Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)

كلية البيان
BAYAN COLLEGE

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

Big Data Science in Astronomy Especially with respect to the Search for Exoplanets aka Earth Analogue 2.0

Mohammed Sabahuddin Ansari
Research Scholar
Modern College of Business and Science
Muscat, Sultanate of Oman
Email: shafqans@gmail.com

Abstract

Big data Science in Astronomy for the Hunt for Earth Analogue as the current research embodies the rationale for such scientific data driven projects as far as the dystopian future of our planet earth is concerned. The research provides the conceptual framework for the various terminologies or jargons associated with exoplanet hunting or searching for the Earth's analogue along with the challenges embedded in the procedure to carry out the project. The research emphasis mainly on exoplanet hunting using transit method of planet hunting deploys the data science and its underlined algorithms like the KNN algorithm to illustrate the process hunting or searching an analogous earth 2.0. The research work also successfully demonstrates based on the data modeling techniques the ways and means to differentiate between an actual data set against outliers or anomalous data. The research deploys research methods namely observations, experiments, document screening, empirical data and scientific illustration in forms of graphs to illustrate the underlying research concepts pertaining to Big Data in astronomical science as an intensive data driven field of science. The results and findings are synonymous to the hypothesis of the research that Astronomical Science has undergone a paradigm shift from a theory based field to more of an intensively data driven field of science in the current day and age of Big Data Era.

Keywords: Exoplanet, 'Goldilocks Zone, habitable zone, Earth's Analogue, data mining, outliers, anomalous data, KNN algorithm

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

1. Introduction

1.1 Overview

Astronomy has undergone paradigm shift with respect to Big Data science. The ever developing advancements of ground and space-based observatories such as the large sky surveys brought Astronomy to the Big Data era. Gaia or Euclid are examples space bound initiatives but new ground based projects, such as the LSST or SKA never the less are undergoing state of the art upgrades. Data Mining has always been envisioned as the backbone of astronomical discoveries, in fact it won't be wrong to call Astronomers as true data miners in the field of data science. Data mining has always been the bedrock of astronomical discoveries, and it is the reason why astronomers are known as true data miners.

1.2 Research Objectives

Astronomers globally are trained to data mine in the following dimension:

- To investigate the importance of data science in astronomical discoveries.
- To characterize the known or Clustering.
- To devise models or algorithms for supervised learning & Classification.
- To establish the importance of machine learning in astronomical discoveries for semi-supervised learning.
- Outlier detection or Anomaly detection.
- To find any trend pattern in the data set.

The areas as mentioned above are absolutely critical as the field of astronomy is evolving as a **Data - Intensive** scientific field with extensive use of the **computer assisted data mining technology** (Y. Zhang, Y. Zhao, C. Cui.2002).

1.3 Rationale for the study

Challenges faced by Astronomy in the Big Data Era:

Astronomy as a scientific and research field has emerged into **space exploration industry** intensively driven by **Big Data** or **volumes of data sets** in the very fabric of astronomy as a science. The data or statistical appetite of astronomy based research or sky survey is huge and

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

unending flow of raw data in the form of Astronomical Imagery and empirical data further refined into large volume catalogues adding up to the science being done.

1.4 Problem Statement

Hence astronomical researches are changing from being just theoretical or hypothesis-driven to being data-driven to being data-intensive in this ever evolving age of Big Data. In astronomy **Data Volumes** are measured in **terabytes, petabytes, and even hexabytes** transforming the field of astronomy as a **Data – Intensive and Deep Data** field of science.

1.5 Research Questions

The following are the hypothesis or questions for the current research study:

- How important is data science in the field of astronomy?
- Has astronomy undergone any paradigm shift over recent two decades?
- Is the amount of data produced in astronomical observations and science difficult to manage?
- Has the level of data produced in astronomy heading towards Big Data?
- Does anomalous data exists in astronomical data science?
- What data mining models can be applied to analyze the astronomical data for exoplanet detection?
- Does the model analysis and findings indicate any trend pattern in the data?

1.6 Thesis Outline

- Acquiring raw image data.
- Cleaning, de-noising, stacking and sharpening the raw image data set.
- Acquired raw image data set integration.
- Safe Storage of the data.
- Processing the raw data volumes into processed information.
- Indexing heaps and heaps of astronomical image data.
- Searching, sorting and classifying the raw the data.
- Storing or indexing astronomical data into Catalogs.

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

- Sharing the astronomical data to the public domain.
- Transporting heaps of data over fiber optics every night from the observatory to the control center for data processing and analysis.
- Data Mining and Data modeling or visualization.

2. Literature Review

It is very evident and clear based on the challenges mentioned above that the traditional or manual tools cannot deal with such large amounts of data with various ground and space based telescopes doing all sky survey projects bring a data avalanche that the field of astronomy has to deal with (W. J. Frawley, G. Piatetsky-Shapiro, C. Matheus. 1991).

2.1 Proposed Solutions for Data Mining and Data Warehousing in Astronomy

- Optimized data acquisition.
- Precision raw data de-noising, cleaning and integration.
- Optimized astronomical object clustering and classification.
- Efficient handling of data dimensionality.
- On the spot and processed anomaly detection in the raw data.
- Better redundancy corrections.
- Effective management of huge astronomical catalogs.
- Faster and efficient data searching as well as fetching.
- Timely alerts of astronomical events.
- Enhanced data sharing in the public domain.
- Data accessibility 24x7 anytime or anywhere facilitating remote observing and data analysis without the need to be physically present in the observatory.
- Facilitating cross survey validations in order to ensure quality and precision in data analysis.

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

2.2 What are Exoplanets?

Exoplanets are planets beyond our own solar system. Thousands have been discovered in the past two decades, mostly with NASA's Kepler Space Telescope**.** These exoplanets come in a huge variety of sizes and orbits. Some are gigantic planets hugging close to their parent stars; others are icy, some rocky. NASA and other agencies are now looking for earth analogue, orbiting a sun-like star like ours in the habitable zone where it is not too hot and not too cold for liquid water to exist on the surface of the planets.

2.3 Searches for Exoplanets & Big Data in its Discovery

There are about 100,000,000,000 stars in our Galaxy, the Milky Way. How many exoplanets planets outside of the Solar System, do we expect to exist? Why are some stars surrounded by planets? How diverse are planetary systems? Does this diversity tell us something about the process of planet formation? These are some of the many questions that motivate the study of exoplanets (Y. Zhang, Y. Zhao, C. Cui.2002). However, detecting exoplanets is a huge and challenging task in itself. Exoplanets do not have light of their own rather it reflects the light of their host star. Thus, exoplanets due to its feeble light as compared to the extremely glaring light of their host star gets lost in its glare and hence it becomes a huge challenge for astronomers to spot an exoplanet around a distant star visually. There is a famous firefly analogy among astronomers to this fact that spotting an exoplanet is like spotting a firefly in front of a giant flood light 1000s of kilometers away. This is where data science comes into play where astronomers worldwide uses heaps of observational astronomical image data to hunt exoplanets (Aigrain S, Favata F. 2002).

