International Journal of Research in Engineering, Science and Management
Volume-3, Issue-9, September-2020
journals.resaim.com/ijresm | ISSN (Online): 2581-5792 | RESAIM Publishing

Exoplanet Hunting in Deep Space with Machine Learning

Shivam Pratap Singh[1*], Devendra Kumar Misra[2]
[1]*Student, Department of Information Technology, HMR Institute of Technology & Management, Guru Gobind Singh Indraprastha University, New Delhi, India*
[2]*Assistant Professor, Department of Information Technology, HMR Institute of Technology & Management, Guru Gobind Singh Indraprastha University, New Delhi, India*
*Corresponding author: shivamp.singh27@yahoo.com

***Abstract*: This project focuses on the application of various machine learning algorithms on NASA's Kepler data for prediction of exoplanet habitability disposition. A comparative study of the performance of various algorithms will also be performed. The results obtained will be used to identify algorithms which are suitable for performing prediction about exoplanets. It is the need of the hour to utilize machine learning to expedite the process of exoplanet detection. This will provide greater insights in the study of planet habitability, stellar bodies and the variety of exoplanets that exist in our galaxy. As space telescopes return new data the model can be further tuned for a further improvement of accuracy. The proposed model will be able to operate on data generated by different ground and space observatories and classify exoplanet candidates as habitable or non-habitable.**

***Keywords*: Astronomy, Exoplanet, Habitability, Kepler, Machine learning, NASA, SMOTE.**

1. Introduction

The search for extraterrestrial life and habitable planets has fascinated humans for centuries. We live in an era where this search can be aided by technology to provide faster as well as accurate outcomes. The search for new habitable worlds could not be more relevant than it is today. Scientists believe that this search could result in discovery of worlds similar to our planet which have the essential conditions to support life or it could help us realize the irreplaceable nature of our planet and hence the need to conserve its resources.

The Milky Way Galaxy has about 100 to 400 billion star. Each star has potential to host Earth-sized planets at a habitable distance. These planets, which revolve around a star outside our solar system, are called exoplanets. NASA has launched several space observatories with the objective of discovering planets outside our solar system, which orbit another star. The observatories use techniques such as direct imaging, gravitational microlensing, transit method and radial velocity method to detect potential exoplanet candidates.

NASA Kepler space telescope was a space observatory launched on 7 March 2009.The purpose of Kepler was to identify Earth sized planets orbiting another star. The primary mission of Kepler ended in 2013 due to break down of its second reaction wheel. The space observatory was then given an extended mission as K2 to shift the field of view and attempt to map new portions of the sky. K2 was retired on October 30, 2018 as it ran out of fuel. It has spent 9.6 years in space, observing 530,506 stars, confirming 2,662 planets and documenting 61 supernovae [1]. The NASA Exoplanet Archive provides data collected during Kepler and K2 missions. The Planet Habitability Laboratory's Habitable Exoplanets Catalog (HEC) uses data collected by Kepler and K2 missions and other ground and space observatories to identify potentially habitable exoplanets.

Several algorithms have been devised to detect exoplanets effectively and machine learning has been used to detect whether an object of interest is a confirmed exoplanet or a false positive. However, the task of identifying whether a confirmed planet can be potentially habitable is performed by manually vetting the planetary and stellar features of each exoplanet. There are varied definitions of habitability but a general operational definition considers a planet to be habitable if has the right size and orbit to support liquid surface water. Other definitions compare the mass, radius and orbit of exoplanets to that of earth to determine whether they are potentially habitable. Habitability metrics such as Earth Similarity Index, Habitable Zone Atmosphere and Standard Primary Habitability are also used to measure the ability of exoplanet to hold a habitable atmosphere [2].

