable zone for a given star portrays the extent of circumstellar partitions from the star inside

which a planet could have liquid water on its surface, which depends on the remarkable properties. The points of confinement of the HZ for an explicit structure are resolved subject to both the exceptional properties and assumptions as for the response of a planetary air to great progress. Means HZ limits have encountered critical change. The begin that remotely unmistakable livability anticipates that water should exist on a planetary surface in a liquid state allows a quantitative depiction of the HZ to be surveyed. This is developed by uniting

Fig. 2.1 A Example of Habitable zone

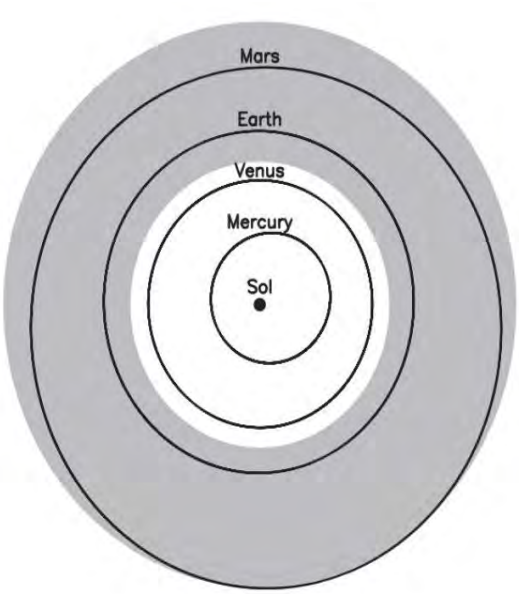


the superb properties with those of a hypothetical planet at various partitions from the star. The change motion gotten by the planet is a component of this detachment and the shine of the star. Kasting et al. (1993)[16] used one-dimensional (rise) air models to discover the climatic response of the earth to the progress by considering conditions whereby the concordance would impact to a runaway nursery affect or to a runaway snowball affect. The runaway nursery cycle is sketched out as seeks after: extended CO2 and also extended superb movement; temperature constructs; temperature rise prompts an extension in water vapor; extended ozone hurting substance obsession prompts high temperatures. Note that the extension in water vapor prompts fast loss of hydrogen and extended absorption at the wavelengths through which a considerable amount of Earth's infrared radiation escape, thusly spreading the runaway warming of the surface. The runaway snowball show is illustrated

as seeks after: reduced ozone exhausting substances and furthermore lessened brilliant movement; temperature lessens; occurring ice constructs surface reflectivity; diminished warming prompts further ice plan. Note this similarly results in enormous proportions of air CO2 uniting as dry ice and thusly further diminishing the nursery affect. The general tenability is furthermore perplexed by the distinctive segments trademark to the system. The change from the host star depends upon unearthly sort and luminosity class yet furthermore depends upon the age of the star. As the age of the star increases, so does the shine, and as such the HZ moves outward from the point of convergence of the system. One by then can consider a perpetually reasonable zone in which an explicit planetary circle may remain while the star is on the rule gathering, since the advancement of the HZ continues even after the star leaves the essential game plan [18]. As portrayed over, the planetary barometrical and albedo properties expect a basic occupation in controlling the glow re-arranged at the surface. The redistribution of that warm is in this way related to the planetary turn rate, particularly fundamental for planets in the HZ which may be tidally darted around late-type stars [24]. Furthermore, the planetary mass will effect such perspectives as the proportion of held atmosphere and the measurement and rationality of volcanic development. Too little a mass will most likely incite a nonappearance of these characteristics, and too high a mass will provoke a possibly reserved thick hydrogen envelope that will also cloud the detectable quality of bio marks. Along these lines, perfect planetary masses exist which result in a dynamically suitable modifying of these impacts.At that point the most basic parameter is The Habitable Zone (HZ) of a star. The livable zone for a given star delineates the extent of circumstellar divisions from the star inside which a planet could have liquid water on its surface, which depends on the remarkable properties. The points of confinement of the HZ for an explicit system are resolved subject to both the remarkable properties and assumptions as for the response of a planetary air to superb progress. Means HZ limits have encountered critical change. The begin that remotely conspicuous livability anticipates that water should exist on a planetary surface in a liquid state allows a quantitative depiction of the HZ to be evaluated.

Resulting to getting the data of the satellite various specialist made their own specific manner to manage make their own one of a kind record to choose the probability of life on those planets. Quantifiable examination of the hopefuls shows that at any rate half of stars have planets with orbital periods under 200d and that Earth-to Neptune-gauge planets are unquestionably more different than Jupiter-gauge planets [13] [8]. Some Kepler-perceived planets circumnavigating to a great degree cool (late-K-and M-type) small stars are close or inside the speculative 'livable zones' of these stars where an Earth-like planet could have liquid water on its surface [7] [11] [17] [21]. In any case, Kepler planet-encouraging stars are

Fig. 2.2 Distribution of habitable zone of our Solar System



usually far away (hundreds or thousands of pc) and pass out (V15), making estimation of mass by Doppler winding pace (RV, e.g. Marcy et al. 2014) or development, for instance, travel spectroscopy or impression of discretionary obscuration troublesome or incomprehensible.

The Kepler field covers just 0.25 percent of the sky and, by possibility, we know amazingly less about Earth-to Neptune-measure planets around close to stars, including those around to an extraordinary degree cool diminutive people. The RV framework can be immediately connected with close to stars which are usually dispersed on the sky, yet since of the scaling between planet mass and length, it is less unsteady to more minor planets than the development approach. The most touchy RV considers have discovered a few super-Earths on close circles around M smaller people [23], yet such examinations have been hampered by the faintness of such stars at distinguishable wavelengths. Ground-based, wide-field travel considers are affected by related ('red') commotion from the air and can just perceive brief period brute planets around F and G midgets. Travel surveys of adjacent enormously cool small stars utilizing individual pointing have met with restricted achievement [2] [12].

A machine learning approach

Disregarding the way that there were various papers on machine adjusting anyway executing machine making sense of how to find a valid planet is absolutely new. So we read a huge amount of paper on machine adjusting, especially on boosting. Boosting is a general

technique for improving the exactness of some arbitrary learning count. Focusing basically on the Boosting computation, we audits a part of the progressing take a shot at boosting including examinations of boosting's arrangement mix-up and theory botch; boosting's relationship with beguilement speculation and straight programming; the association among boosting and vital backslide; extensions of AdaBoost for multiclass portrayal issues; systems for solidifying human data into boosting; and preliminary and associated work using boosting. Building an extremely exact desire rule is verifiably a troublesome endeavor. On the other hand, it isn't hard at all to consider undesirable reliable rules that are simply nicely correct. Boosting, the machine-learning procedure that is the subject of this part, relies upon the recognition that finding numerous unforgiving general rules can be substantially less requesting than finding a singular, exceedingly correct desire rule. To apply the boosting approach, we start with a method or count for finding the cruel reliable rules. The boosting computation calls this "frail" or "base" learning figuring more than once, each time sustaining it a substitute subset of the planning models (or, to be dynamically correct, a substitute transport or weighting over the arrangement examples2). Each time it is called, the base learning estimation delivers another weak figure rule, and after various rounds, the boosting computation must solidify these feeble standards into a single desire choose that, in a perfect world, will be impressively more exact than anyone of the fragile principles. To make this approach work, there are two noteworthy request that must be answered: first, by what strategy should each scattering be singled out each round, and second, by what technique should the weak standards be joined into a lone standard? As for choice of scattering, the framework that we advocate is to put the most load on the models much of the time unclassified by the previous slight rules; this has the effect of compelling the base understudy to focus on the "hardest" points of reference. As for uniting the slight precepts, basically taking a (weighted) lion's offer vote of their figures is ordinary and effective[22].

We in like manner used controlled learning features of the Machine learning. The essential focal point of controlled learning is to fabricate a brief model of the transport of class in imprints the extent that marker features. A champion among the most tremendous use of machine learning is data mining. Various people much of the time submits mistakes in the midst of examinations or maybe, when endeavoring to make the most ideal relationship between various features. This makes it slippery responses for explicit issues. Controlled machine learning can as often as possible be viably associated with these issues, improving the capability of the structure and moreover better precision for the result.