2.4 Transit Method to Discover Exoplanets

Astronomers and scientists worldwide have devised techniques driven by data science whereby they measure star light or star brightness and looks for a tiny dip in its brightness that indicates the presence of a planet or planets around the distant stars they orbit. Considering this fact into account scientists at NASA developed a method which they called Transit method in which a

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

digital-camera-like technology is used to detect and measure tiny dips in a star's brightness as a planet crosses in front of the star. The diagram below illustrates the Transit Method, where a star is orbited by a planet or exoplanet. The figure bellows clearly depicts the transit event where it is visible that the starlight intensity drops because it is partially obscured by the planet. However, the starlight rises back to its original value once the planets crosses in front of the star or completes its transitory revolution around its host star.

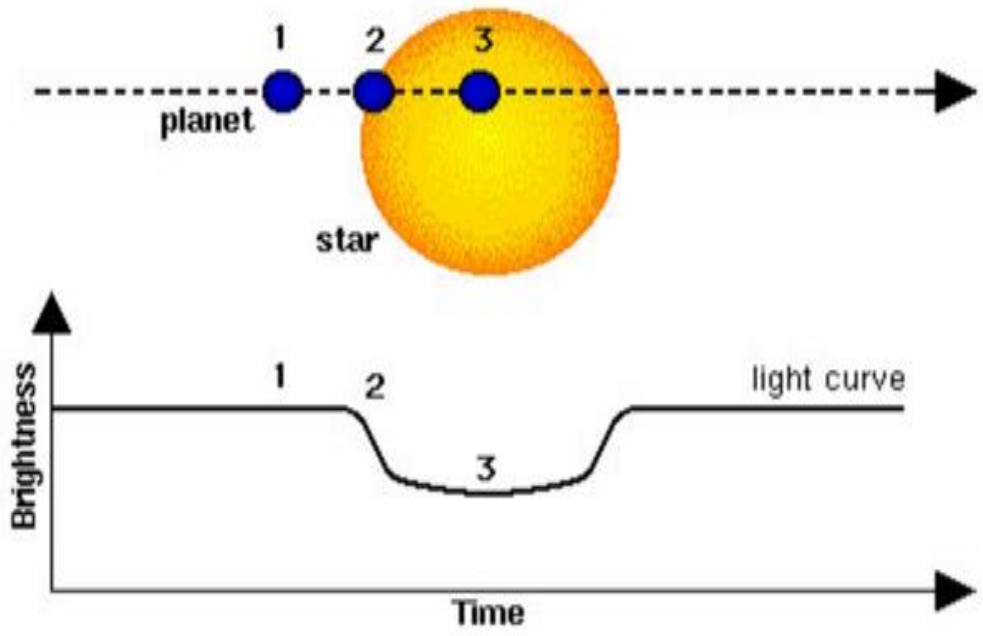


Figure 1 showing an exoplanet transit event (http://www.esa.int/Science_Exploration)

2.5 Data Mining Techniques in Exoplanet Hunting

Data Normalization is a technique often applied as part of **data** preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values.

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

2.6 The NASA Kepler Exoplanet Project

In (2009) NASA launched the telescope called Kepler, and the mission of this telescope is to discover planets similar to Earth and outside the solar system. As its main task is to send data, thermal and optical readings, everything that is similar to Earth's cups, and this represents a great challenge for researchers on a planet similar to Earth(Armstrong D, et al. 2015).

The mechanism of the Kepler telescope is to record the intensity of the light emanating from the stars, and when a planet passes around the star, it records the amount of change that occurs between them. As we can see in the images below, it indicates the passage of a planet around the star, and it has been named the candidate star system or probable candidate exoplanet signal.

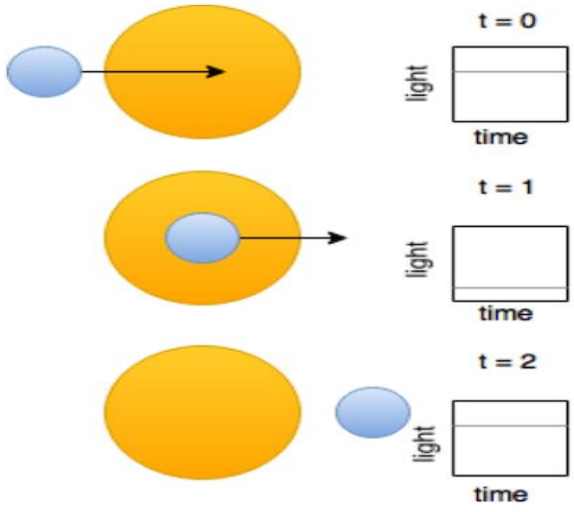


Figure 2 showing planet orbiting a star lowers light intensity (source: Kaggle)

2.7 Machine Learning and Exoplanet Hunting

As it is mentioned in above lines that exoplanets as compared to their host stars are cold, small and dark just like the firefly analogy where it's tricky to spot a firefly flying next to a searchlight from thousands of miles away. This is where data science and machine learning comes in very handy as state of the art technology to discover distant exoplanets (Armstrong D, et al. 2015).

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

One of the main ways astrophysicists search for exoplanets is by analyzing large amounts of data from NASA's Kepler mission with both automated software and manual analysis. Kepler observed about 200,000 stars for four years, taking a picture every 30 minutes, creating about 14 billion data points. Those 14 billion data points translate to about 2 quadrillion possible planet orbits! It's a huge amount of information for even the most powerful computers to analyze, creating a laborious, time-intensive process. To make this process faster as well as more effective, astronomers and data scientists turned to machine learning. Machine learning is a way of teaching computers to recognize patterns, and it's particularly useful in making sense of large amounts of data. The key idea is to let a computer learn by example instead of programming it with specific rules.

2.8 Advantages of Using ML in the Exoplanet Search

Deploying machine learning algorithms gives a flexibility because it learns using examples and not by concrete programming. The basic idea is teaching an algorithm to classify whether the input intensity or transit graph is of an exoplanet or not. It will save time that astronomers use to analyze these exoplanets to come with a conclusion (H. Zheng, Y. Zhang. 2008).

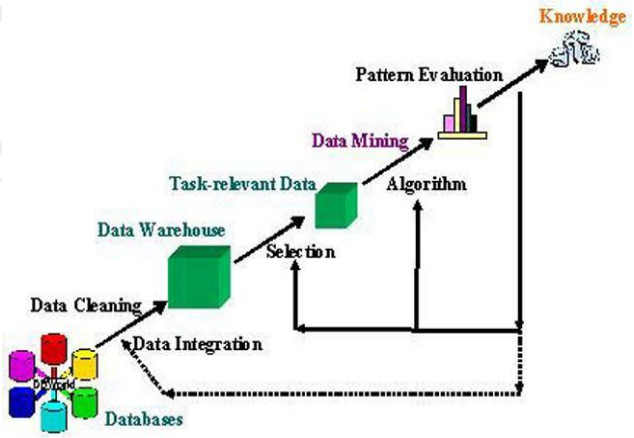
- **Neural networks:** A deep convolutional neural network is trained to test whether the transiting object causing an intensity dip is an exoplanet or not. There has also been a successful discovery of two new exoplanets using this neural network, after testing the Kepler data.
- **Known light curves processing:** Raw Light Curves were used as inputs to the neural network. Three types of neural networks can be used for classifying the dip as exoplanets and non-exoplanets. However for each type, three different input models are used namely global, local as well as global and local together.
- **Detecting anomalous data to target data differentiation:** Exoplanets and non-exoplanets are separated by a linear decision surface in the input image, so they are both linearly separable. Baseline model is a neural network with zero hidden layers.