2. Literature Review

The field of astronomy consists of several data intensive problems, which can be solved by applying techniques of data science and machine learning. Machine Learning has been effectively used to extract necessary information and identify patterns in data. The paper [3], aims at understanding the effectiveness of machine learning in analysis and inference using algorithms such as K Nearest Neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naïve Bayes, Decision Tree and Random Forest on decision theoretic problems in Astronomy. They have also developed a tool kit ASTROMLSKIT and utilized it for analysis on HabCat

and supernovae data. This paper is the first instance where Habitable Exoplanets Catalog by PHL has been used for machine learning as it provides data for exoplanets observed by several space observatories.

The NASA Exoplanet Archive presents data primarily for Kepler objects of interest and the dataset contains fewer features as compared to HEC. The execution of various machine learning algorithms by the authors show promising results with accuracy as high as 98.7% for HabCat data using Random Forest and Naïve Bayes algorithm and 98.86% for supernova classification data using Decision Tree and Naïve Bayes algorithm. LDA and SVM were able to achieve the lowest accuracy of 65.9% for supernovae data due to geometric boundary constraints. The paper concludes by drawing attention to the rapid rate of exoplanet discovery and the complex task of habitability classification, which can be automated using machine learning.

In the paper [4], Hora has used machine learning to build models for prediction of habitability and has applied CART to create a regression model to predict Earth Similarity Index value.

The author uses six supervised learning algorithms-CART, Random Forest, SVM, Logistic Regression, Feed-Forward Neural Network and Naive Bayes. The CART 5 model is able to achieve the highest accuracy of 99.89%. The samples in habitable class are very low as compared to non-habitable class and the data therefore suffers from class imbalance problem. The author states that the model can be optimized further to aid in classification of exoplanets.

The paper [5] states that machine learning can be applied in the field of exoplanet research however; the existing approaches have several limitations. The author provides a novel method for habitability prediction using Gradient Boosted Regression Trees with experimental results showing 100% accuracy. The paper uses exoplanet database published by Kyoto University and the SEAU habitable definition of habitability. The SEAU habitable definition is relaxed as compared to requirements of NASA and Kopparapu habitable therefore, a greater number of samples were available under SEAU habitable definition of habitability. Thus, the author is able to avoid the problem of imbalanced classes caused by very few positive samples. The GBRT algorithm is able to perform well as it is suitable for small data sets. The features selected by the author cover both physical as well as astronomical parameters that may possibly affect habitability. The paper claims that the definition of habitability must be carefully selected before making prediction. If a strict definition is selected, the potentially habitable samples will be close to actually habitable samples but the data set for machine learning will be too small for prediction. On the other hand, if the definition is too loose the data set obtained will be large enough for convenient training and prediction. The author concludes by saying that better models can be developed if more exoplanet data is available.

In [6], several statistical techniques have been used on PHL-EC to explore classification ability of machine learning in astronomical problems. The authors have described the algorithms applied on the dataset as well as the motivation beside application of these algorithms with an effort to construct a primer in machine learning with respect to habitability detection.

A follow up of ASTROMLS KIT called ExoPlanet has also been developed during the course of research. This software allows user to select a machine learning method of choice, which would be applied on dataset, and resulting graphs would be generated after algorithm execution. The incorporation of more analysis, preprocessing and post processing techniques is still pending. The author highlights the limitation of missing 6 data in dataset as well as the high bias towards non habitable class and asserts that machine learning can be used to address the problem of habitability instead of deep learning. The task of automation of habitability detection could aid in conserving a significant portion of time which is wasted in manually studying parameters and labelling. The future scope of the paper lies in achieving an automated system which is sustainable and effectively discriminates between habitable and non-habitable exoplanet classes.

In paper [7], algorithms were applied without balancing the data at first. The results are not up to mark, such as the accuracy for SVM was 97.84%, for LDA, it was 93.23%, etc. But after balancing, the data resulted in more accurate results. Random forests gave the accuracy of 96.667%, and then decision trees gave 96.697%, while KNN gave the worst accuracy of 72.191%. Ten folded validation is done on a more balanced data such that there are more number of collection of habitable confirmed exoplanets in the data.