Chapter 3

Finding Exo Planet and Their Characteristics

3.1 What is Exo Planets?

Since the late of the twentieth century in excess of 3,000 planets have been found outside our very own nearby planetary group, changing the way in which we think about the universe and its size, and opening up our minds to contemplations, for instance, the probability of life elsewhere.

The majority of the planets in our excellent framework circle around the Sun. Planets that circle around different stars are called exoplanets. Exoplanets are difficult to see specifically with telescopes. They are covered up by the splendid glare of the stars they circle.

These affirmed exoplanets, which circle their own one of a kind stars, are directly expected to exist in abundance all through the Smooth Way, and research is in advancement into how a part of these may have the ability to help life.

It is believed that various elements should be perfect for starting and keeping up the advancement of life on different planets. These components incorporate the planet's temperature, which ought to be neither too hot nor excessively cool (and ought not differ too drastically), and – essentially – the accessibility of fluid water. Be that as it may, our comprehension of these limitations is just in its early stages.

"With respect to, the ebb and flow – and extremely oversimplified – see is that the planet must have fluid water at its surface," says Prof Aline Vidotto, at the school of material science, Trinity School Dublin.

"This is predominantly dictated by the temperature of the planet, which is set by how much light it gets from the star which it circles," she says.[20]

A planet's temperature in this way depends generally on how far it is from its star, and the area around a star where simply enough radiation is gotten for fluid water to exist is known as the livable, or Goldilocks, zone.[15][14]

Most exoplanets have been found by the Kepler Space Telescope, an observatory that began work in 2009 and is required to finish its focal objective in 2018, when it misses the mark on fuel. As of mid-Walk 2018, Kepler has discovered 2,342 insisted exoplanets and revealed the nearness of possibly 2,245 others. The total number of planets found by all observatories is 3,706.

3.2 How are they detected?

Exoplanets orbiting different stars are excessively far away to be straightforwardly imaged from Earth with presently accessible telescopes. There are a couple of special cases of immense planets orbiting far from their host star and a portion of things to come telescopes as of now under development will be significantly more ready to picture exoplanets than should be possible today. Be that as it may, coordinate pictures of exoplanets are as of now special cases and, regardless, we won't have the capacity to see substantially more than a spot of light; nothing that could uncover any surface subtleties of the planet.

But here is the good news: we are able to detect these planets anyhow, and in some cases we will even be able to analyze the composition of their atmospheres! So how about we view the two fundamental strategies for recognizing exoplanets.

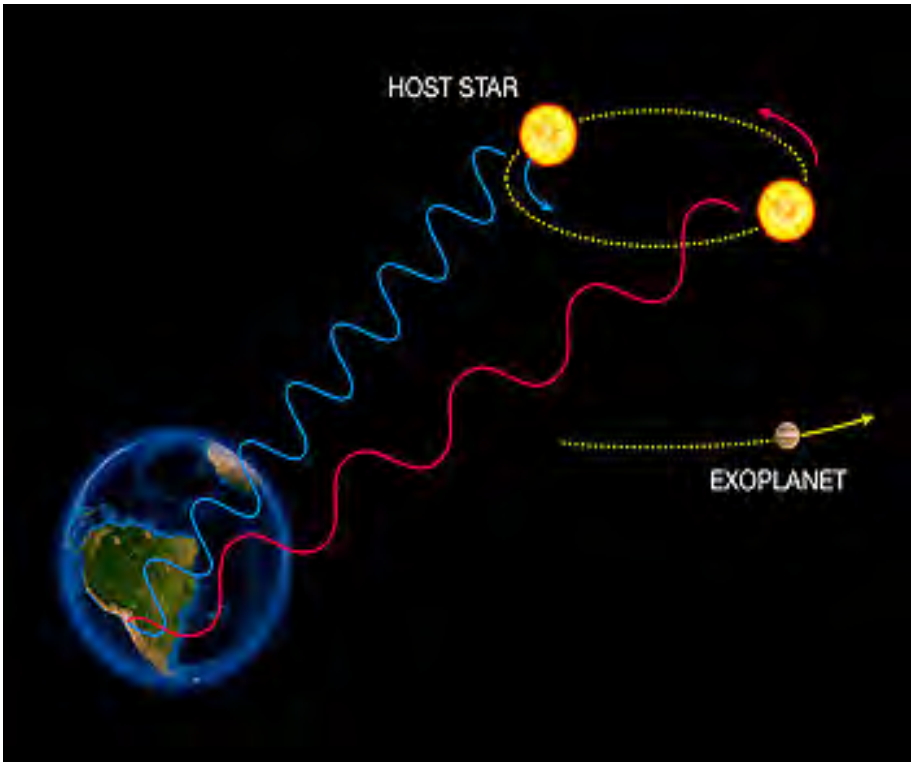
3.2.1 Radial Velocity (Doppler Method)

To be extremely exact: a planet does not circle around a star as we generally will in general say. Actually both the star and the planet circle around their normal focus of mass. Since the star is considerably more enormous than the planet this middle is near the star or even inside its volume - however it is never the focal point of the star itself. This implies a star with planets around it is never totally still; it is circling around the regular focus of mass of the whole outstanding framework. This wobbling of the star can be seen from Earth with help of the Doppler impact. All of you realize the Doppler impact since it causes the sound of a moving toward question like a vehicle or a train to sound sharp (the recurrence of the sound waves are higher) and that of a leaving item to show up lower (bring down sound recurrence). This impact works for sound waves as well as for electromagnetic radiation. Amid its circle around the focal point of mass, the star moves towards us and far from us, accordingly the discharged light of the star marginally changes its recurrence intermittently as observed from

Earth. The extent and recurrence of the wobbling focuses to the presence, and even the base mass, of the planets in circle around the star.

Several planets have been found utilizing this strategy. In any case, just enormous planets like Jupiter, or considerably bigger can be seen along these lines. Littler Earth-like planets are a lot harder to discover on the grounds that they make just little wobbles that are difficult to identify.

Fig. 3.1 Example of Searching exo-planets using Doppler Method



As a star moves towards an observer, the interim between back to back light waves decline, and the light is identified at a higher recurrence (or shorter wavelength). As it moves away the interim increment and the light is distinguished at a lower recurrence (or longer wavelength). The sum by which this progressions relies upon how rapidly the star moves, which thus relies upon the mass of the planet.

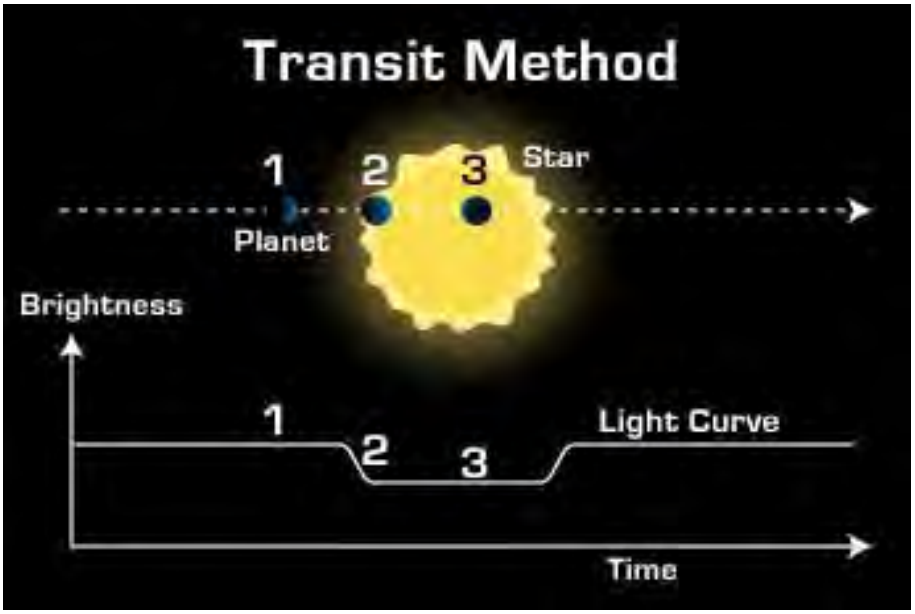
3.2.2 Transit Method

Another technique for identification is the transit method. At the point when a planet goes before a star, the planet will obstruct a modest measure of light. Telescopes can distinguish this slight darkening and can work out the sweep of the planet by how much light is blocked

and the circle by how as often as possible the planet goes before the star. One disadvantage is that a star needs the correct circle so it goes before the star.

