Machine Learning for Astronomical purposes

By

AKSHA THAKKAR - 17BIT003

DEPARTMENT OF INFORMATION TECHNOLOGY
Ahmedabad 382481

Machine Learning for Astronomical purposes

Minor Project

Submitted in fulfillment of the requirements

For the degree of

Bachelor of Technology in Information Technology

By

AKSHA THAKKAR- 17BIT003

Guided By

Dr. Swati Jain

[DEPARTMENT OF INFORMATION TECHNOLOGY]

DEPARTMENT OF INFORMATION TECHNOLOGY

Ahmedabad 382481

CERTIFICATE

This is to certify that the project entitled "Machine Learning for Astronomical Purposes" submitted by AKSHA THAKKAR(17BIT003) towards the partial fulfillment of the requirements for the degree of Bachelor of Technology in Information Technology of Nirma University is the record of work carried out by him/her under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination.

Dr. Swati Jain
Professor
Department of Information Technology,
Institute of Technology,
Nirma University,
Ahmedabad

Dr. Madhuri Bhavsar
Head Of Department , CSE
Institute of Technology,
Nirma University,
Ahmedabad

ACKNOWLEDGEMENT

I would like to express my most profound thankfulness to each one of the individuals who provided me any kind of help to finish this project. I acknowledge with thanks, the gentle support rendered by Dr. Swati Jain, under whose guidance I am able to complete the task in a given period of time.

I would also like to appreciate the suggestions given by my friends for make this project more enhanced. At the home front, I am very thankful to my relatives for the backing and consolation I got from them in effectively finishing the project.

ABSTRACT

The accompanying report is about Machine Learning for Astronomical Purposes. At first, we talk about the basic concept of Exoplanets, how to access its data and its characteristics. At that point we take a leap to the various indexes required to obtain information from this data. This module perceives the discourse with the assistance of various techniques to detect exoplanet possibility and habitability detection to get insightful observations.

CONTENTS

1. Introduction

1.1 General

In recent times there has been a lot of developments in Machine Learning and Deep Learning. This has given a boost to various fields of scientific study too like life sciences, particle physics, astronomy etc.. It has become one of the essential tools to study and analyse large data sets, helping to identify certain patterns and perform large scale research work which wasn't possible earlier. For e.g. one of the most famous missions, the Large Synoptic Survey Telescope (LSST) is expected to generate 15 TB of data each night from its tools to capture images and other astronomical information. Hence it becomes important to design ways to properly categorize this data and use it to make our work easy. Implementation of Deep Learning Techniques are getting better and better with time which has allowed us to overcome challenges which seemed impossible earlier. Detecting Exoplanets from far distant galaxies seemed to be a very far fetched goal earlier but now not only detecting but we have enough tools to check if that planet is habitable.

1.2 Scope of Work

The Study covers the basic concepts of Astronomical Data Analysis using Machine Learning and its basic working. It also covers the theoretical foundations which are needed to understand and apply techniques to detect Exoplanets. We also learn about various techniques to preprocess Astronomical data. Finally we dive into some applied examples to check the habitability of those planets to get the proper understanding of applications of Astronomy using Machine Learning.

2. Literature Survey

The Literature survey covered papers, books and online resources that explained Astronomical data analysis and Machine Learning in detail and discussed the methods to apply them. Papers and blogs that focused primarily on the technology underlying Astronomy were studied to determine the topics to be covered for the Minor Project; subsequently those papers were chosen that aligned with the topics decided previously.

3. Exoplanets

3.1 Introduction

An exoplanet or extrasolar planet is a general term for a planet which is found to be outside the Solar System. In the search for extraterrestrial life, detereming and identifying exoplanets has become very important. There is extraordinary interest in planets that circle in a star's orbitable zone, where it is feasible for fluid water, an essential for life on Earth, to exist on a superficial level. The investigation of planetary livability additionally thinks about a wide scope of different factors in deciding the appropriateness of a planet for facilitating life. The following is the picture of Exoplanet HIP 65426b is the primary found planet around the star.