Maxwell et al in [8], state that machine learning can be effectively used for remotely sensed imagery classification due to the ability of algorithms to handle high dimensional data and map classes to complex characteristics. The paper discusses decision tree, support vector machines, random forest, artificial neural networks and k-nearest neighbor's algorithms for classification. Random Forest, boosted decision trees and support vector machines produce higher accuracy as compared to single decision tree and k-nearest neighbors and default parameters of algorithms result in satisfactory performance with the remote sensing dataset. However, parameter optimization results in best classification performance. The paper states that tree based algorithms are able to perform well even when the training sample size is small. Class imbalance problem can affect accuracy of rare classes significantly. Machine learning algorithms may be embedded in remote sensing tasks in the future, as they are a powerful tool for extracting information.

In paper [9], it is shown that ROC curves do not prove to be a better performance measure when the data set is strongly imbalanced. Rather, use of Precision- Recall curves is proved to provide better evaluation of performance measure.

Performance of Binary classifiers is evaluated based on specificity and sensitivity. ROC curves can illustrate misleading results and it might seem more pleasing than Positive predictive value or Precision-Recall curve, but actually, they provide wrong accuracies because it might get biased towards the more significant class present in the data set. PRC plot is proved better than ROC because of the fact that it evaluates only the fraction of true positives among all the positive predictions. Only PRC changes the ratio of positives to negatives, while curves like CROC, CC, and ROC are not capable of doing that.

Table 1
Evaluation measure of Confusion Matrix [9]

Measure	Formula
ACC	(TP + TN) / (TP + TN + FN + FP)
ERR	(FP + FN) / (TP + TN + FN + FP)
SN, TPR, REC	TP / (TP + FN)
SP	TN / (TN + FP)
FPR	FP / (TN + FP)
PREC, PPV	TP / (TP + FP)
MCC	$(TP * TN—FP * FN) / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))^{1/2}$
$F_{0.5}$	1.5 * PREC * REC / (0.25 * PREC + REC)
F_1	2 * PREC * REC / (PREC + REC)
F_2	5 * PREC * REC / (4 * PREC + REC)

ACC: accuracy; ERR: error rate; SN: sensitivity; TPR: true positive rate; REC: recall; SP: specificity; FPR: false positive rate; PREC: precision; PPV: positive predictive value; MCC: Matthews correlation coefficient; F: F score; TP: true positives; TN: true negatives; FP: false positives; FN: false negatives

3. Methodology

The dataset used in this project was obtained from National Aeronautics and Space Administration, "Kepler and K2". This catalog provides data on exoplanets collected by ground and space observatories. The workflow of project is shown in the figure 1.

Data preparation is of utmost importance before application of any machine-learning algorithm. Data preprocessing involves construction and transformation of dataset so that machine-learning algorithms can be applied to understand accurate patterns in the data. Every machine-learning problem focuses on prioritizing improvement of the quality and size of dataset.

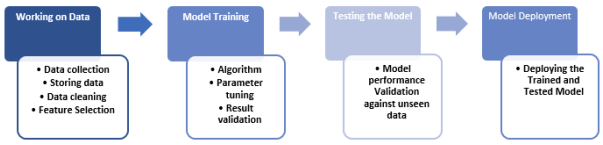

Fig. 1. Working model

1. Import all the libraries necessary for the execution of the project, i.e. Numpy, pandas, sktlearn, matplotlib.pyplot, etc.
2. Import warnings to filter out or ignore the unnecessary warning messages while execution.
3. Load the csv file and after that find out the total number of null values in each column. Delete all the columns whose null values are more than 50% of its data, and use remaining columns for the data pre-processing.
4. Cleaning the data is done. Firstly, take out the datatypes of all the columns. Check datatypes of the attributes and correct those datatypes if necessary.
5. Change the datatypes of those columns, which are object to category so it will be easier to work on those columns as well.
6. Now, duplicates are handled. Here, duplicates mean those rows which have same values for every column. Though in the data that we used, the number of duplicate rows was just 1. After handling the duplicates, categorical data is handles. The categorical data can be handled on many techniques.
7. Missing values were imputed by replacing missing values with mean value of column data.
8. A function is defined which is used with every machine learning algorithm we used in our project. This function includes Accuracy score, precision score, recall score, f1 score, confusion matrix, classification report, training score, and testing score.
9. The data is scaled using Min Max Scaler to transform all values between 0 and 1 and standardized using Standard Scaler to transform data having mean=0 and standard deviation=1.
10. The dataset was imbalanced as minority class was only 0.7% of the total data. Class Imbalance was handled using oversampling, under sampling and SMOTE technique.