With this strategy we can just locate a minor part of the current exoplanets since the Earth, the exoplanet and its star must be consummately adjusted so as to watch an exoplanets transit. For instance, the chances of recognizing the Earth from any arbitrary position outside our nearby planetary group with the transit strategy are simply 0.3 percent, the shot of finding Mars would be simply 0.18 percent. By the by, space observatory Kepler has constantly estimated the light bends of 145 000 fundamental succession stars over a time of quite a long while, discovering many exoplanets utilizing the transit method.

Fig. 3.2 Example of Searching Exo-planets using Transit Method



With the assistance of the transit method we can not just distinguish at least one planets in an outstanding framework, we can likewise increase other significant data:

• The size of the planet can be assessed by the measure of darkening. Greater planets cause a more grounded drop of the light bend than do littler planets.

• With the correct instruments we can even break down the climate of the planet. How is this conceivable? A piece of the light from the star goes through the air of the planet. Diverse particles in the environment square explicit wavelengths of the star's range. By estimating the purported transmission range here on Earth we can decide the creation of the planet's environment. With the new telescopes that will work inside the following decade we may even discover supposed bio-markers, certain particles that could be an indication of additional earthly life!

3.3 Characteristics of Exo-Planets

• The known exoplanets fall along a scope of sizes, masses, and orbital positions. Sizes and masses run from littler and less monstrous than Earth to super-Jupiter kinds of universes. Orbital positions go from near the parent star to exceptionally far off.

• Astronomers are beginning to discover and gauge environments around removed exoplanets. This enables them to comprehend what gases exist in those vaporous envelopes.

• Among different attributes, stargazers can quantify the surface temperatures, circles, attractive fields, and shades of exoplanets. As discovery strategies enhance, they will have the capacity to discover progressively about removed universes.

• The locale around a star where fluid water could exist on the surface of a strong planet is known as the livable zone. Universes circling in that zone are viewed as prime applicants where life could be upheld.

• More than 22 percent of Sun-like stars have Earth-sized planets in their livable zones. These are critical spots to focus a look for conceivable life-bearing universes.

The Kepler Mission was jump started to look out removed universes. It proceeds with its inquiry today. Different missions that have discovered removed universes incorporate the Hubble Space Telescope, the CoROT mission from the European Space Office, the Savvy mission, and the Herschel rocket. Ground-based observatories keep on being a vital piece of the scan for far off universes.

Chapter 4

Data Analysis

4.1 Description of Dataset

We use NASA's extra solar planets dataset to predicting the habitable planets existence. For our research purpose we have built our data set according to our goal. We had searched for datasets in some popular data repositories like Planck, IPAC. But those datasets were confidential and there were some ethical obligations to publish the data in open sources for public access. So after searching every nook and corner we failed to get any secondary data. But still there was so many obligations and confidentiality that we need to go through. Data collection became the most challenging part of our research. After several discussions between our teammates and the supervisor, we finally decided to take help one of our elder brother. He gave us the link of NASA's Extra Solar Planets Dataset. We studied over 300 published papers of NASA's about these planets different features. It took a long time to get all the parameters we needed for our dataset. But when we combined the whole dataset there were some missing values under some important features and same planets had different values in different papers. There are several parameters in our data set such as

- Planet: Name of planet,
- Mp: Planet mass (Jupiter masses),
- Rp: Planet radius (Jupiter radii),
- P: Orbital period,
- e: Orbital eccentricity,
- w: Longitude of periastron,
- tHZc: Percentage of the orbital phase spent within the Conservative Habitable Zone,
- tHZo: Percentage of the orbital phase spent within the Optimistic Habitable Zone.There

are some planetary equilibrium temperatures assumptions also such as

- Teqa:Periastron,hot-dayside,

• Teqb:Periastron,well-mixed,

• Teqc:Apastron,hot-dayside,

• Teqd:Apastron, well-mixed.

The "hot-dayside" display expect there is no warmth redistribution by the environment thus the planetary temperature is determined dependent on the warm vitality being confined to the half of the globe confronting the star. The "very much blended" display then again expect the warmth is redistributed equitably over the whole surface by fast zonal breezes. Presently we should talk about HZ (Tenable Zone) Parameters :

Mp: Planetary mass is a proportion of the mass of a planet-like question. Inside the Close planetary system, planets are typically estimated in the cosmic arrangement of units, where the unit of mass is the sunlight based mass (M), the mass of the Sun. In the investigation of extrasolar planets, the unit of measure is regularly the mass of Jupiter (MJ) for expansive gas monster planets, and the mass of Earth (M) for littler rough earthbound planets.

Rp: Radius is the planetary range. It is ordinarily determined from photometry information of traveling planets as an element of excellent span, which can be determined from ghostly vitality conveyance fitting to multiband outstanding photometry. For planets with just outspread speed information, the span is either obscure, or it must be assessed from the planetary least mass. For instance: one Jupiter span is roughly 11 Earth radii.

P: The orbital period is the time a given galactic protest takes to finish one circle around another question, and applies in space science more often than not to planets or space rocks circling the Sun, moons, circling planets, exoplanets circling different stars, or twofold stars.

e: The Observed Eccentricity is the orbital flightiness. This amount is commonly surmised from Keplerian orbital fits to spiral speed information, coordinate imaging, or smaller scale lensing. Erraticism is a major parameter portraying the condition of a circle (the orbital way for a bound planet, with the host star at one focal point of the oval). Whimsy for bound circles falls somewhere in the range of 0 and 1. A capriciousness of 0 relates to a roundabout circle. An unpredictability of 0.7 relates to an exceedingly erratic planet circle.

where **a** is the semi-major axis, **r** is the radius vector, is the genuine inconsistency (estimated anticlockwise) and **e** is the eccentricity. An ellipse has an unconventionality in the range $0 < e < 1$, while a circle is the extraordinary case e=0.

w:The Longitude of Periastron or Contention of periastron is the orbital edge at which a planet experiences its periastron (time of nearest approach). This amount is a compound point, and it characterizes the circle of the planet as for the plane of the sky. This amount is normally induced from Keplerian orbital fits to spiral speed information. Longitude of Periastron characterizes the introduction of the circle as for the plane of the sky. The Longitude of Periastron is specifically identified with the Season of Periastron Section.

Subsequent to getting the informational index we chose to break down the information well ordered. First we center around our information structure and our information type. Picking a calculation is a basic advance in the machine learning process, so it's an essential assignment to recognize the information structure first. The can be two conceivable sorts of information types, for example, Organized Information and Unstructured Information. Presently we should talk about the organized which has been sorted out into a designed vault, normally in a database or dataset so its components can be made addressable for progressively powerful preparing and examination. Then again unstructured information type is a model of not sorted out in a pre-characterized way. Unstructured data's are commonly message substantial, yet may contain information, for example, dates, numbers, and realities too. Our parameters and also the entire informational collection are efficient for executing any calculation so our informational collection contains organized information.

We actualized regulated learning process where the information needs to prepare and test to foresee an expected outcome. Directed learning is a kind of machine discovering that empowers the model to foresee future results after they are prepared dependent on past datahere the client need to give an objective variable for prediction.On the other hand both the info variable and the objective variable requirement for each preparation information which is known as labeleddata. At first preparing information incorporates a lot of models with combined information subjects and wanted yield. This learning calculation at that point takes in the connection between the info and the yield factors from the information where the given info X can deliver the relating yield y.

4.1.1 Training Dataset

The training data set in Machine Learning is the actual dataset used to train the model for performing various actions. This actual data develop process models to learn with the help of algorithms to train the machine to work automatically. We used 2860.8 data for training from the whole dataset, which is 80 percent.

4.1.2 Validation of Dataset

While tuning model hyper parameters, the sample data used to provide an unbiased evaluation of a model fit on the training dataset. Among the 80 percent of training data, we used 70 percent of them for creating the model and remaining 30 percent is used to fit the model. Once the model is well -fitted then the model is considered as trained data.

