# Exoplanet Classification

Shaunak Pal
shaunak18098@iiitd.ac.in

Abhinav Suresh Ennazhiyil
abhinav18003@iiitd.ac.in

Rajat Prakash
rajat18078@iiitd.ac.in

## Abstract

*Existing work on characterizing exoplanets are based on assigning habitability scores to each planet which allows for a quantitative comparison with Earth.*
*Over the past two decades, discoveries of exoplanets have poured in by the hundreds and the rate at which exoplanets are being discovered is increasing the actual number of planets exceeding the number of stars in our galaxy by orders of magnitude.*
*The research is based on classifying exoplanets as Habitable and Non-Habitable. The research uses the dataset provided by NASA's Exoplanet Archives which is from the TESS satellite. We tried various preprocessing techniques and classifiers along with it such as KNN (K-Nearest Neighbors), Hard-Boundary , SVM (Support Vector Machines) and, Tree based classifiers like Random Forests and ensemble classifiers like XGBoost to thoroughly analyze what works best for this domain. The trained models can be referred to [here].*

## 1. Introduction

Wondering about the existence of life outside the solar system has led humans to unravel many secrets of the universe. Launched in April 2018, TESS is surveying the sky for two years to find transiting exoplanets around the brightest stars near Earth. Exoplanets are divided into two category, Habitable and Non-Habitable. The research includes observations and results from multiple classification techniques to classify whether the planet is habitable or non-habitable.

## 2. Literature Survey : Suryoday Basak

### 2.1. Objective

Artificial Undersampling and KDE using Parzen Window Estimation is used on (All features, Mass and Radius) with different family of classifiers to classify exoplanets into thermal habitability and characterize them based on potential habitability. (PHL-EC dataset [3])

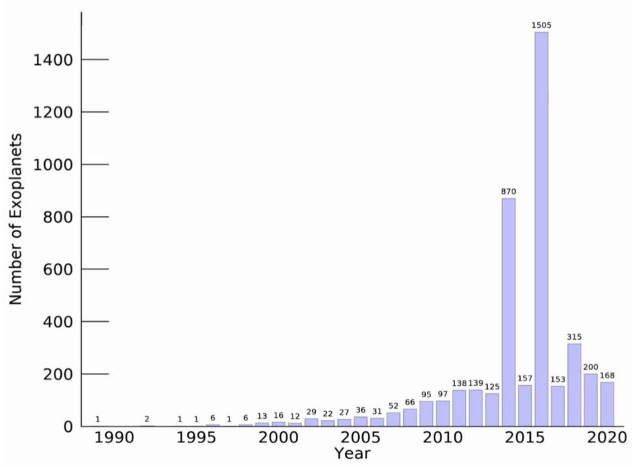


Figure 1: Planet Discovery(years) Vs Count

### 2.2. Features Selection

Stellar data from the Hipparcos catalog the PHL-EC dataset [3]. It consists of a total of 68 features (of which 13 are categorical and 55 are continuous valued) and more than 3800 confirmed exoplanets. The dataset consists of observed and derived attributes. Atmospheric type, mass, radius, surface temperature, escape velocity, earth's similarity index, flux, orbital velocity are the important features.

### 2.3. Prediction Models

Gaussian Naive Bayes evaluates the classification labels based on class-conditional probabilities with class a priori probabilities while GNB works on the assumptions that the features are independent of each other (Gaussian Distribution). The k-nearest neighbor classifier used where the distance between the neighbors in the input space is used as a measure for categorization . k = 3 while the weights are assigned uniform values. Creating hard-boundaries i.e. n-dimensional Hyper-planes. SVM without a kernel is and with a radial basis kernel tried. The parameters setup for linear discriminant analysis classifier was implemented by the decomposition strategy similar to SVM . No shrinkage metric was specified and no class prior probabilities were

assigned. The discrimination which leads to the best value of gain is used to develop a rule; should the rule not result in a perfect discrimination of the data , then the criterion is recursively applied to the partitions in the data created by the previous rule. Random forest classifier is used with splitting criteria as Gini and elastic Gini. (augmented variable from top 85% features based on their importance determined) [2]

## 3. Literature Survey : George, Brychan, Sohail

### 3.1. Objective

Out of over 9500 objects discovered by Kepler Satellite , 2000 are confirmed exoplanets. Kepler satellite can detect star brightness with high accuracy. It is cumbersome task to detect whether the exoplanet is in HZ(Habitable Zone) or not. Model is trained to predict the HZ.