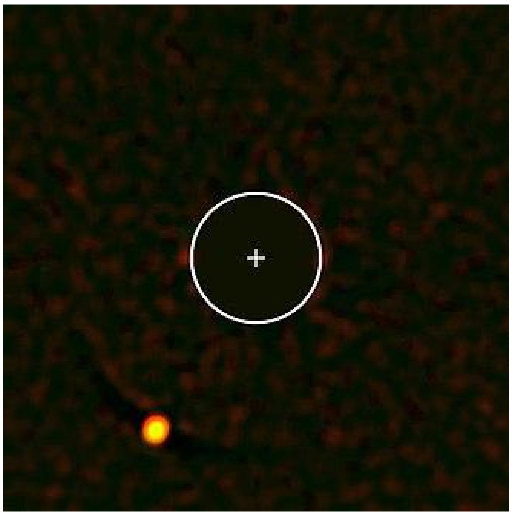


Figure 1 Confirmed Exoplanet HIP 65426b

3.2 Datasets used in the project

Here we make use of two different datasets in the project.

- The first dataset is extracted from the NASA's Exoplanet archives. Here in this particular model we have used Kepler's Object of Interest Dataset. It contains stellar features of celestial bodies which are necessary to determine if that particular object is exoplanet or not. The dataset mainly contains parameters like the diameter, distance, temperature, weight etc.. of the exoplanet. A KOI is a defined as a target which is identified by the Kepler Project after years of research. It generally displays at least one transit-like sequence within Kepler time-series photometry which is identified to be of astronomical origin only and also consistent with a planetary transit hypothesis.
- To accompany the above dataset we use STScI's (Space Telescope Science Institute) dataset for time-series flux of exoplanets The above mentioned dataset only contains stellar features of the celestial bodies so to determine habitability this dataset is useful. We make use of api MAST(Mikulski Archive of Space Telescope) using lightkurve package.

4. Methodology

We have two major objectives

4.1 Exoplanet detection

- Exoplanet dataset from NASA exoplanet archive.
- In addition to stellar features, time series flux values are taken STScI via lightkurve package
- Tried classifying using basic classifiers and boosting classifiers will try to use ExoNet or AstroNET for increasing efficiency.

4.2 Checking the habitability

- Use PHL's Exoplanets Catalog Data with NASA archive to predict habitable planets and handle missing data.
- Preprocess the data and oversample using SMOTE.
- Will make ExoNet model to check habitability and evaluate it.

5. Preprocessing the data

5.1 Basic preprocessing of the dataset

Kepler's exoplanet data from NASA was extracted. Dataset contains 9564 rows with 141 columns. The columns in this dataset are about stellar's features but we need time-series flux values for each star to achieve my goal. So, I used the lightcurve package to extract time-series flux data via the kepid. After gathering flux data, I got a dataset that contains 7888 rows and 1627 columns.

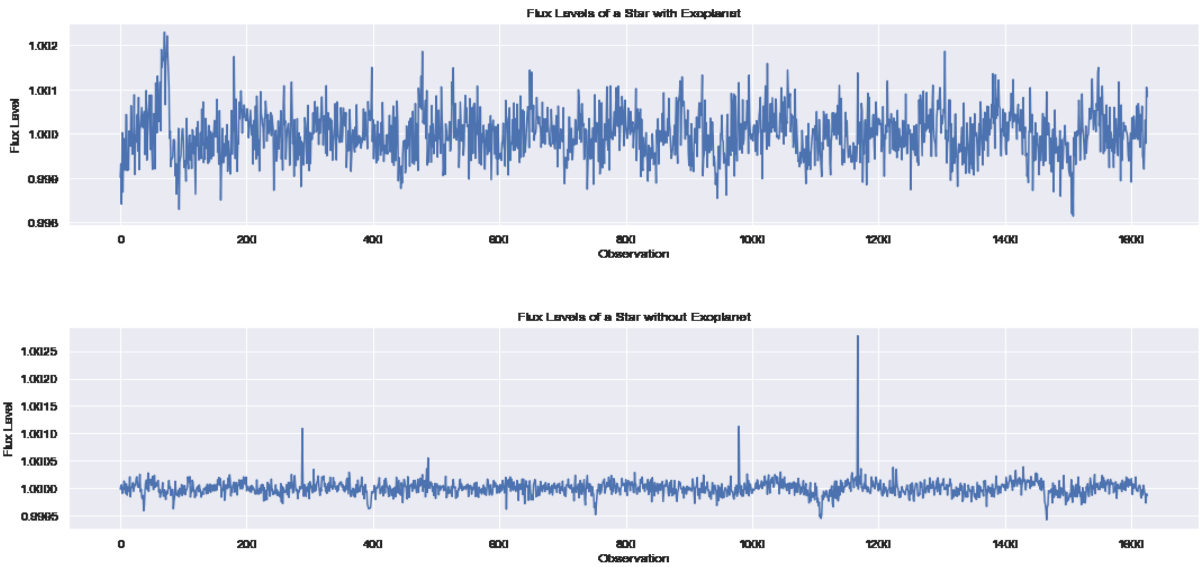


Figure 2 Flux levels of a star with or without exoplanets

Now here we have two datasets to handle: Nasa Exoplanet Archive and PHL data which contains habitability situations. These datasets contain the physical features of planets and stars. First task was to remove the missing values in the dataset.