We split the train data into 70 percent and 30 percent for creating the unbiased model.

After creating the model neither for emotion detection we tested some data, which was, nor in the training part.

4.1.3 Test Dataset

To evaluate the model the test dataset provides us the standard. When our model is completely trained, the test dataset is used. We got different type of accuracy from different type of algorithms .we used 400 data (which is 20 percent data in whole data set) for testing. This was not in training part.

Chapter 5

Applying Machine Learning

5.1 Choosing best machine learning algorithms:

5.1.1 Classification

Classification and Regression are two different types of algorithms that we can apply to our data set depending upon the nature of the input data and output prediction that we make. If we want to say it in a descriptive way, we need to focus on how classification and regression works .Classification is a machine learning technique of identifying to which of a lot of classifications another perception has a place, on the basis of a training set of data containing observations whose category membership is known. Classification predictive models do the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y). The output variables are often called labels or categories. The mapping capacity predicts the class or classification for a given perception. As for an example we have a condition like Astronomers have been cataloguing distinct objects in the sky using long-expose CCD images. The objects need to be labeled as star, galaxy, nebula etc. The data is highly noisy and the images are very faint. The cataloguing can take decades to complete. The Physicists need to automate the cataloguing process and improve its effectiveness. Here the classification method can help then to improve their effectiveness and also for cataloguing process. If the Physicists select their feature section and analyses their data set, they can use then various classification model and decision tress to classify their objects effectiveness and will get an estimated accuracy to catalogue their objects. The general equation for linear regression is:

$$F = f(W.X) = f\left\{ Sum \sum_{n=1}^{\infty} (W_n X_n) \right\} \tag{5.1}$$

5.1.2 Regression

On the other hand Regression in machine learning is a technique from statistics that is used to predict values of a desired target quantity when the target quantity is continuous. And regression predictive models do the task of approximating a mapping function (f) from input variables (X) to a continuous output variable (y) and the continuous output variable is a genuine esteem, for example, a whole number or drifting point esteem. Similarly these are often quantities, such as amounts and sizes. As for an example a person has decided to start a hot dog business. He has hired his cousin to help him with hot dog sales. There is a problem occurred that is his cousin can only work 20 hours a week. But he wants to have his cousin working at peak hot dog sales hours. In order to find the information he needs to collect the data and use regression analysis to find the optimum hot dog sale time.Regression analysis is the investigation of two factors trying to discover a relationship, or connection. So he needs to apply a reasonable relapse examinations model to discover the relationship and gauge the ideal deal time in this procedure, if the dependent variable is continuous, the independent variable(s) is continuous or discrete and nature of regression line is linear, linear regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line which is known as regression line then the represented equation will be:

$$Y = a + b * X + e \tag{5.2}$$

Where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable.

5.2 Why Choose Classification?

Classification is regarding predicting a label and regression is regarding predicting a amount. Classification algorithms are used once the required output may be a separate label. In different words, they're useful once the solution to your question regarding your assigned drawback underneath a finite set of doable outcomes . If these are the queries you're hoping to answer with machine learning in your assigned drawback, think about algorithms like naive Bayes, call trees, provision regression, kernel approximation, and K-nearest neighbors. On the opposite hand, regression is helpful predicting a continual amount. meaning the solution to your question is painted by a amount that may be flexibly determined supported the inputs of the model instead of being confined to a group of doable labels. A classification algorithmic rule might predict a continual price, however the continual price is

within the sort of a likelihood for a category label .But A regression algorithmic rule might predict a separate price, however the separate price within the sort of AN number amount. Classification predictions is evaluated victimization accuracy, whereas regression predictions cannot. Regression predictions is evaluated victimization root mean square error, whereas classification predictions cannot. that the key to settle on is to work out what sort of output we tend to primarily would like. 1st if we wish to settle on whether or not we must always use Classification or Regression, we might have to examine or information wherever X is input information|input file|computer file and Y is output data .Regression algorithms are used for continuous information wherever X and Y each are continuous variables. On the opposite hand Classification algorithms are useful once we have categorical information that will|we will|we are able to divide in 2 or additional classes as any drawback with solely 2 doable output 'Yes' or 'No' wherever Y can have solely 2 values zero for each No and one for each affirmative .Our information set may be a categorized and that we assume the target (parameter) price as "0" and "1".That's why selected classification for obtaining out expected results of livable planet existence. In our information set we'd like to specialize in that planet is additional livable then others. therefore we tend to labelled the characteristic "HZ"(average temperature) and "Period"(each year time) as "1" so as to assume that there can be any chance to measure and haven't any "HZ"(average temperature) and "Period"(each year time) as "0" so as to assume the living isn't doable there.

We assumed one single classification drawback isn't enough to predict the correct results of livable planet existence therefore we tend to used 5 classifier models build|to form|to create individual accuracy and make mix estimation by victimization sensory and sensitivity formula .We conjointly used cross validation and learning curve to form the accuracy more economical.

5.2.1 Ada Boost Classifier

Ada-boost classifier train weak classifier algorithms and transform them into strong classifier. A single algorithm is not enough for classifying the objects properly so we need to combine many classifiers all to gather to avoid the poor result. So we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good accuracy score for overall classifier .It retrains the algorithm attractively by choosing the training set based on accuracy of previous training. The weight-age of each trained classifier at any iteration depends on the accuracy achieve[19]. Ada Boost works on improving the areas where the base learner fails. The base learner is a machine learning algorithm which is a weak learner and upon which the boosting method is applied to turn it into a strong learner. Where $f_m stands for the m_{th} weak classifier and the ta_m is the corresponding weight. It is an exactl$

$$F(x) = sign\left\{\sum_{n=1}^{M}(\theta_n f_n)\right\} \tag{5.3}$$

5.2.2 KNeighbor's Classifier (KNN)

KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms. KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. When we say a technique is non-parametric, it means that it does not make any assumptions on the underlying data distribution. In other words, the model structure is determined from the data. If you think about it, it's pretty useful, because in the "real world", most of the data does not obey the typical theoretical assumptions made (as in linear regression models, for example). Therefore, KNN could and probably should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution data. KNN can be used for classification—the output is a class membership (predicts a class—a discrete value). An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. It can also be used for regression—output is the value for the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbors.

Pros:

• No assumptions about data; useful, for example, for nonlinear data,

• Simple algorithm, to explain and understand/interpret,

• High accuracy (relatively), it is pretty high but not competitive in comparison to better supervised learning models,

• Versatile, useful for classification or regression.

Cons:

• Computationally expensive;because the algorithm stores all of the training data

• High memory requirement,

• Stores all (or almost all) of the training data,

• Prediction stage might be slow (with big N),

• Sensitive to irrelevant features and the scale of the data.

5.2.3 Decision Tree Classifier

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, decision tree algorithm can be used for solving regression and classification problems too. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data(training data).

The understanding level of Decision Trees algorithm is so easy compared with other classification algorithms. The decision tree algorithm tries to solve the problem, by using tree representation. Each internal node of the tree corresponds to an attribute, and each leaf node corresponds to a class label.

Decision Tree Algorithm Pseudo-code

1. Place the best attribute of the data-set at the root of the tree. 2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute. 3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

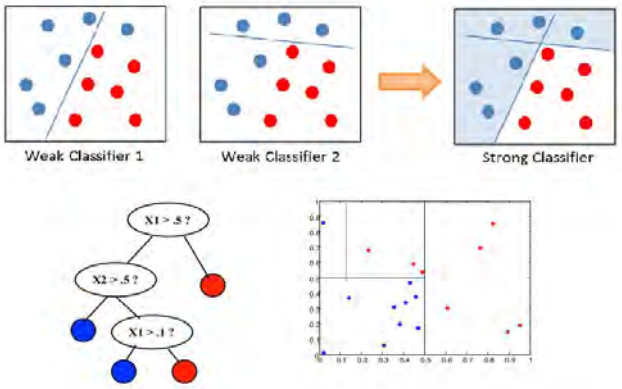
In decision trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. We continue comparing our record's attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value. Now let's understanding how we can create the decision tree model.