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

- **Fully connected neural network:** Integrated Neural Networks helps to wipe put guess work or wrong data assimilation or inference thus keeping assumptions to the least for the input.
- **Convolution neural network:** It therefore "blends" one function with another for example, in synthesis imaging subtracting dirty data or anomalous data from the targeted clean data. It is also used for spatially structured input data like speech synthesis or in particular in the current research case of **image classification**.

2.9 Data mining in astronomy

Astronomy as field of science has faced a data avalanche due to advances in telescopes and detectors or camera technology in synergy with the exponential increase in computing capabilities, improvements in data-collection methods, and successful applications of theoretical simulations the expected data volumes will add up to terabytes, soon to be followed by petabytes (Armstrong D, et al. 2015). Proper management and processing of massive data sets requires efficient techniques for optimizing database technologies. However, mining knowledge from huge data volumes is the ultimate goal and development of data-mining techniques like the Knowledge discovery in databases (KDD) is the process of extracting useful knowledge from data. Data mining deploys specific algorithms to discover rare or previously unknown types of object or phenomenon or pattern. Common KDD functions are classification, cluster analysis, and regression.



ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

Figure 3 showing knowledge discoveries in databases (https://webdocs.cs.ualberta.ca/)

Cluster (or clustering) is data technique whereby objects are placed or categorized into more or less homogeneous groups such that the relationship or pattern between groups is revealed. It lacks an underlying body of statistical theory and is heuristic in nature, requiring decisions to be made by individual users manually (H. Zheng, Y. Zhang. 2008).

Cluster analysis is particularly useful when groups or objects are classified more objectively than subjectively and can help astronomers find outliers or anomaly or even an unusual pattern within a flood of data. Examples include discoveries of high-red shift quasars, type-2 quasars (highly luminous active galactic nuclei whose centers are obscured by gas or dust and not planet or any rocky object), and brown dwarfs thus ruling out the existence of any exoplanet signal what so ever.

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
كلية البيان
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

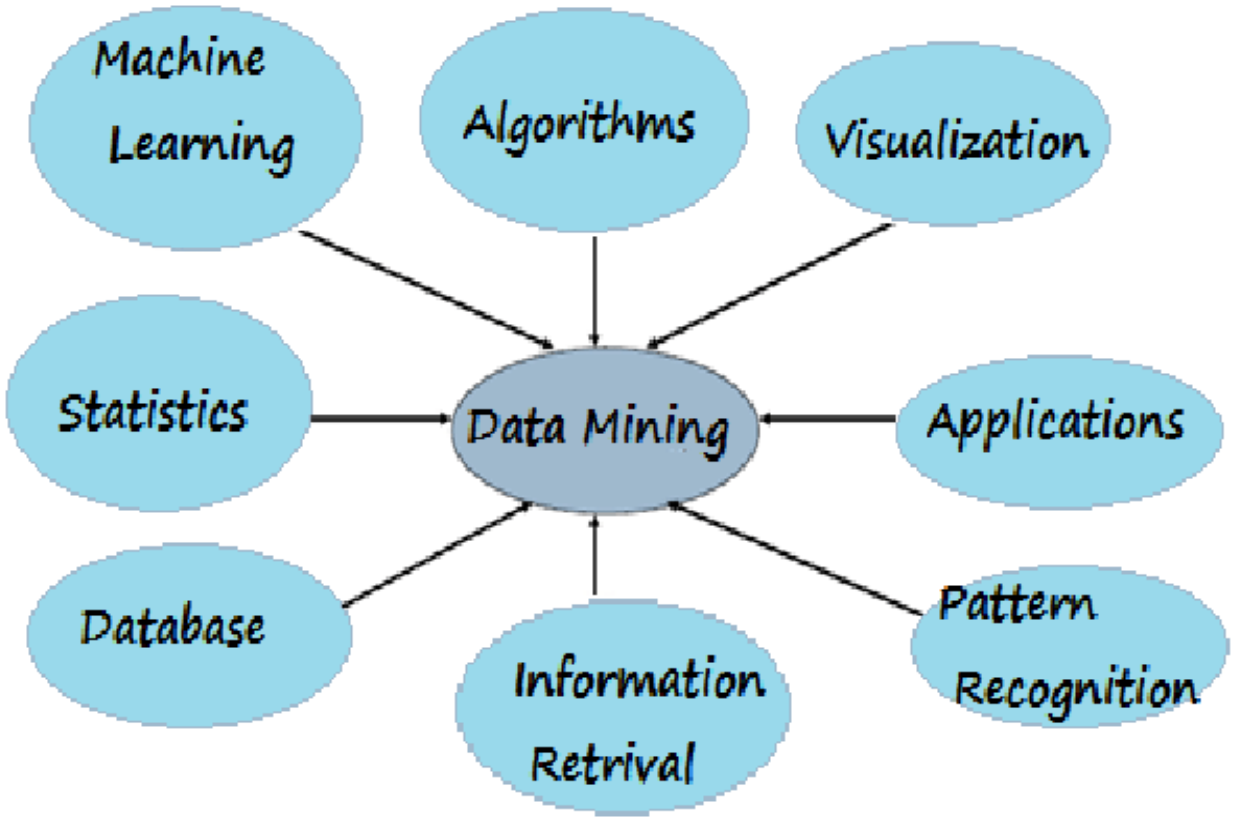


Figure 4 illustrating the Data Mining Architecture

2.10 Big Data in Astronomy

There are many different definitions provided by researchers for the characteristics of big data, and the most important of these definitions provided by the researcher Kirk Burney called (10V): which are vagueness, truthfulness, speed, variety, location, vocabulary, change, value, validity, size and ambiguity (Y. Zhang, Y. Zhao, C. Cui.2002).

Volume its big data, measured in petabytes, terabytes, and Exabyte. Big data always presents challenges in processing, analyzing, storing, indexing, searching, sharing, and a lot of services. All traditional data management tools cannot handle this big data. There are a lot of projects that

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

take out a survey of the wide sky. This table displays the data volumes resulting from sky surveys.

2.10.1 Data volumes of different sky survey projects.

Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected

Table 1 showing actual data volumes generated by sky surveys

Variety refers to the data complexity and variable nature of the data sets itself. Astronomical data mainly include images, spectra, time-series data, and simulation data. Most of the data are saved in catalogues or databases. The data from different telescopes or projects have their own formats, which causes difficulty with integrating data from various sources in the analysis phase such **as** structured, semi-structured, unstructured, and mixed (W. J. Frawley, G. Piatetsky-Shapiro, C. Matheus. 1991).