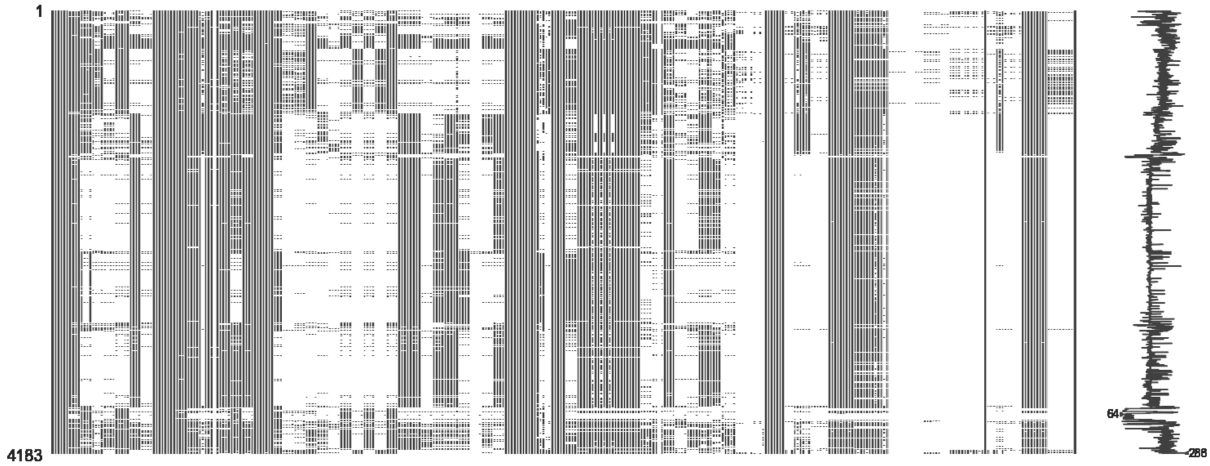


Figure 3 Missing values in the dataset

This problem can be solved by

- Removing columns with more than 40% missing data.
- Removing categorical features with more than 10 unique values.
- Iterative imputing on numeric values.

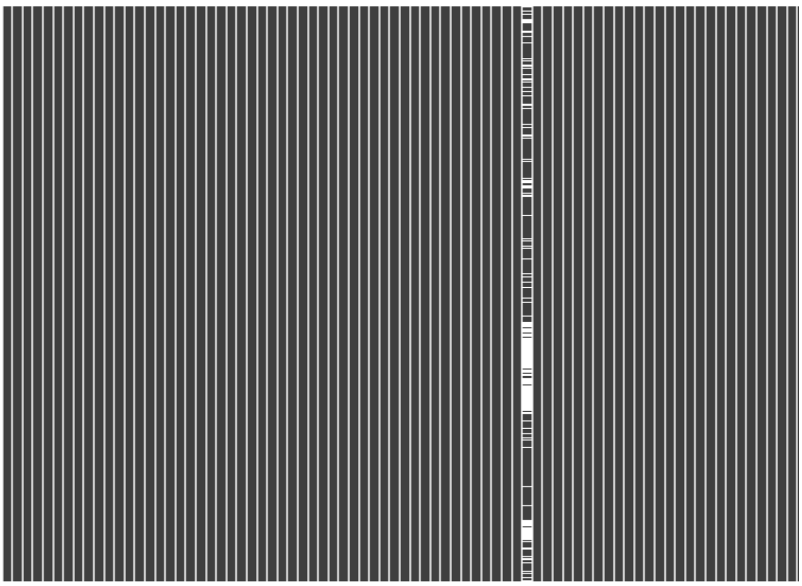


Figure 4 Removed columns with missing values and dataset ready to use

Here we have an imbalanced dataset so to solve that issue we SMOTE (Synthetic Minority Oversampling Technique) oversampling technique which is a common data augmentation method to balance the dataset.