5.2.4 Boosting Algorithm

"Boosting" is a general method for improving the performance of any learning algorithm. In theory, boosting can be used to significantly reduce the error of any "weak" learning algorithm that consistently generates classifiers which need only be a little bit better than random guessing.Despite the potential benefits of boosting promised by the theoretical results, the true practical value of boosting can only be assessed by testing the method on "real" learning problems [9]. We implemented boosting algorithm to verify our accuracy more efficiently. But before implementing the algorithm first we need to focus on which boosting algorithm would give the best prediction. Boosting is actually family of algorithms or methods of converting a set of weak learners into strong learners. The idea of boosting is to train weak learners sequentially, each trying to correct its predecessor. Suppose we have a

binary classification task. As for an example a weak learner has an error rate that is slightly lesser than 0.5 in classifying the object, the weak learner is slightly better than deciding from a coin toss. A strong learner has an error rate closer to 0. To convert a weak learner into strong learner, we take a family of weak learners, combine them and vote. This turns this family of weak learners into strong learners. We can also explain this with a figure:

Fig. 5.1 Example of Boosting Model



Let Blue and Red be different classifiers. Their area of Blue represents where the classifier blue miss classifies (goes wrong) and area of Red represents where the classifier Since there is no correlation between the errors of each classifier, combining them and using a technique of democratic voting to classify each object, this family of classifiers will never go wrong. This would have provided a basic understanding of boosting. Actually here the wrong decisions will pass the wrong answer to train and test for each model. And the end the combine estimated result will be the prediction.The different types of boosting algorithms are Ada Boosting, Gradient Boosting and XG Boosting.

5.2.5 Ada Boosting Algorithm

Ada Boosting is basically a variation of bagging.Methods for solving classification algorithms, such as Bagging and Ada Boost, have been shown to be very successful in improving the accuracy of certain classifiers for real-world data-sets. Bagging reduced variance of unstable methods, while boosting methods reduced both the bias and variance of unstable methods [1].The pseudo of Ada Boosting algorithm is [10]:

Input: sequence of m examples h(x_1, y_1), ...(x_m, y_m)

with labels $y_i \in Y = 1, ..., k$

weak learning algorithm Weak Learn

integer T specifying number of iterations

Initialize $D_1(i) = 1/m for all i.$

Do for t = 1,2,.....T

1. Call Weak Learn, providing it with the distribution D_t.

2. Get back a hypothesis $h_t : X \to Y$.

3. Calculate the error of $h_t : \in_t = \sum_{i:h_t(x_i) \neq y_i} (D_t(i)). If \in_t > 1/2,$

then set T= t-1 and abort loop.

4. Set $\beta_t = \in_t /(1- \in_t)$

5. Update distribution $D_t : D_{t+1}(i) = D_t(i)/Z_t * (\beta if h_t = y_i$

or 1 otherwise) where Z_t :

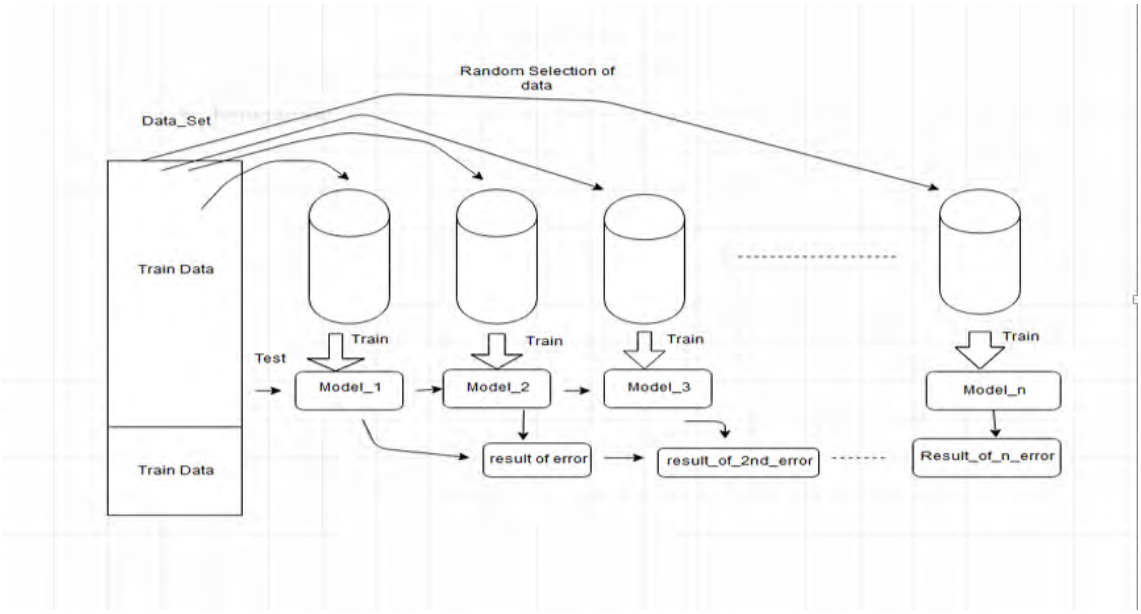
is a normalization constant (chosen so that (D_{t+1})

will be a distribution).

Output :the final hypothesis: $h_f in(x) = \arg max_(y \in Y) \sum_{t=1}^{T} (log 1/\beta_t)(h_t(x,y))$

Here is the picture which describes the implantation process in brief where the training data set makes model of weak learners and combine them to calculate error sequentially .The figure is below:

Fig. 5.2 Figure of Ada Boosting technique



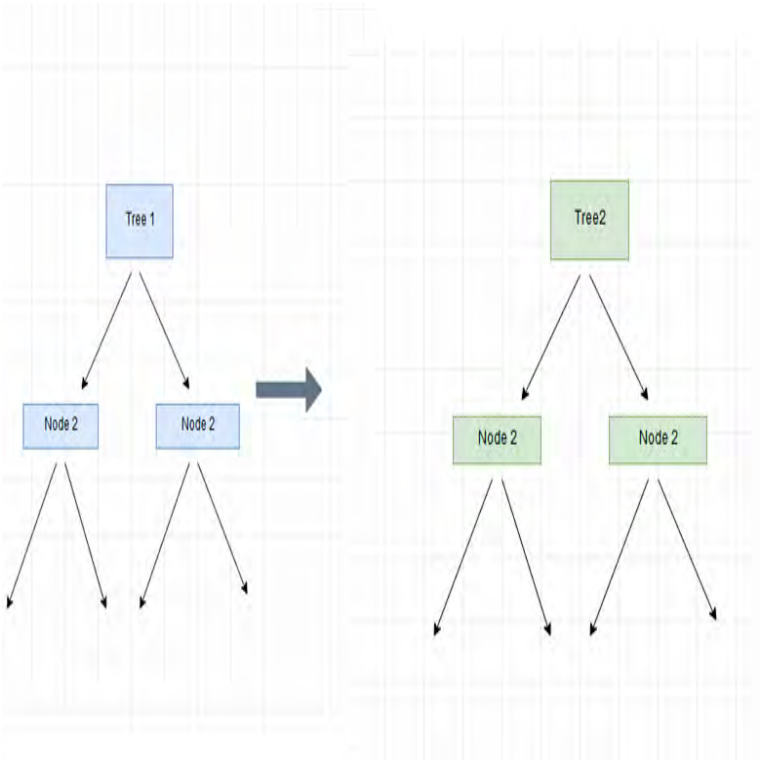
We implemented Ada Boosting algorithm but the accuracy result was always near to 26.98 % which was not sufficient for our possible outcome .In order to get better result we also implemented XG boot algorithm

5.2.6 XGBoost(Extreme Gradient Boosting)

It is one among-st the quickest implementations of gradient boosted trees. It will try that is one among-st the main inefficiencies of gradient boosted trees considering on the potential loss for all doable splits to make a replacement branch. XG Boost tackles this un skillfulness by staring at the distribution of options across all information points in a very leaf and victimization this data to cut back the search area of doable feature splits.It has several features [6]: 1. Speed: xgboost can automatically do parallel computation on Windows and Linux. It is generally over 10 times faster than gbm. 2. Input Type: xgboost can take several types of input data 3. Sparsity: xgboost accepts sparse input for both tree booster and linear booster, and is optimized for sparse input. 4. Customization: xgboost supports customized objective function and evaluation function 5. Performance: xgboost has better performance on several different

Although XG Boost implements a few regularization tricks, this speed up is by far the most useful feature of the library, allowing many hyper parameter settings to be investigated quickly [5]. This is helpful because there are many, many hyper parameters to tune. Nearly all of them are designed to limit over fitting. There are two reasons why we choose XG Boost which are Execution speed and Model performance.The most helpful part in this iterative process is that it passes the error to its second tree which has the same weight of the first tree.The tree figure of the tree is given below :

Fig. 5.3 Weight transfer one Extreme Gradient Tree to another



Chapter 6

Results

6.1 Performance Analysis

We used three classifiers for predicting results depending on each how many days each year can take and average temperature .