Velocity however refers to the speed with which data is produced, transmitted and analyzed. As far as data volume is concerned, LSST observatory will generate one SDSS each night for 10

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

years. However how efficiently mining is carried out like correctly classify candidate objects as against the target candidates while making follow-up observations for the subsequent decade time making a huge challenge for astronomers.

Value refers to the value adding factor in terms of uniqueness, authenticity and correctness to astronomical data being collected each observational night as well as its target data mining. It is interesting and inspiring in astronomy to discover surprising, rare, unexpected, and new objects or phenomena. Thus, the discovery of a new distribution trend or pattern or law provides great value to the data driven astronomical sciences.

2.10.2 Recent Exoplanet Discoveries Using Big Data & Neural Network-Machine Learning

Through the (Kepler) space telescope, planets were recently discovered through a neural network to analyze data and identify the signals of promising planets with great accuracy. This was a preliminary analysis of hundreds of stars and planets. We consider this discovery a success of the concept of using machine learning to discover exoplanets. Machine education is used in many different scientific disciplines such as chemistry, physics and health care.

2.11 A Planet Hunting Primer

The plot below is known as a light curve showing the brightness of the star (as measured by Kepler telescope's photometer) over time. Thus, as a planet passes in front of the star, it

temporarily blocks some of the light, causing a dip as shown in the diagram below in the star brightness and eventually increasing again shortly thereafter, causing a "U-shaped" dip in the light curve(Y. Zhang, Y. Zhao, C. Cui.2002).

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

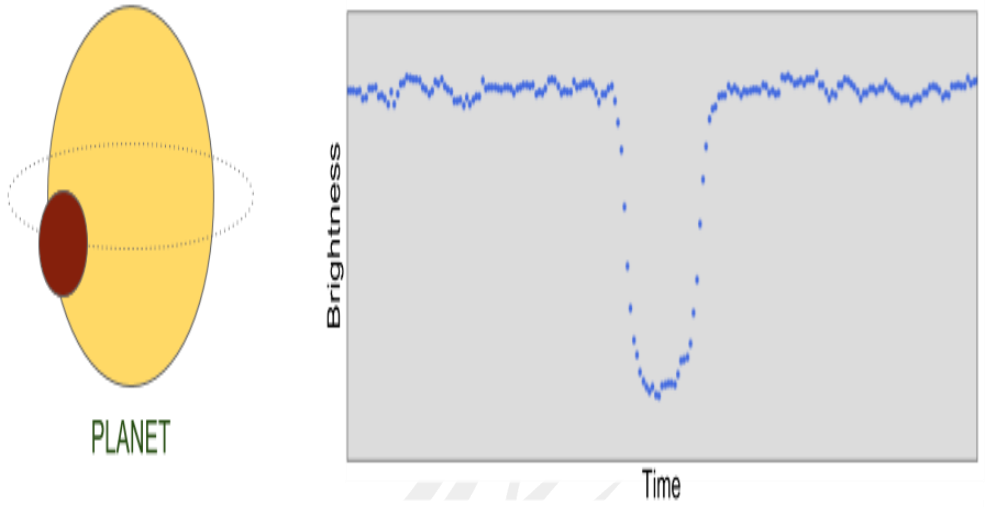


Figure 5 showing a light curve from the Kepler space telescope with a "U-shaped" dip thatindicates a transiting exoplanet.

2.11.1 Anomalous Data Mimicking Exoplanet Transit or False Positive Detections

Astronomical phenomena and many devices can cause star measured brightness to decrease, such as mechanical noise, star spots, binary star systems and cosmic-ray strikes on the Kepler optical scale.

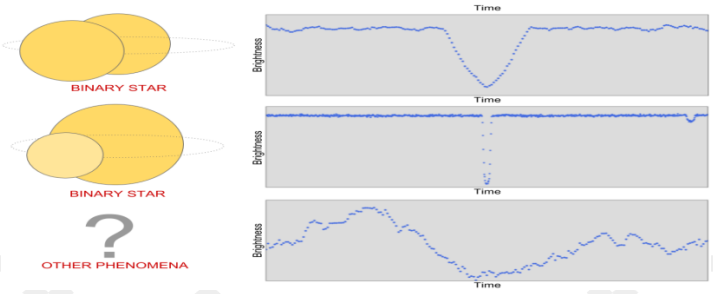


Figure 6 showing an anomalous or outlier event micmicking an exoplanet transit

The figure above illustrates the first light curve has a **"V-shaped"** pattern indicating that a very large astronomical object passed in front of the star as observed that Kepler telescope. However the**second light curve contains two dips** as shown in the diagram above indicating a **binary star system revolving around each other** where the larger dip is caused by the dimmer star

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

passing in front of the brighter star, and vice versa. Finally the **bottom most light curve in the diagram above is an example of the miscellaneous other non-planetary object signals** causing uneven variable dips in the star brightness light curve (Y. Zhang, Y. Zhao, C. Cui.2002). The search for the planet signal in the Kepler data, starts with scientists using automated software (e.g. the Kepler data processing pipeline) detecting candidate brightness dip signals that might be caused by planets, and then manually following up to come up with the result **whether the signal is a planet or a false positive or anomaly or outlier data.**

2.11.2 Applying the Machine Learning Approach

Applying the automated approach to analyzing the large amount of data, as the (Google Brain) team applied machine learning to a lot of varied data. And because of the large amount of data that arrives through the experiments, the best solution to processing this data is a machine learning system. In one experiment, Kepler collects billions of data that researchers have to process. The machine learning system was used to analyze old data, and the system proved effective in the analysis. We have also been working collaboratively with Andrew Vanderburg and have designed a network to assist in the search for the discovery of low-signal planets.

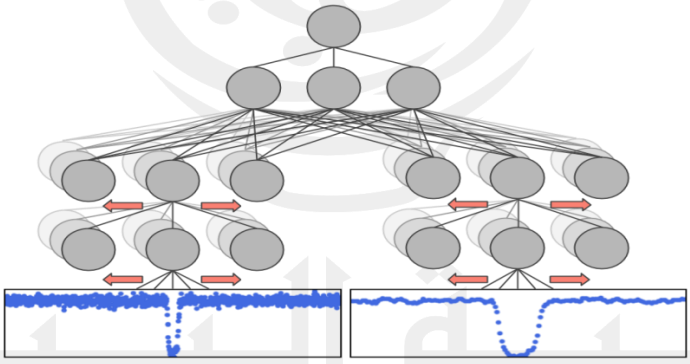


Figure 7 demonstrating the machine learning approach

Many of the signals have been manually checked and exceed 30,000 Kepler signals. In addition, a lot of subgroups have been used in research to find out which planets are real from that do not

exist. More than 2000 planets have been scientifically verified. There are two different widths in each light curve. The first is a wide width curve through which researchers can examine the signal from another place on the light curve, while the second is a magnified width curve and researchers can examine the shape of the signal, the sign shapes are a letter (V) and Letter U(H.