### 3.2. Features Selection

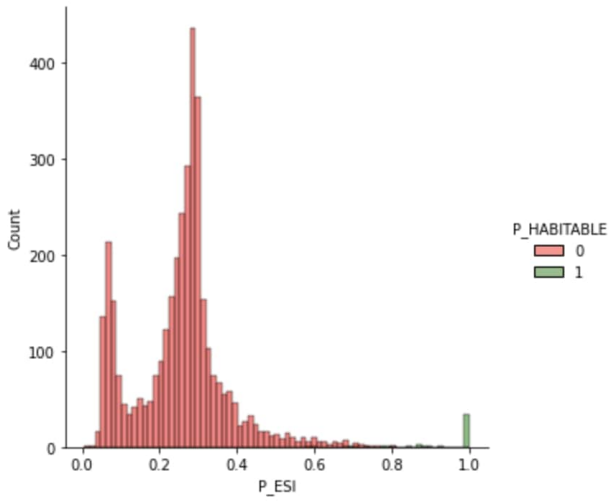


Figure 2: Importance of Earth Similarity Index

Many features are dropped because they are ( All Zeroes, Leakage, Free form text, Zero variance or Duplicate) . Koi_prad (Planet Radius) , koi_dicco_msky (Angular Offset), koi_fpflag_nt (Transit), koi_fpflag_ss (Transit), koi_fpflag_ec (Similarity to confirmed exoplanets) are in descending order of highest importance in Random Forest classification providing our final model.

### 3.3. Prediction Models

The research paper used following Prediction models KNN, SVM, and random forest are selected Methodologies as they are naturally robust to high dimensionality.
(Metric - Accuracy, Precision, Recall : A, P, R)

SVM Did not produce expected prediction results in terms of the proportion of observations classified as exoplanets. Various (feature,parameter) combinations with SVM. A, P, R = 0.9681, 0.9309, 0.973 .

The KNN Training performance and Prediction performance are quite good. A, P, R = 0.9371, 0.854, 0.9704

Random Forest Produces scaled feature importance values which clearly show the most and least important features in classifying data. A, P, R = 0.9896 0.9955 0.9721.

All models performed well well, Random Forest showing more superior results. Following is F1 score respectively F1 (0.9515, 0.9085, 0.9837) which favours Random Forest. [1]

## 4. Dataset and Preprocessing

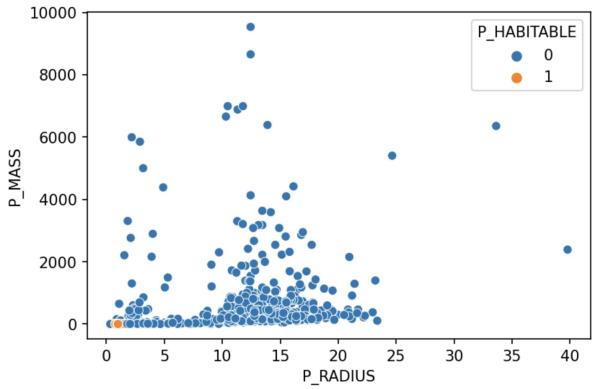


Figure 3: Planet Radius vs Planet Mass

The original dataset [3] consists of 112 features. Some of the important features are planet mass, radius, surface temperature, ESI (Earth Similarity index), flux, orbital velocity, escape velocity, etc. The features starting with "P_" indicates it's about the planet while "S_" is about the star.

The research will use these features to predict whether the planet is habitable (P_Habitable=1,2) or not (P_Habitable=0). P_Habitable has values 0,1,2 where 1 and 2 represent optimistic and conservative habitable zones respectively, while 0 means the planet is non-habitable. The research will convert this into a binary variable and train our model based on this. The description and visualization of some important features are shown in the following figures.

As seen in Figure 2, life favourable conditions are harder to maintain with larger radius or mass, as it leads to very high gravity, thick atmosphere and higher atmospheric pressure. Hence planet density plays an important role. At the outset, the research removes the columns that are irrelevant for the classification such as the names of the planets, the mode of discovery, the errors in the measurements, etc. Then since many columns have a majority of missing columns, we drop them as they would provide little insight to the model. The

remaining missing values we filled using the mean (for numerical attributes) and the mode (for categorical attributes). We dealt with the categorical data in 2 ways. The ordinal features i.e, features which had some kind of order but objects where encoded using Label Encoding. Rest of the features without any order were one hot encoded to get the best out of the features. The data was then scaled to prevent the features with astronomical values from overshadowing the features with microscopic values. This is a must for distance and probability based classifiers like K Nearest Neighbours or Naive Bayes. The data-set is highly imbalanced as well, and to combat that we used SMOTE for artificially oversampling the minority class. This will make sure the model does not return negative for all the planets, and instead give a much better F1-score will results in a better model.

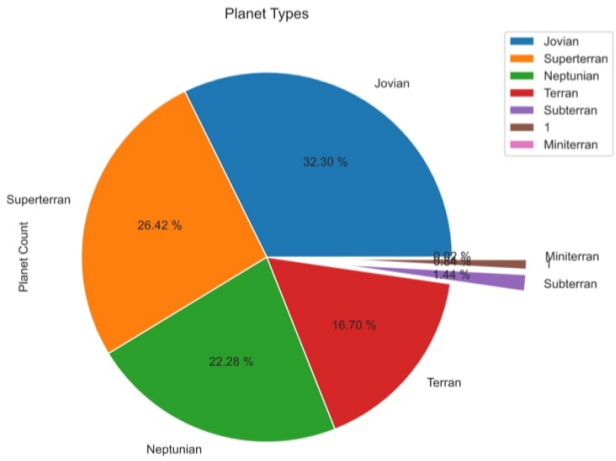


Figure 4: Planet Types

## 5. Methodology

Before training the models, a stratified train-test split was done in the ratio 15:85 for testing the models and prevents an unfortunate split, and comparing them based on metrics like f1-score, precision, recall etc. Since the output is a binary label (0 is Non-Habitable and 1 is Habitable), classification algorithms were used which include Instance-Based classifiers like KNN (K-Nearest Neighbors), Hard-Boundary classifiers like SVM (Support Vector Machines), Tree based classifiers like Random Forests and ensemble classifiers like XGBoost.