5.2 Identifying features to be used for checking habitability

Habitability of a planet is determined by the probability of its potential habitability according to features. The planet should be at a right distance from its parent star. It must have liquid water to exist. Hence these features help to determine whether a planet is habitable or not.

Here we see the relation between mass and temperature with habitable planets from the PHL's dataset.

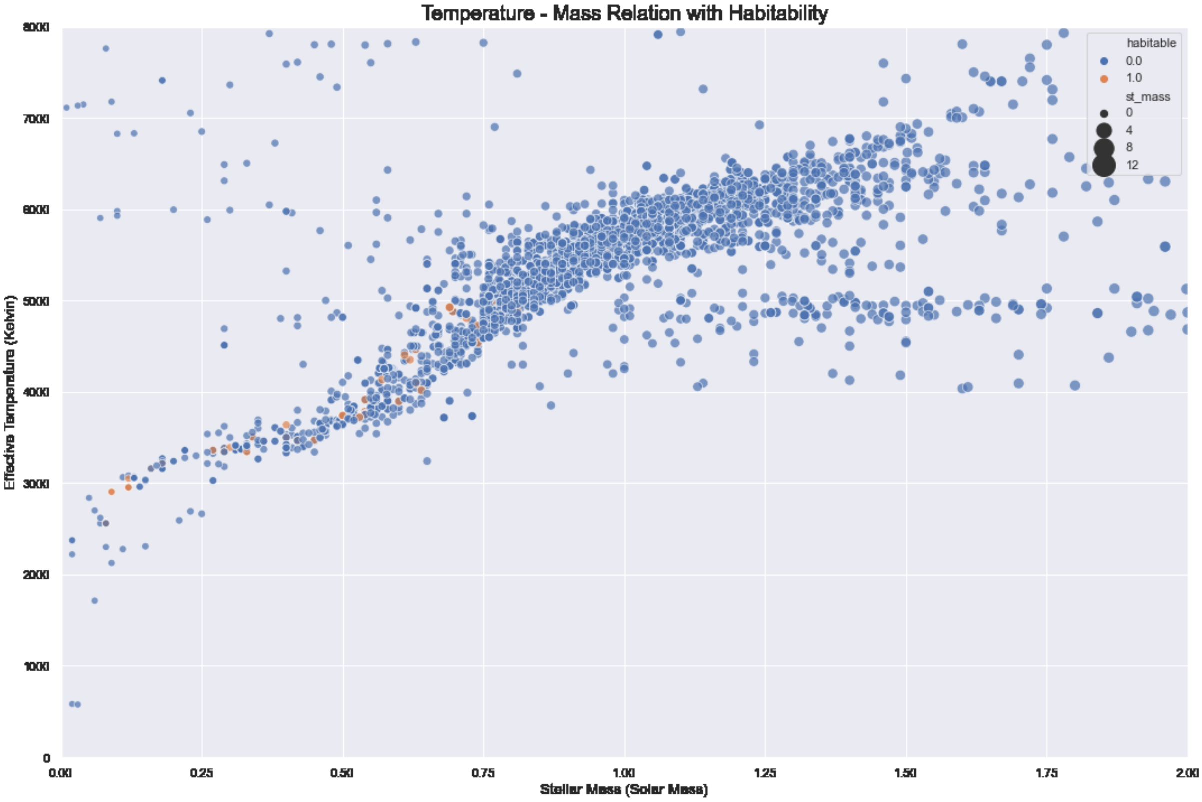


Figure 5 Relation of mass and temperature with habitable planets

Also below is the relation between Stellar Metallicity with exoplanets.