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

Zheng, Y. Zhang. 2008). So in this experiment it becomes clear to us from the light curve that it is not a planet. This is inferred by the green points heading downwards, which coincide with the secondary declining indicator of the binary system. The model's prediction changes when these points are obscured, and the possibility may be that it is a real planet. When the zoom in the main central depression appears in fact in the form (V), it indicates the binary system.

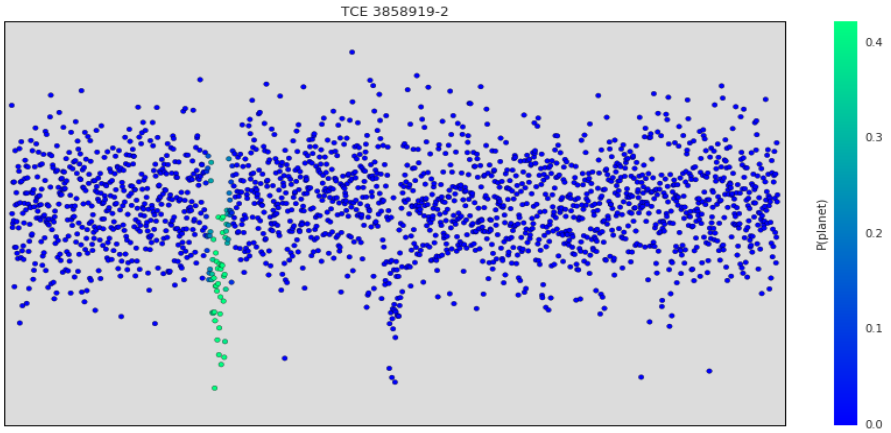


Figure 8 showing an Anomalous Light Curve Data of Binary Stars mimicking an Exoplanet transit event

2.12 Applying Big Data & Machine Learning to Find Our Own Exoplanet

There are many stages to announce the discovery of a real planet, for example Kepler-90i planet has gone through the following stages of discovery, first making a clear plan and preparing and training a complete model through a special code. Upon completion of the first plan, enormous data is collected and downloaded, which may take a long time. Upon completion of this stage, researchers work to make models and predictions about the new data. A popular way to find and display new mentions is to use Box Least Squares. This method works as follows, it searches for the periodic decrease in brightness (as shown in the following picture). The algorithms of this method show the (U-) shape as the planetary signals, and the (V-) shape as the binary star signals. There are also many false shapes that are not known as to what they indicate. The BLS

method at the end of the analysis allows you to see the discovered planets with the naked eye (Y. Zhang, Y. Zhao, C. Cui.2002).

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

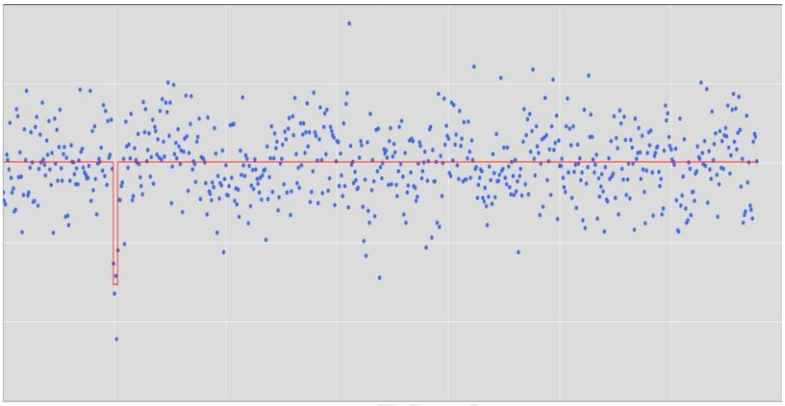


Figure 9 showing the dip as an indication of an exoplanet transit

Through the experiment, a decrease in the signal of the previous star was detected in the light curve. The period taken for revealing the results is (14.44912) days, and for (2.70408) hours, and the start will be from (12:00) on the date (1/1/2009). This next command triggers the detected signal through the demo model.

```
python predict.py  --kepler_id=11442793 --period=14.44912 --t0=2.2
--duration=0.11267 --kepler_data_dir=$HOME/astronet/kepler
--output_image_file=$HOME/astronet/kepler-90i.png
--model_dir=$HOME/astronet/model
```

Through the previous results, it appeared that the forecast result is less than 1%. Which means that the model is confirmed by a large percentage, which is more than 90%, It indicates the existence of a real planet outside the solar system. There are many models and predictions for the discovery of planets outside the solar system, and this is based on only one model, and this signal that reached us has proven correct, and the name (Kepler-90 i) has been launched on this planet as a real planet(Armstrong D, et al. 2015).

3. Methodology

3.1 Research Methods

Research method refers to the **ways and means of doing the research in terms of data collection, analysis or findings and deriving inference.**

Thus, research methods to uncover new information or create a better and deeper understanding of the topic are strategies, processes, or techniques used in collecting data or evidence. There are many different ways to collect data through specialized tools, including:

- **Qualitative Research**

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

This type of research is based on the collection and analysis of data that have been conducted in previous experiments. These data may be behaviors, meanings, or live experiences. This type of experiment is useful for researchers to gain the best concepts or interactions. It is also useful for exploring what happened, how it happened and why it happened. We can say that it helps to describe events and explain experiences.

- **Quantitative Research**

  This type of research is based on the collection of numerical data that can be measured or classified through analyzes. This type of research helps to make generalizations and uncover relationships and patterns. This type of research is characterized by knowledge of number, quantity and repetition (Y. Zhang, Y. Zhao, C. Cui.2002).

- **Mixed Methods Research**

This method is considered as a combination of the best research methods, namely qualitative and quantitative research. This research helps to provide a comprehensive approach, and by virtue of it combining the best types of methods in the search, it collects and analyzes data in depth to obtain the best results. Also, one of the advantages of this type of research is that it works to verify data through several sources at the same time.

3.2 Data Collection Procedure

Data can be either primary data or secondary data, where primary data can further be quantitative or qualitative or may be both dependent on the research being pursued. Nevertheless whatever be the type of data needed to be collected it usually should follow a pre-set of procedures for its gathering (Baluev R. 2018).

3.3 Methodology adopted for the current research study

The current research study employed observations, document screening and experiments conducted by the researcher to simulate the actual astronomical data acquisition, cleaning, sorting, stacking and integrating to convert it as a valid input for data analysis.

3.3.1 Observations

Observing the process of astrodata mining and visualizations done by field scientists together reading with their subsequent papers published on it along with viewing observantly their public talks on the topic serving the current research with the needed conceptual framework as well as the required resource to fuel the research study.

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

3.3.2 Document Screening

Document screening refers to the documents and text information whether that be in the form of published paper or journal or even articles as published by NASA media sites bounding the topic of research providing the treasure trove of materials to empower the research study under consideration (Kovacs G .2017).

3.5.3 Experiments

The current research study deals with data science with respect to exoplanet hunting. The study elaborates and delves into the significance of data in astronomical discoveries while at the same also illustrates the process of exoplanet hunting right from the get of getting raw images of the distant star systems, cleaning data, sorting, classifying , stacking and last but not the least analyzing or visualizing the data.