4. Data Preprocessing and Normalization

Data preprocessing is an important step in Machine Learning as the quality of data used and the information required for that can be gathered from it directly affects the capability of our model to learn; hence, it is very important that we preprocess our data before providing it into our machine learning model.

In machine learning, Data Normalization is a method, which is mostly used as part of preparation of data. By changing the values of numeric columns in the dataset to a common scale, without influencing the differences in the ranges of values can be done by using the method of data normalization.

A. Data normalization

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.

B. PCA (Principal Component Analysis)

Need to decrease the number of features (Dimensionality Reduction) to remove the possibility of Curse of Dimensionality.

C. SMOTE (Synthetic Minority Over-Sampling Technique)

It is an over-sampling method. It creates synthetic (not duplicate) samples of the minority class. Hence making the minority class equal to the majority class.

5. Classification

In Machine learning, mainly there are three primary

categories of machine learning solutions: regression, clustering, and classification. Classification is a way of supervised machine learning in which observations are given a known class value based upon their explanatory variables. Binary or multi-class values can be used in classification. This model focuses on a binary classification of objects of interest as "FALSE POSITIVE" or "CONFIRMED" exoplanets. NASA to indicate the satellite tracked an object of interest incorrectly uses the classification of "FALSE POSITIVE". The meaning of the term in machine learning classification terms is a bit different.

For the majority class, classifying imbalanced data is biased. This bias is even greater for high-dimensional data, in which the number of variables greatly exceeds the number of samples. The problem of biasing can be fixed by using the method of under sampling or oversampling, which provides class-balanced data. Usually method of under sampling is useful, whereas random oversampling is not. Synthetic Minority Oversampling Technique (SMOTE) is a much known oversampling method that was proposed for the random oversampling improvisation but its working on high-dimensional data has not been thoroughly tested.

SMOTE is helpful for k-NN classifiers for high-dimensional data in case the number of variables is reduced performing some type of variable selection; otherwise, the k-NN classification is biased for the minority class.

6. Results and Conclusion

Supervised learning algorithms were implemented using Python 3.7.0 and the performance of the algorithms is summarized below.

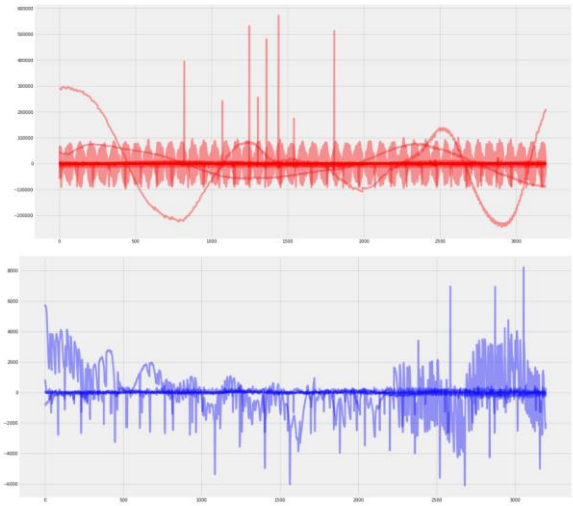

Fig. 2. Light intensity of the exoplanets and non-exoplanets

The receiver operating characteristic curve is plotted and areas under the ROC curve is computed.

Following performance measures are used to differentiate among the entire algorithms used.