Fig. 6.1 Comparison of Different Classifier of HZ

```
In [147]: models = []
          models.append(("AdaBoostClassifier:",AdaBoostClassifier()))
          models.append(("Decision Tree:",DecisionTreeClassifier()))
          models.append(("K-Nearest Neighbour:",KNeighborsClassifier(n_neighbors=3)))

          print(models)

          [('AdaBoostClassifier:', AdaBoostClassifier(algorithm='SAMME.R', base_estimato
                    learning_rate=1.0, n_estimators=50, random_state=None)), ('Decision
          one, criterion='gini', max_depth=None,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                      splitter='best')), ('K-Nearest Neighbour:', KNeighborsClassifier(a
          ski',
                     metric_params=None, n_jobs=1, n_neighbors=3, p=2,
                     weights='uniform'))]

In [148]: from sklearn.model_selection import KFold
          from sklearn.model_selection import cross_val_score
          results = []
          names = []
          for name,model in models:
              kfold = KFold(n_splits=10, random_state=0)
              cv_result = cross_val_score(model,X_train,y_train.values.ravel(), cv = kfo
              names.append(name)
              results.append(cv_result)
          for i in range(len(names)):
              print(names[i],results[i].mean()*100)

          AdaBoostClassifier: 100.0
          Decision Tree: 99.95833333333334
          K-Nearest Neighbour: 99.24843096234311
```

This is the accuracy result depending of the feature "HZ" which is the planets average temperature where the adaptive classifier shows 100 percent accuracy, Decision tree is about

99.9 percent just close to adaptive and K-NN is 99.24 percent .So we can say that there is at least one habitable planet existing in the data set.

This is the accuracy result depending of the feature "period" which is the day each year takes where the adaptive classifier shows 100 percent accuracy, Decision tree is about 100 percent and K-NN is 97.78 percent .Here majority is saying the possibility is 100 percent. So the day is not far away when we will be able to see another potential planet where life can be formed.

Fig. 6.2 Comparison of Periods classifier

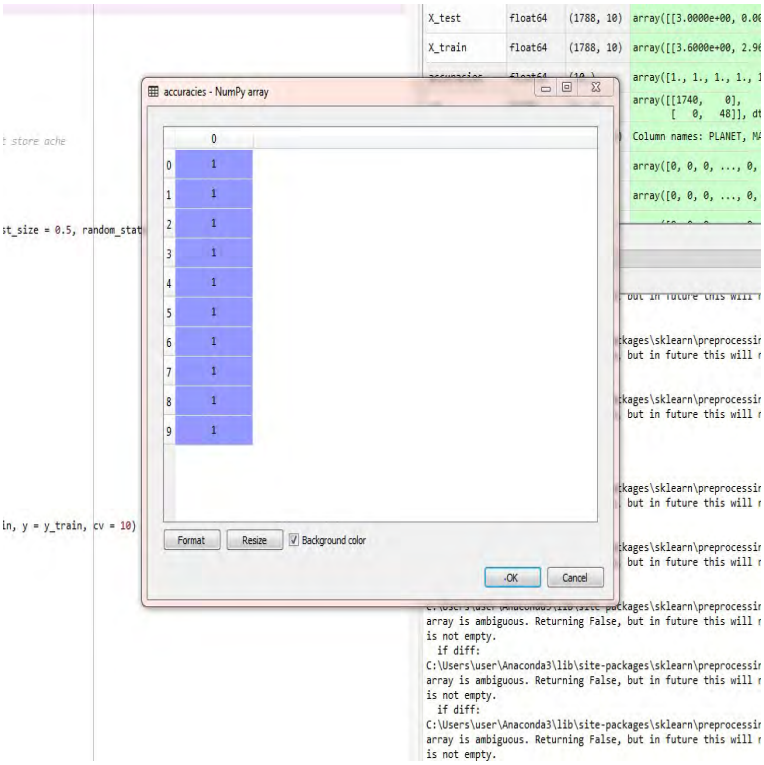
```
In [172]: models2 = []
          models2.append(("AdaBoostClassifier:",AdaBoostClassifier()))
          models2.append(("Decision Tree:",DecisionTreeClassifier()))
          models2.append(("K-Nearest Neighbour:",KNeighborsClassifier(n_neighbors=3)))

          print(models2)

          [('AdaBoostClassifier:', AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None,
                    learning_rate=1.0, n_estimators=50, random_state=None)), ('Decision Tree:', DecisionTreeClassifier(class_weight=N
          one, criterion='gini', max_depth=None,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                      splitter='best')), ('K-Nearest Neighbour:', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkow
          ski',
                     metric_params=None, n_jobs=1, n_neighbors=3, p=2,
                     weights='uniform'))]

In [171]: from sklearn.model_selection import KFold
          from sklearn.model_selection import cross_val_score
          results = []
          names = []
          for name,model in models2:
              kfold = KFold(n_splits=10, random_state=0)
              cv_result = cross_val_score(model,X_train,y_train.values.ravel(), cv = kfold,scoring = "accuracy")
              names.append(name)
              results.append(cv_result)
          for i in range(len(names)):
              print(names[i],results[i].mean()*100)

          AdaBoostClassifier: 100.0
          Decision Tree: 100.0
          K-Nearest Neighbour: 97.78748256624826
```

6.2 Results

We used five factors to predict a possible result the are HZ,period,mass and radius ,metallicity and less Eccentric Orbit.Among them some of them are easy to justify the result but some of are not. In our data set we don't have any dummy variable that's why we do not need to use the label encoder or one hot encoder. For avoiding the dummy variable track we omitted the one and only string in our data set which is Planets Name at the very first of our coding part. Here 20percent of our data is remaining in our test part and 80 percent data will go for train part. We fit the XG boost model and make a pre prediction with the of training and test

models. With the help k-Fold cross validation we can train our predicted variable also .Our accuracy result shows the possibility of having habitable planet is nine. The following figure is containing the code result where we used "Hz" (average temperature of habitable zone) parameter as our target variable:

Fig. 6.3 Planets from HZ characteristic



We also used confusion matrix which proves how the confusion has been done properly by the help of a cm variable. From the cm variables we can measure the accuracy performance of confusion matrix which is: (1683+0)/(1683+0+0+102) =.9428

The performance result is 94.28 percent so that we can consider it a perfect the performance result is fear enough .The following figure is containing cm variables which we got from our accuracy performance:

There is another figure where we use "Period"(per year total day) parameter as our target variable: Here the period variable accuracy result shows the possibility of having habitable