The following are the process were conducted under the current research study to simulate the actual process of data gathering and mining in the field of astronomy to discover newer planets: **Image acquisition** via 80 mm **Sky Watcher telescope** paired with computerized **Celestron AVX** mount for pointing and maneuvering the telescope to target area in the sky.

Fig 10 Showing Observatory grade equipment were used for astronomical data acquisition

- **Physical skimming of the images taken** for dirty image data detection.
- **Classifying and verifying data** by plate solving the image for its coordinates or location in the sky to ensure the image frame contains the target object or target star in the research case.

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

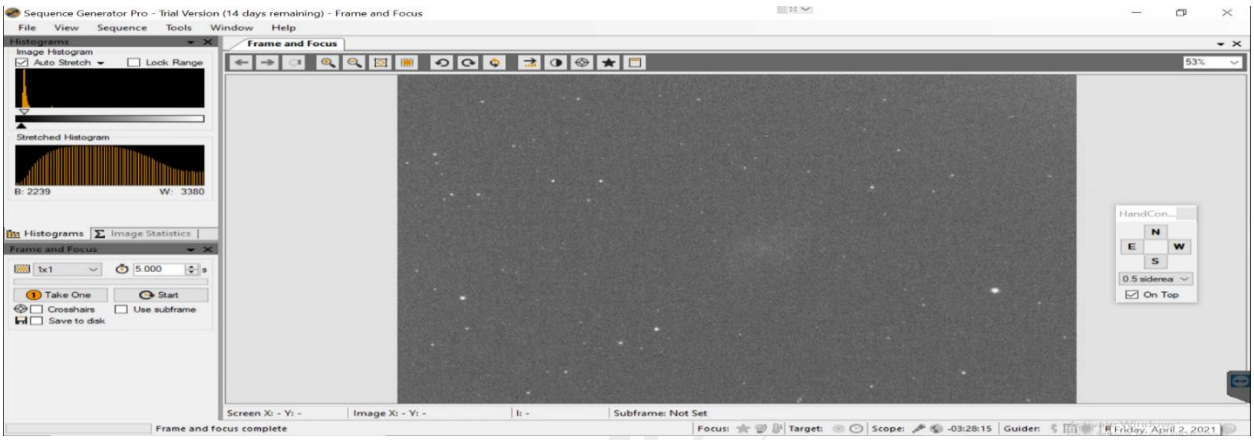


Figure 11 showing Image acquired of the target object area in the sky using Sequence Generator Pro

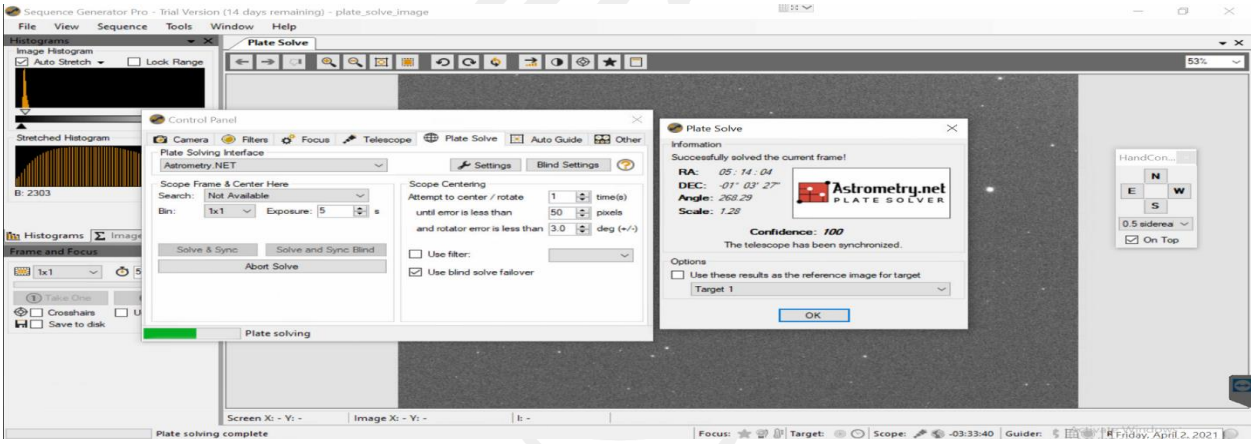


Figure 12 showing Image classification & verification via plate solving that compares the image with the catalogue database of Astrometry.net online.

- **Stacking data** using the Deep Sky Stacker software for aligning and stacking image data to improve signal to noise ratio or the SNR.
- **Preparing the light Curve** from the acquired data over a time period of images.
- **Plotting the Light Curve using AstroImageJ light curve visualization software or Excel Sheet** to see the regular dips in the curve to indicate the presence of planet orbiting the target star in regular intervals.
- **Documenting** the entire process and verifying the results with previous successful experiments already done by fellow researchers.

4. Data Analysis and Findings

Data Analysis. **Data Analysis** It is a process by which data is prepared in various mathematical or logical methods

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

available to study each component of the research. It is also a process of systematic application of describing, summarizing and evaluating data. This is to access useful information that is useful for research and analysis.

4.1 Data Mining Models Applied

4.1.1 K-Nearest Neighbor Algorithm Overview

The k-Nearest Neighbors (KNN) algorithm stores the training data and searches for the closest k records (Euclidean distance) of this new record as and when a new object is submitted for classification. Thus, the new record is ranked in the most common class among all the k closest records.

Thus the K-Nearest Neighbor algorithm applies the concept of centroids. Therefore, for a given a data set, the algorithm randomly selects k records, each representing a cluster. For each remaining record the similarity between the analyzed record and the center of each grouping is calculated. The object is inserted in the cluster with the shortest Euclidean distance, that is, greater similarity and, with each new element inserted, the centroid is recalculated (Armstrong D, et al. 2015). Then a distance matrix is created between each point and the centroids. Each point is placed in the classes depending on the distance of the centroid from the class. Thus, all new centroids are calculated for all categories with repetition until convergence is achieved. It should be emphasized that different variations implement optimizations for the choice of the k value, measures of dissimilarity and strategies for the calculation of the cluster center. One variation of this is the k-Means algorithm that uses the mode to calculate the centroids.

4.1.2 Advantages of K-Nearest Neighbor

The advantage of nearest-neighbor classification are as follows:

- One of the most important features of K-NN is ease: as it is easy to use and easy to implement, the K-NN algorithm detects the nearest neighbors by classifying the new data point.
- There are no assumptions in K-NN: they are considered nonparametric algorithms. It also contains a set of parametric models and example models: Linear Regression.
- In K-NN there is no training step: in general, K-NN is not based on building new models but rather relying on previously existing models. As it relies on entering new data based on historical data.
- One of the strongest features of K-NN is that it can be used for both regression and classification problems (Y. Zhang, Y. Zhao, C. Cui. 2002).
- Continuous improvement that K-NN relies on like-based learning. Which represents the memory-based approach. As it allows algorithms to quickly respond to variables during use in real time.
- Most of the algorithms are easy to implement and do not require effort to work on binary problems.
- K-NN has a set of criteria for choosing the distance, among them: Models are built in selecting the distance

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

followed in the K-NN algorithm on the basis of:

- 1. Euclidean
- 2. Hamming
- 3. Minkowski
- 4. Manhattan

4.1.3 Disadvantages of K-Nearest Neighbor

However, K-NN has several advantages but there are certain very important cons of K-NN that are as follows:

- The first negative in K-NN is slow: K-NN is easy to implement and use, but when the data set increases, the speed of the algorithms decreases dramatically.
- Curse of Dimensionality: When there are a small number of input variables, K-NN works fine, but if there are too many input variables, it is difficult for K-NN to find the new data point.
- K-NN needs homogeneous features: K-NN relies on the use of common distance, as an example of distances is Manhattan and Euclidean, and the features must or must have the same scale.