Accuracy: It is computed using the formula,

(True Positive+True Negative)/Total Population.

It is an intuitive performance measure which indicates the ratio of correctly predicted observations to total number of observations.

ROC Curve: It stands for Receiver Operating Characteristic curve. The curve is used to determine the performance of model under all classification thresholds. It is plotted using True Positive rate and False Positive rate. True positive rate is the ration of true positives to total no. of positives in dataset. True negative is the ration of false positives to the total no. of negatives. The area under the ROC Curve is used to evaluate the goodness of model. The area under ROC curve (AUC) is scale invariant and classification threshold invariant.

Table 2
Performance of the algorithms

Algorithm	Accuracy	F1_Score	Confusion Matrix
Logistic Regression	0.9745	0.9744	[442 23] [0 438]
KNN(n=3)	0.9922	0.99207	[458 7] [0 438]
Naïve Bayes	0.9579	0.95768	[435 30] [8 430]
Decision Tree	0.9944	0.99432	[460 5] [0 438]
Random Forest	0.9867	0.98648	[453 12] [0 438]
K- means Classifier	0.602	0.63981	[356 538] [0 478]

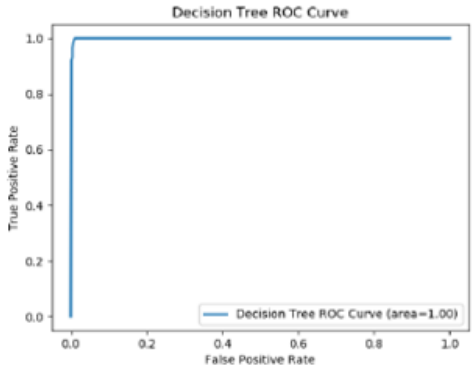

Fig. 3. ROC curve for Decision Tree

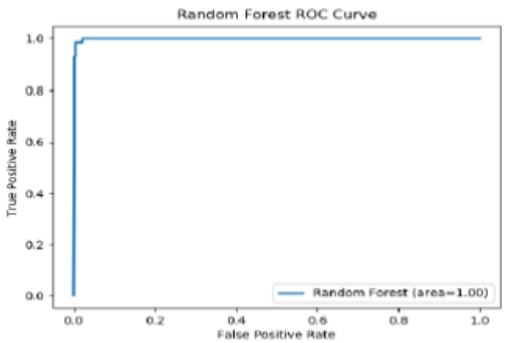

Fig. 4. ROC curve for Random Forest algorithm

F1-Score: It is computed using the formula,
(2*Precision*Recall)/(Precision+Recall).

Here Precision=True Positive/Total Predicted Positive and Recall is True Positive/Total Actual Positive.F1 Score is a function used to measure a balance between precision and recall which provides an indication of the efficiency of model when classes are imbalanced.

Confusion Matrix: It is a measure to evaluate a model's performance which shows the number of true positives, false positives, false negatives and true negatives in the test data. True Positives are the values in test data which were positive and were correctly predicted by model. Similarly, True negatives are values which were negative and were correctly predicted as negative by model. False positives are the values which were negative but were predicted as positive by the model. False negatives are the positive values which were incorrectly predicted as negative.