Fig. 6.4 Cm variables data table of HZ

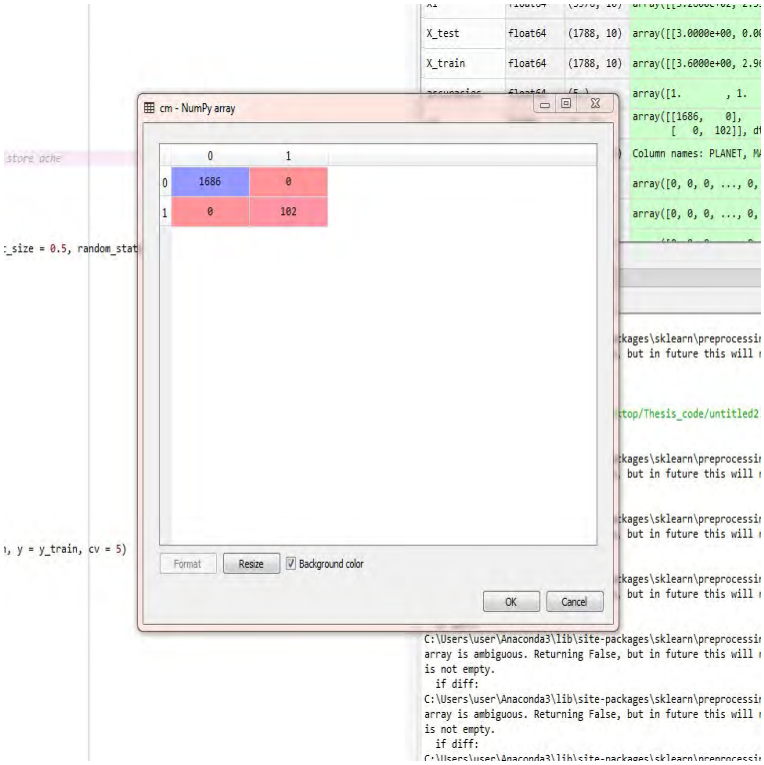
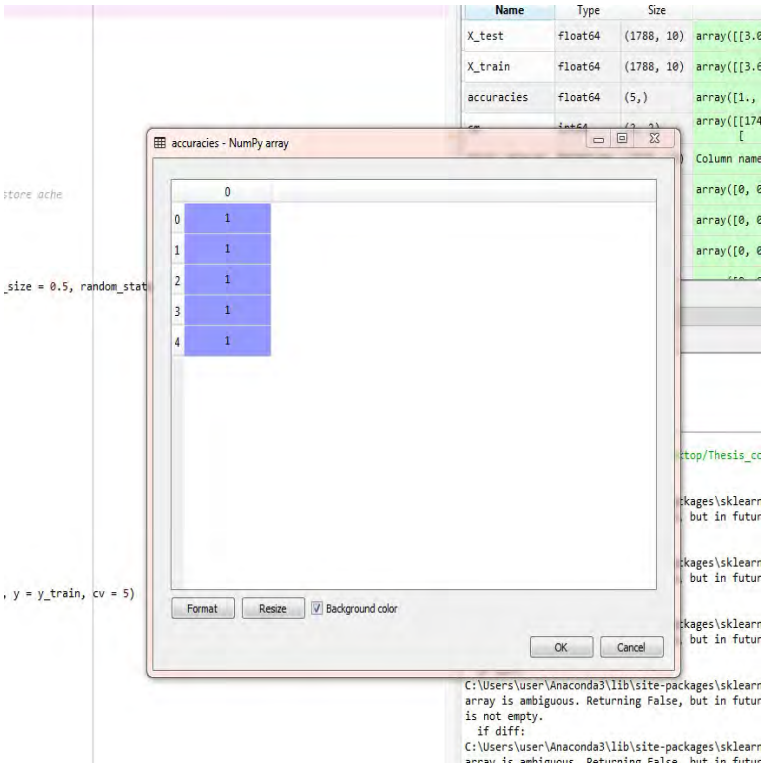


Fig. 6.5 Planets list based on Period parameter



planet is four. By seeing this estimation we can say that all of the planets have less days in a year than the Earth have days in one year .But we got 5 more planets when we consider "HZ"(average temperature) as our target variable. So we can say according to temperature this habitable planets have more efficacy. The accuracy performance of confusion matrix is: (1740+0)/(1740+0+48+0) =.9731

The performance result is 97.31percent which is more than the result we got from "HZ" target variable. So there be a prediction like the habitable planets must have less days then earth in a year. The following figure is containing cm variables which we got from our accuracy performance:

Fig. 6.6 Cm variables Data table(Period)

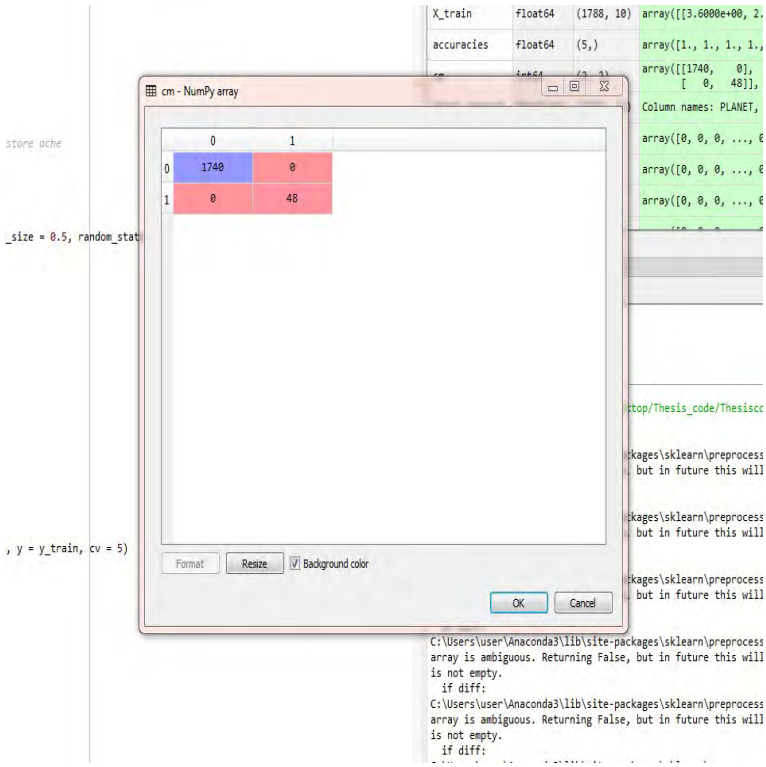


Fig. 6.7 Approximate Planet list with Mass and Radius value

Fig. 6.8 Approximate Planet list with Metallicity value

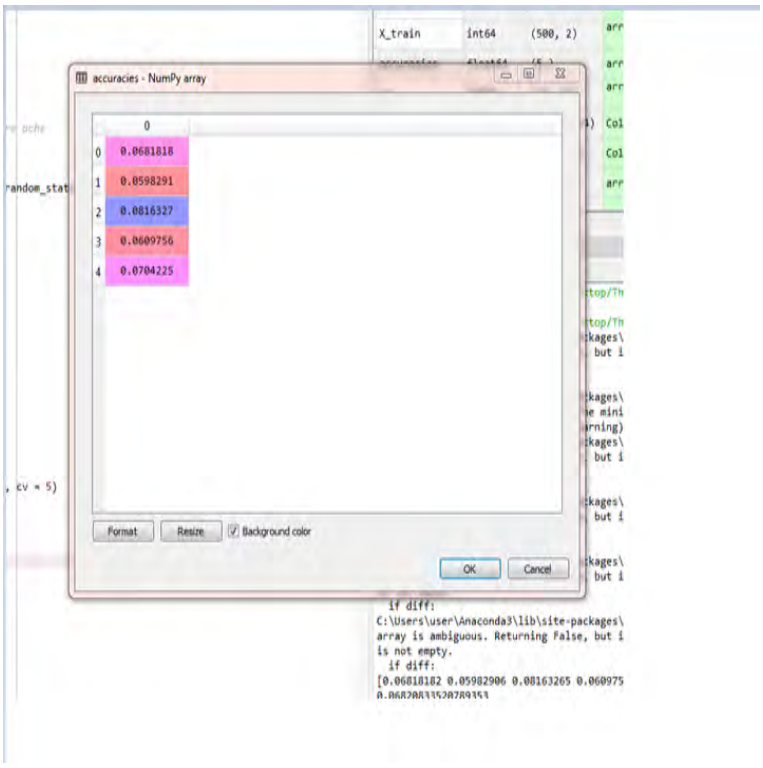
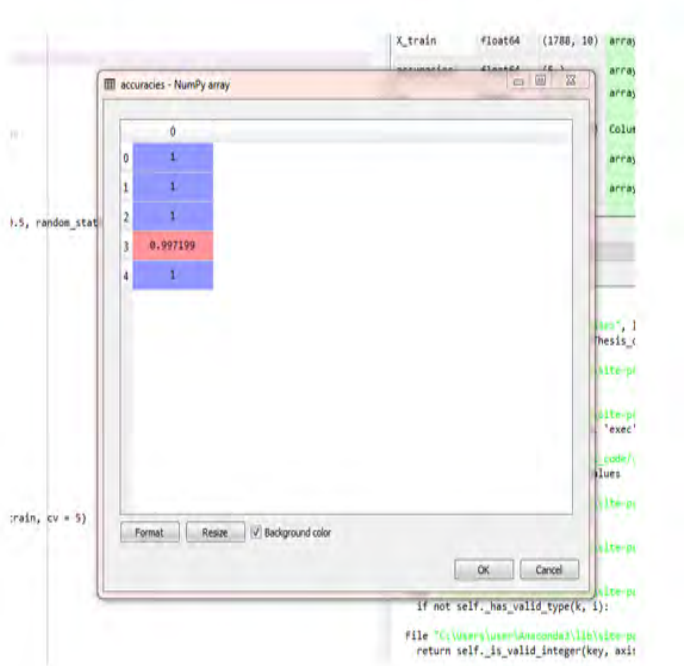