- Optimal number of neighbors. One of the biggest problems with K-NN is the number of neighbors that must be taken into account when selecting K-NN for data classification.
- Outlier sensitivity: The K-NN algorithm is sensitive to outliers due to its reliance on distance in selecting neighbors.
- Missing Value treatment: K-NN cannot deal with the problem of missing value, which is one of the big drawbacks.

4.1.4 Comparison of various Data Mining Models that can be applied in the current research

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
كلية البيان
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

Comparison of various classifiers

Algorithm	Features	Limitations
C4.5 Algorithm	- Models built can be easily interpreted - Easy to implement - Can use both discrete and continuous values - Deals with noise	- Small variation in data can lead to different decision trees - Does not work very well on small training dataset - Over-fitting
ID3 Algorithm	- It produces more accuracy than C4.5 - Detection rate is increased and space consumption is reduced	- Requires large searching time - Sometimes it may generate very long rules which are difficult to prune - Requires large amount of memory to store tree
K-Nearest Neighbour Algorithm	- Classes need not be linearly separable - Zero cost of the learning process - Sometimes it is robust with regard to noisy training data - Well suited for multimodal classes	- Time to find the nearest neighbours in a large training dataset can be excessive - It is sensitive to noisy or irrelevant attributes - Performance of the algorithm depends on the number of dimensions used

Table 3 showing comparisons of various classifiers

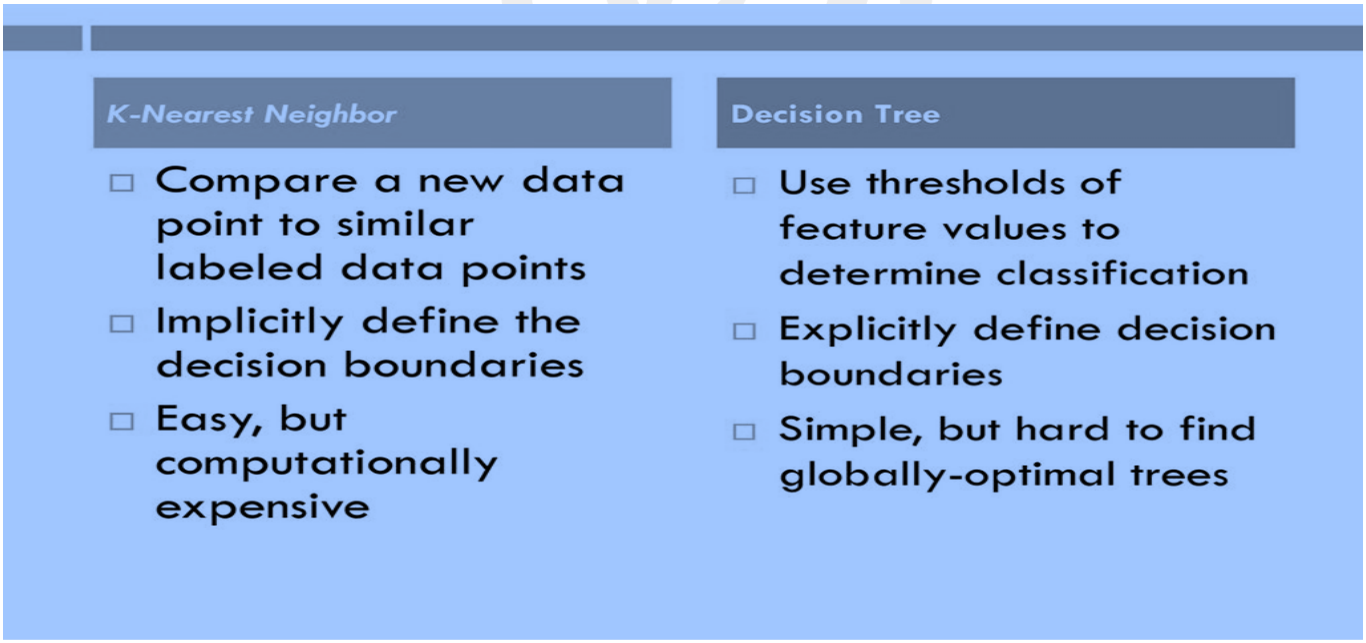


Table 4 showing comparisons of various data mining models

4.2 Major Findings Based on the Models

The findings based on the Decision Tree and K-Nearest Neighbor Models with respect to two major attributes of

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

the data set namely Mass and Radius of the planet are as follows:

- Through the initial verification of this analysis, an important result was obtained, which is that there is no correlation between the stars and the planetary parameters. This means that the planet's mass or radius has no direct relationship with the host star.
- Classification results models refer to two dimensions, namely the radius of the planet and its mass with several different classes: Neptunian, Jovian, Superterran, Subterran and Terran types.
- The algorithms KNN can be used in the surrounding environment of processing centers, but most of the time the kNN algorithm is approved and compiled (Kovacs G 2017).
- Although there is a low data engagement index and high correlation.
- Decision tree provides instant result based on conditional parameters unlike KNN

4.3 Other Findings

Thus, data analysis refers to converting raw data into meaningful relational data or information for deriving target results from it. In the current research case under consideration it refers to establishing the significance and indispensability of data in the field of astronomical sciences. The data analysis takeaways for the current study based on the conceptual detailing in the literature review based on the published papers, journals , articles and previous researches are as follows:

- Astronomy has evolved from a theoretical or hypothesis based science to a data driven science.
- Data is in the very foundation of astronomical science without which it’s virtually impossible to proceed.
- Exoplanets are so far into space that it’s virtually impossible to even sniff its very existence in front of the very bright parent star that it is orbiting around (Aigrain S, Favata F 2002).
- It’s like conferring or detecting a fly flying in front of a lighthouse 1000s of kilometers away.
- It is the data science that makes this monumental task possible by applying data mining techniques.
- Once the astronomical data are collected in the form of set of images over a time period with respect to the target star under observation these image data is then plotted in an excel sheet to produce a light curve.
- This light curve based on the data acquired depicts whether or not a planet exists around a distant star under observation by measuring a feeble dip in brightness of the parent star when the planet passes directly in front of the star.
- This ever so tiny dip is almost impossible to detect by naked eye or via telescope alone unless and until images acquired (via the telescope and dedicated astrocams with sensitive sensors or detectors) are inputted into computer equipped with machine learning algorithms geared towards detecting the dip.
- This image data when cleaned, aligned and stacked to plot in an excel sheet to come up with a light curve it is then the dip which the sensors detected but unable to visualize for the observers to actually see finally

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

comes into view by naked eye in the form of dip curve in the plot.