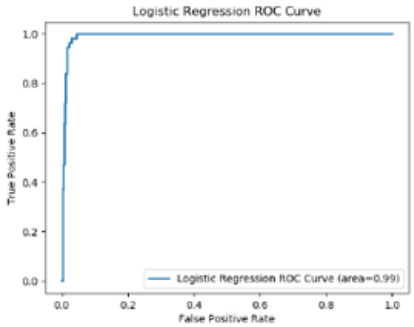

Fig. 5. ROC curve for Logistic regression

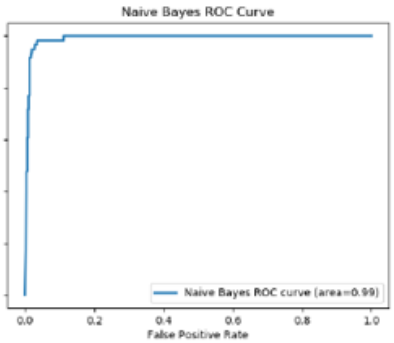

Fig. 6. ROC curve for KNN

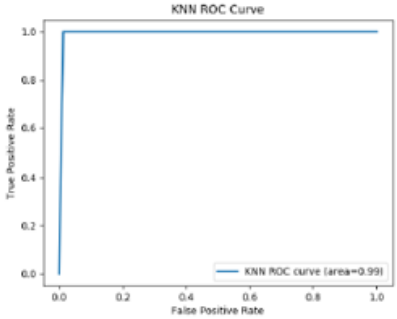

Fig. 7. ROC curve for Naïve bayes

The objective of identification of a suitable machine learning algorithm for predicting habitability of exoplanet was achieved. This project involved the implementation of supervised machine learning algorithms which included Logistic Regression, K Nearest Neighbor, Naive Bayes, Decision Tree and Forest. The performance of each of these algorithms was analysed using the performance metrics accuracy, f1 score, confusion matrix and ROC curve. The research problem addressed in this study was a binary classification problem of classifying a celestial body as habitable or non-habitable. The evaluation of the implemented supervised learning algorithms on the basis of the above mentioned performance metrics indicate Decision Tree algorithm as a suitable model for predicting habitability. Decision Tree is a robust machine learning algorithm. It has the following advantages:

1. This algorithm is able to implicitly perform feature selection in analytics. Hence it was able to extract the necessary features for learning from the 15 features which were supplied.
2. The algorithm does not show sensitivity towards outliers and splitting at each decision node takes place on the basis of the proportion of samples within the split ranges.
3. Decision trees are not affected by nonlinear relationship among parameters. Regression algorithms suffer from failing checks when the relationship between variables is nonlinear. The parameters in our research problem had high nonlinearity which did not affect the model.
4. Non-parametric method applied by decision trees makes no assumptions about the structure of classification or the space distribution. Therefore, the model shows a better performance in comparison to its counterparts.

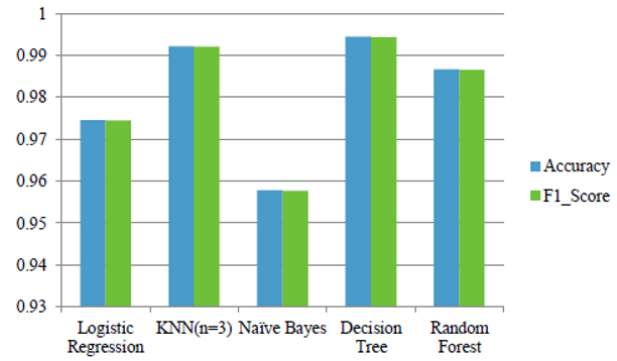

Fig. 8. Graphical representation of the above mentioned table

The K nearest neighbours' algorithm also provides satisfactory performance for the same research problem. This algorithm does not create any model for analyzing data but relies on storing the data points. It then identifies the nearest neighbors of the unknown data point and predicts its class. The consistency of training data is very important for the effective prediction by this algorithm. Since our dataset did not contain any erroneous data and maintained its consistency, KNN was also able to produce accurate predictions.

The ROC curve of decision tree and KNN algorithm are close to an ideal plot with provide evidence of suitability of these algorithms. The data which is generated by Kepler and other space observatories and crafts can now be used to predict habitability by these automated machine learning algorithms rather than manually examining the plethora of data.

7. Future Scope

In the future, huge space agencies will provide detailed maps of planets. Probably, interstellar probes will be launched towards the nearest exoplanets to take close- up images.

Exoplanets data would be present in large number in future. Machine learning would assist us in gathering the existing knowledge about habitability via machine learning model and using it on new data to make the list of habitable planets. More accurate model would be possible when more training data about habitable planets become available.