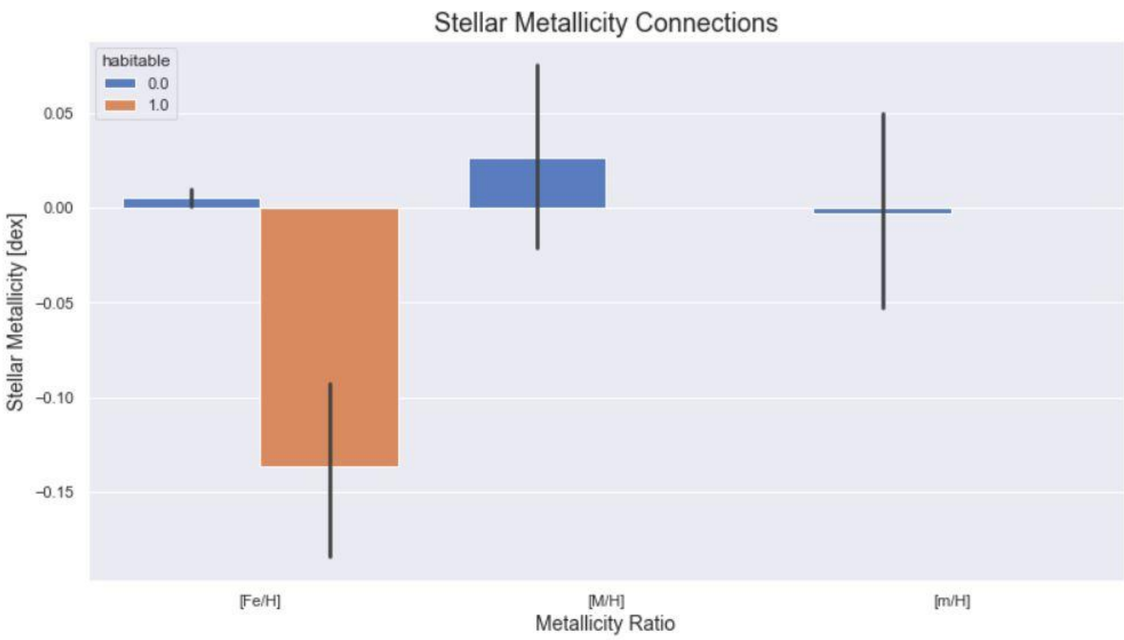


Figure 6 Relation of stellar metallicity with habitable planets

Hence after due consideration, we take various features from the dataset like the planet mass, planet’s radical velocity, equilibrium temperature, planet radius, orbital period etc.

6. Classification using basic machine learning and ExoNet

6.1 Classification using basic machine learning techniques

Here we get the data from various features of the derived datasets. We can test them with various classifier algorithms to predict the possibility of habitable planets. In the dataset the KOI’s are divided into classes like not habitable, potential and confirmed habitable planets.

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1
Bagging Classifier	98.28	66.89	0.67	0.66	0.66
Random Forest	98.97	65.97	0.66	0.65	0.65
Gradient Boosting	64.74	64.40	0.70	0.63	0.60
Logistic Regression	61.61	62.47	0.75	0.60	0.55
Linear SVC	61.58	62.32	0.73	0.60	0.55
Decision Trees	98.97	62.02	0.62	0.62	0.62
GaussianNB	61.16	61.82	0.75	0.59	0.53
KNN	75.78	60.40	0.60	0.60	0.60

Figure 7 Various machine learning techniques to classify

6.2 Classification using ExoNet

The ExoNet used here is a Convolutional Neural Network(CNN) specially made for detecting exoplanets in space. CNN is one of the most commonly used model as it has the capability of identifying complex features in the data and also has lower risk of overfitting.