- The decrease in the curve of the light indicates the rotation of a planet around a distant star, which is impossible to discover except with the help and use of data science tools.

5. Conclusion

Scientists worldwide are encouraged to analyze light curve data setsas produced by Kepler and similar missions, the most important of which is Kepler emission and Transiting Exoplanet Survey Satellite mission. This research emphasizes on the astronomical discovery based on data mining in other words analyzing the data to discover many exoplanets outside of our own solar system. Also, from this data, scientists were able to better distinguish the real planets through the radial velocity curves and the light curves. All research bears error or right, and checking or analyzing data manually is one of the most difficult things and needs a lot of effort. Therefore, the use of semi –supervised machine learning is deployed as the most widespread use technique or procedure in planetary discovery research, as the basis for scientific accuracy of the information or astronomical data acquired. Through this research, several planets were analyzed through algorithms and large data that helped the success of this research. To this day, many extra solar planets have been discovered through this process (Armstrong D, et al. 2015).

There are four steps in which algorithms are analyzed and they are as follows:

* Determine and indicate if the planet is in the solar system or outside the solar system group.
* Determine the strength of the detected signal.
* Probable initial detection of exoplanet signals.
* Processing the detected optical curve.

The diagram below illustrates the process of data acquisition to data analysis in astronomical discoveries especially with regard to exoplanet hunting.

Bayan College International Journal of Multidisciplinary Research
(Quarterly, Peer Reviewed International Journal)
BAYAN COLLEGE
ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

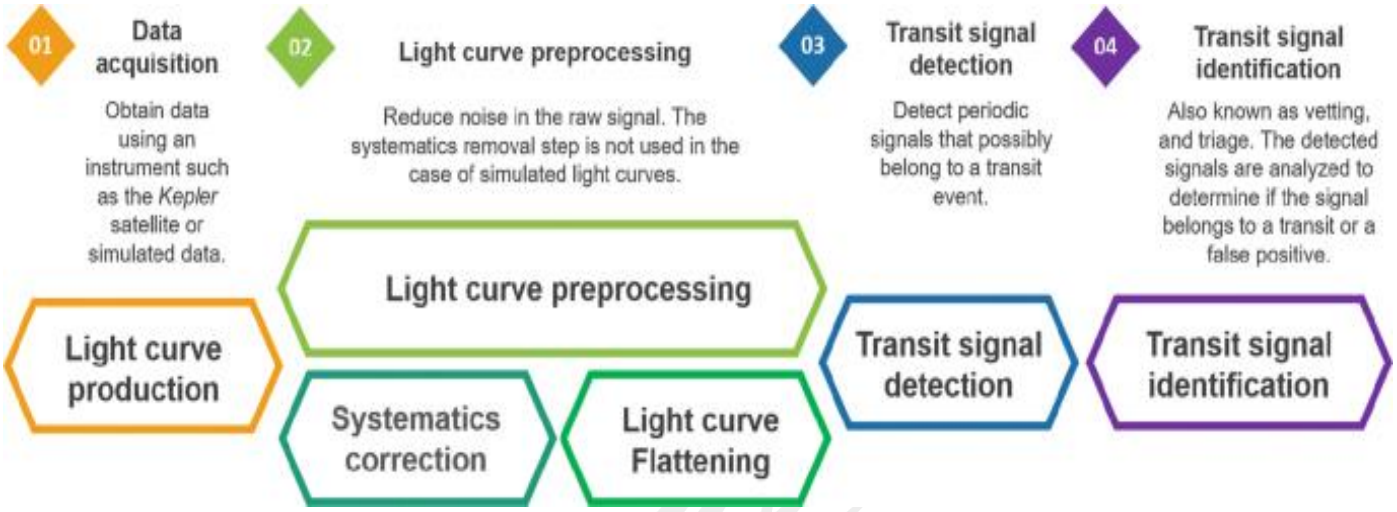


Figure 13

Showing the process of raw data acquisition for light curve analysis to detect candidate exoplanet signal

6. Future Work

The research under consideration has fetched its researchers with treasure trove of knowledge into the field of astronomical science and its deep synergistic connection with data science. The research not only concluded that data remains the backbone of astronomical discoveries but it also illustrated ways and means to mine data to produce target data visualization results.

However, the researchers of the current research study would like to pursue the following areas in this field of astronomical data science:

- Acquiring spectral data of distant star systems and sniff atmospheric data of exoplanets.
- Detecting near earth objects.
- Applying for observation time with leading observatories for a high resolution observation of astronomical objects.
- Hunting asteroids and comets as they pass by earth.
- Probing those stars that are not known to have exoplanets to find data otherwise.
- Calculating distances between our solar system and galaxy to other distant galaxies or solar systems.
- Start to learn astronomical data collection via radio astronomy.

References:

1. W. J. Frawley, G. Piatetsky-Shapiro, C. Matheus, Knowledge discovery in databases: an overview, *Knowledge Discovery in Databases*, pp. 1-30, AAAI Press/MIT Press, Cambridge, MA, 1991.
2. H. Zheng, Y. Zhang, Feature selection for high-dimensional data in astronomy, *Adv. Space Res*. 41, pp. 1960-1964, 2008.

ISSN (Online)- 2710-2432
ISSN (Print)-2730-2424
Vol. 04, Issue 01, January 2024

3. Y. Zhang, Y. Zhao, C. Cui, Data mining and knowledge discovery in database of astronomy, *Prog. Astron*. 20, no. 4, pp. 312-323, 2002.
4. D. Wang, Y. Zhang, Y. Zhao, An automatic system for photometric redshift estimation based on sky survey data, *Proc. SPIE* 7019, pp. 701937, 2008. doi:10.1117/12.788429.
5. Aigrain S, Favata F (2002) Bayesian detection of planetary transits. a modified version of the Gregory-Loredo method for bayesian periodic signal detection. Astron Astrophys 395:625–636. https://doi.org/10.1051/0004-6361:20021290.
6. Armstrong D, et al. (2015) K2 variable catalogue II: Machine learning classification of variable stars and eclipsing binaries in K2 fields 0-4. Mon Not R AstronSoc 456:2260–2272. https://doi.org/10.1093/mnras/stv2836.
7. Baluev R (2018) Planetpack3: A radial-velocity and transit analysis tool for exoplanets. Astronomy and Computing 25:221–229. https://doi.org/10.1016/j.ascom.2018.10.005.
8. Foreman-Mackey, et al. (2015) A systematic search for transiting planets in the K2 data. Astrophys J Lett 806:215. https://doi.org/10.1088/0004-637x/806/2/215.
9. Kovacs G (2017) Synergies between exoplanet surveys and variable star research. EPJ Web of Conferences 152:01005. https://doi.org/10.1051/epjconf/201715201005.