Space exploration involves significant investment in terms of time, effort and funding. It is imperative that this expenditure is made to provide better understanding and advancement in the fields of science and technology. The application of machine learning models to classify exoplanets will provide assurance that the planets being explored are highly likely to be habitable. The space industry is no longer restricted to a few nations but has witnessed an expansion to not only several nations of the world but privately owned companies such as SpaceX, Virgin Galactic, Orion Span, Blue Origin as well. This has led to a surge in investments to fund commercial space travel and make space tourism a reality.

Machine Learning provides new hope to identify habitable worlds and analyze plethora of complex data. It offers tremendous help in classification with numerous complex inter-connected parameters to identify which exoplanets are favorable candidates to be potentially habitable. The exploration and study of habitable exoplanets can enhance our understanding and enable us to discover new properties of exoplanets.

The scalability, time efficiency and cost effectiveness of this approach can improve the rate of discovery of habitable worlds and support extensive exploration not only in the area of habitability but in other areas of the data intensive field of Astronomy as well. The groundbreaking research that is driven by Artificial Intelligence and Machine Learning can act as impetus for pioneering discovery and development.

8. Conclusion

The process of identifying the habitability disposition of exoplanets based on data collected by observatories can be extremely cumbersome if performed manually. Machine learning can aid in solving this problem by automating the task of habitability classification. The objective of identification of a machine-learning model for predicting habitability of exoplanets has been achieved. The most effective technique to handle class imbalance problem was Synthetic Minority Oversampling Technique (SMOTE). This technique produced promising results as compared to Random Oversampling and Random Under sampling. The machine-learning model was able to train and generalize well after increasing the minority class samples by synthetically generating new samples. Cost sensitive learning was also performed using Support Vector Classifier however the results were similar to those obtained after applying SMOTE. Accuracy is not a useful metric for imbalanced data hence precision, recall, f1 score and balanced accuracy were computed and ROC curves and Precision Recall curves were plotted for analyzing performance Random Forest was identified as suitable model for classifying exoplanets as potentially habitable or non-habitable. This model showed favorable performance on execution without using any class imbalance handling technique by obtaining f1 score of 95.24% and area under Precision Recall curve value of 0.993. After using SMOTE to handle class imbalance Random Forest obtained f1 score of 90% and area under Precision Recall curve equal to 0.975.

References

[1] National Aeronautics and Space Administration, “Kepler and K2”, https://www.nasa.gov/mission_pages/kepler/main/index.htm
[2] PHL HEC (Habitable Exoplanets Catalog), http://phl.upr.edu/projects/habitableexoplanets-catalog/data/database.
[3] Saha, S., Bora, K., Agrawal, S., Routh, S., & Narasimhamurthy, A. (2015). “ASTROMLSKIT: A New Statistical Machine Learning Toolkit: A Platform for Data Analytics in Astronomy,” (No. arXiv: 1504.07865).
[4] Hora, K. (2018). “Classifying Exoplanets as Potentially Habitable Using Machine Learning,” in ICT Based Innovations, pp. 203-212, Springer, Singapore.
[5] Zhu Weijun, and Wang Xin. “Predicting the Habitability of Exoplanets based on GBRT Algorithm.”
[6] Saha, S., Basak, S., Safonova, M., Bora, K., Agrawal, S., Sarkar, P., & Murthy, J. (2018). “Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets. Astronomy and Computing”, vol. 23, 141-150.
[7] Agrawal, S., Basak, S., Saha, S., Rosario-Franco, M., Routh, S., Bora, K., & Theophilus, A. J. (2015), “A Comparative Study in Classification Methods of Exoplanets: Machine Learning Exploration via Mining and Automatic Labeling of the Habitability Catalog”.
[8] Maxwell, A. E., Warner, T. A., & Fang, F. (2018), “Implementation of machine learning classification in remote sensing: An applied review”, International Journal of Remote Sensing, 39(9), 2784-2817.
[9] Saito, T., & Rehmsmeier, M, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. PloS one, 10(3), 2015.