Fig. 6.9 (Approximate Planet list with Less Eccentric Orbit value)



Chapter 7

Conclusion

7.1 Future Plan

This works introduce us with the huge opportunities that are awaiting for Machine Learning in the fields of Cosmology/Astrophysics. And for this topic, many satellite of different international space organizations are working relentlessly to collect more data and discover new exo planets. As this field is expanding and also the data set of these exo planets. Our Future goal is to set a unsupervised program to collect data continuously and analyzing this data to until that day when we will really need to start our journey towards the infinity, and also so that at least we will have a destination and a chance of survives if we can reach that Planet

7.2 Summery of work

As we got different possible planets for different features, we are planing to do more research with our target possible habitable planet. We also want work in depth with this possible planets. Likewise we also want to implement machine learning algorithms in other cosmology files like like string theory, dark matter etc.

References

[1] Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36(1-2):105–139.

[2] Berta, Z. K., Irwin, J., and Charbonneau, D. (2013). Constraints on planet occurrence around nearby mid-to-late m dwarfs from the mearth project. *The Astrophysical Journal*, 775(2):91.

[3] Brown, T. M., Latham, D. W., Everett, M. E., and Esquerdo, G. A. (2011). Kepler input catalog: photometric calibration and stellar classification. *The Astronomical Journal*, 142(4):112.

[4] Caldwell, D. A., Kolodziejczak, J. J., Van Cleve, J. E., Jenkins, J. M., Gazis, P. R., Argabright, V. S., Bachtell, E. E., Dunham, E. W., Geary, J. C., Gilliland, R. L., et al. (2010). Instrument performance in kepler's first months. *The Astrophysical Journal Letters*, 713(2):L92.

[5] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM @articleruggieri2002efficient, title=Efficient C4. 5 [classification algorithm], author=Ruggieri, Salvatore, journal=IEEE transactions on knowledge and data engineering, volume=14, number=2, pages=438–444, year=2002, publisher=IEEE @articlepatil2013performance, title=Performance analysis of Naive Bayes and J48 classification algorithm for data classification, author=Patil, Tina R and Sherekar, SS, journal=International journal of computer science and applications, volume=6, number=2, pages=256–261, year=2013 @articlechoi2011validation, title=Validation of accelerometer wear and nonwear time classification algorithm, author=Choi, Leena and Liu, Zhouwen and Matthews, Charles E and Buchowski, Maciej S, journal=Medicine and science in sports and exercise, volume=43, number=2, pages=357, year=2011, publisher=NIH Public Access @inproceedingschen2016xgboost, title=Xgboost: A scalable tree boosting system, author=Chen, Tianqi and Guestrin, Carlos, booktitle=Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages=785–794, year=2016, organization=ACM @articlechen2015xgboost, title=Xgboost: extreme gradient boosting, author=Chen, Tianqi and He, Tong and Benesty, Michael and others, journal=R package version 0.4-2, pages=1–4, year=2015 @articlefriedman2002stochastic, title=Stochastic gradient boosting, author=Friedman, Jerome H, journal=Computational Statistics & Data Analysis, volume=38, number=4, pages=367–378, year=2002, publisher=Elsevier @articlemaclin1997empirical, title=An empirical evaluation of bagging and boosting, author=Maclin, Richard and Opitz, David, journal=AAAI/IAAI, volume=1997, pages=546–551, year=1997 @inproceedingsfreund1996experiments, title=Experiments with a new boosting algorithm, author=Freund,

Yoav and Schapire, Robert E and others, booktitle=Icml, volume=96, pages=148–156, year=1996, organization=Citeseer .

[6] Chen, T., He, T., Benesty, M., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4.

[7] Dressing, C. D. and Charbonneau, D. (2013). The occurrence rate of small planets around small stars. *The Astrophysical Journal*, 767(1):95.

[8] Fressin, F., Torres, G., Charbonneau, D., Bryson, S. T., Christiansen, J., Dressing, C. D., Jenkins, J. M., Walkowicz, L. M., and Batalha, N. M. (2013). The false positive rate of kepler and the occurrence of planets. *The Astrophysical Journal*, 766(2):81.

[9] Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969.

[10] Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer.

[11] Gaidos, E. (2013). Candidate planets in the habitable zones of kepler stars. *The Astrophysical Journal*, 770(2):90.

[12] Gaidos, E., Anderson, D., Lépine, S., Colón, K., Maravelias, G., Narita, N., Chang, E., Beyer, J., Fukui, A., Armstrong, J., et al. (2013). Trawling for transits in a sea of noise: a search for exoplanets by analysis of wasp optical light curves and follow-up (seawolf). *Monthly Notices of the Royal Astronomical Society*, 437(4):3133–3143.

[13] Howard, A. W., Marcy, G. W., Bryson, S. T., Jenkins, J. M., Rowe, J. F., Batalha, N. M., Borucki, W. J., Koch, D. G., Dunham, E. W., Gautier III, T. N., et al. (2012). Planet occurrence within 0.25 au of solar-type stars from kepler. *The Astrophysical Journal Supplement Series*, 201(2):15.

[14] Kane, S. R. and Gelino, D. M. (2012). The habitable zone and extreme planetary orbits. *Astrobiology*, 12(10):940–945.

[15] Kane, S. R., Hill, M. L., Kasting, J. F., Kopparapu, R. K., Quintana, E. V., Barclay, T., Batalha, N. M., Borucki, W. J., Ciardi, D. R., Haghighipour, N., et al. (2016). A catalog of kepler habitable zone exoplanet candidates. *The Astrophysical Journal*, 830(1):1.

[16] Kasting, J. F., Whitmire, D. P., and Reynolds, R. T. (1993). Habitable zones around main sequence stars. *Icarus*, 101(1):108–128.

[17] Kopparapu, R. K. (2013). A revised estimate of the occurrence rate of terrestrial planets in the habitable zones around kepler m-dwarfs. *The Astrophysical Journal Letters*, 767(1):L8.

[18] Lopez, B., Schneider, J., and Danchi, W. C. (2005). Can life develop in the expanded habitable zones around red giant stars? *The Astrophysical Journal*, 627(2):974.

[19] Maedche, A. and Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent systems*, 16(2):72–79.

[20] Purcell, C. (2017). What are exoplanets and can they sustain life?

[21] Quintana, E. V., Barclay, T., Raymond, S. N., Rowe, J. F., Bolmont, E., Caldwell, D. A., Howell, S. B., Kane, S. R., Huber, D., Crepp, J. R., et al. (2014). An earth-sized planet in the habitable zone of a cool star. *Science*, 344(6181):277–280.

[22] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.

[23] Udry, S., Bonfils, X., Delfosse, X., Forveille, T., Mayor, M., Perrier, C., Bouchy, F., Lovis, C., Pepe, F., Queloz, D., et al. (2007). The harps search for southern extra-solar planets-xi. super-earths (5 and 8 m _ {}) in a 3-planet system. *Astronomy & Astrophysics*, 469(3):L43–L47.

[24] Williams, D. M., Kasting, J. F., and Wade, R. A. (1997). Habitable moons around extrasolar giant planets. *Nature*, 385(6613):234.