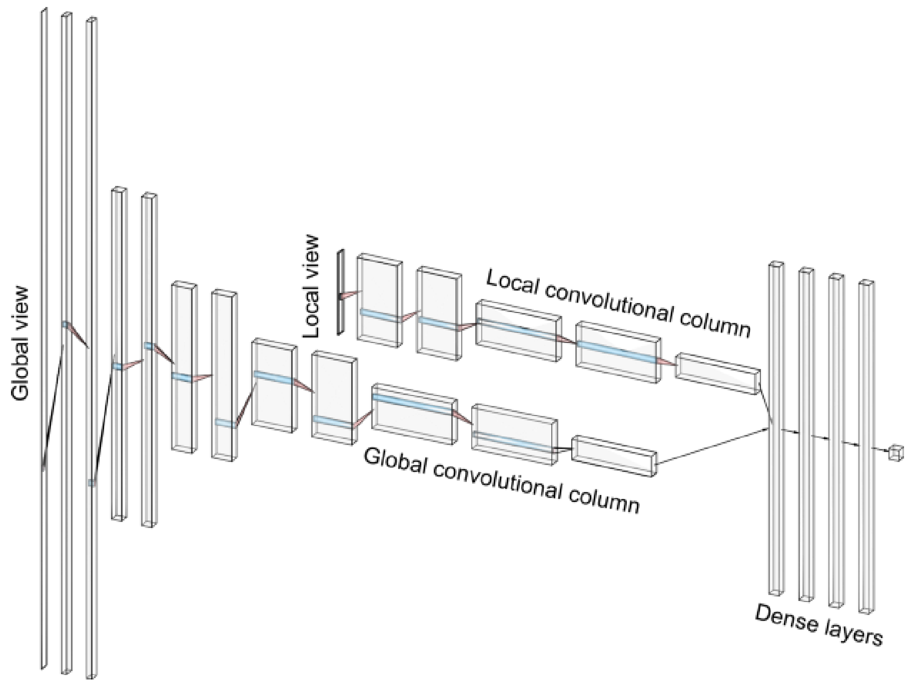
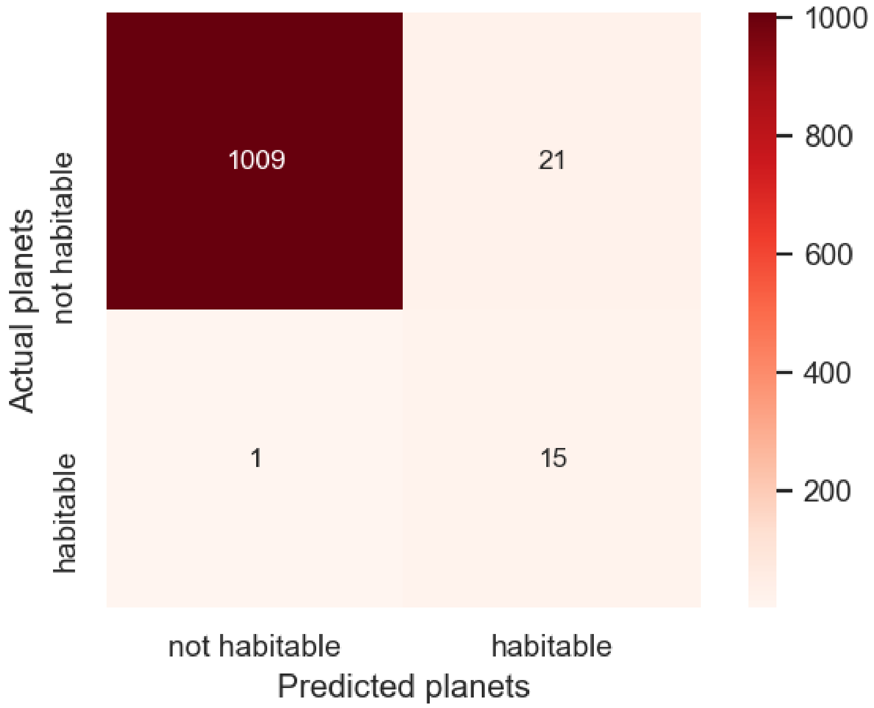


Figure 8 ExoNet model

In ExoNet two one-dimensional CNNs are incorporated, one of the views is global view and the other is local view. These extracted features from every one of the perspectives and afterward four completely associated layers go about as the classifier, doling out a probability that there is a planetary travel in the info light bend. Additionally, the organization joins dropout and batch normalisation layers to limit any overfitting. In TESS dataset this method has shown about 95% accuracy let's observe through a confusion matrix that how efficient is our model.



Here we can see the ExoNet performs well and shows 97.90 test accuracy, which is way better than common machine learning techniques

7. Summary & Conclusion

7.1 Summary

We have talked about Astronomical Data Analysis using Machine Leaning using the above algorithm. First we took our data and preprocessed it. After this point, we analyzed it and removed errors. Finally, we classified the exoplanets and determined their habitability.

7.2 Conclusion

Exoplanet Detection using Machine Learning is one of the latest and most accurate way to predict habitable planets. It has important applications where we want to separate and analyse habitable planets. This is one of the most accurate models and promises to give good results.

REFERENCES:

1) https://www.researchgate.net/publication/344173550_Astronomical_big_data_processing_using_machine_learning_A_comprehensive_review/citations
2) https://medium.com/@miguel.blanco.marcos/exonet-an-ai-saturdays-project-a1bda907bdef
3) https://github.com/alpercakr/Planet-Hunting-and-Habitability-Detection-with-Machine-Learning
4) https://medium.com/@miguel.blanco.marcos/exonet-an-ai-saturdays-project-a1bda907bdef
5) https://exoplanetarchive.ipac.caltech.edu/
6) https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi
7) https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html
8) https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html