To find the optimal parameters, grid search was used with cross validation on 5 folds. This helps in averaging out the score to give a better understanding of how well the model is performing. F1 scoring was used for evaluating models keeping class imbalance in mind.

For KNN the parameters were ('algorithm': 'auto', 'leaf_size': 2, 'n_jobs': -1, 'n_neighbors': 1, 'p': 1, 'weights': 'uniform'). The algorithm was chosen based on the train set. '1' for the value of 'p' indicates that the Minkowski distance is preferred for this problem.

The optimal parameters for the Support Vector Machine were ('C': 0.1, 'gamma': 0.01, 'kernel': 'linear', 'random_state': 42). C=1.0 shows regularization is not needed for the model and the Radial Basis Function is the most appropriate for this data-set.

For random forest the optimal parameters were quite complicated as follows: ('bootstrap': True, 'criterion': 'gini', 'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100, 'oob_score': True). using bootstrap sampling and its score improved the score a bit which can help when the model is low on data and to reduce variance in such imbalanced class problems. Regardless it did not perform as impressively as the other classifiers.

For the XGBoost Classifier, the optimal parameters were ('booster': 'gbtree', 'gamma': 0.0001, 'learning_rate': 1, 'max_depth': 2, 'n_estimators': 100, 'subsample': 0.85). This shows the XGBoost model is very simple as the max_depth required is only 2. XGBoost and RandomForest, both tree based ensemble techniques did not turn out well for exoplanet classification, which could be due to the fact that it assumes that the decision boundaries are always parallel to the axes while that might not be the case actually.

Logisitic regression did not perform very well when using all the features available and reason is curse of dimensionality. When we used the top 50 most important features as given by our best model, there was a big spike in performance with an f1 score of around 0.95, which shows how curse of dimensionality could effect logistic regression. The best parameters where ('C': 0.001, 'class_weight': 'balanced', 'dual': False, 'max_iter': 1000, 'penalty': 'l2', 'solver': 'lbfgs'). Balanced weights being the most important feature for better results.

For Gaussian Naive Bayes the following parameters worked best: ('priors': [0.1, 0.9], 'var_smoothing': 0.000533). The prior was adjusted to give maximum weightage to the minority class. the var_smoothing parameter was adjusted using a linear space to add variances for calculation stability

The feature importance measure for all the models show that star temperature is the most contributing attribute. Other important factors are planet equilibrium temperature, planet flux, star distance, etc. The following graph shows the important features along with their scores. This will give researchers more insight on how these factors affect the habitability of a planet.

## 6. Results

The missing values, imbalanced data and outliers were dealt with as described previously. Initially there were a

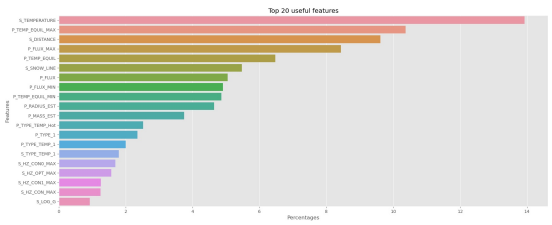


Figure 5: Feature Importance

| Random Forest | | | | |
|---|---|---|---|---|
| Metric | Prec. | Recall | F1 | Support |
| 0 | 0.999705 | 0.997354 | 0.998529 | 3402.0 |
| 1 | 0.808511 | 0.974359 | 0.883721 | 39.0 |
| Accuracy | 0.997094 | 0.997094 | 0.997094 | 0.997094 |
| Macro avg. | 0.904108 | 0.985857 | 0.941125 | 3441.0 |
| Wt. avg. | 0.997538 | 0.997094 | 0.997227 | 3441.0 |

Table 1: Classification Report of Random Forest

| SVM | | | | |
|---|---|---|---|---|
| Metric | Prec. | Recall | F1 | Support |
| 0 | 1.000000 | 0.999705 | 0.999853 | 3395.0 |
| 1 | 0.978723 | 1.000000 | 0.989247 | 46.0 |
| Accuracy | 0.999709 | 0.999709 | 0.999709 | 0.999709 |
| Macro avg. | 0.989362 | 0.999853 | 0.994550 | 3441.0 |
| Wt. avg. | 0.999716 | 0.999709 | 0.999711 | 3441.0 |

Table 2: Classification Report of SVM

| KNN | | | | |
|---|---|---|---|---|
| Metric | Prec. | Recall | F1 | Support |
| 0 | 0.999411 | 0.998822 | 0.999116 | 3396.0 |
| 1 | 0.914894 | 0.955556 | 0.934783 | 45.0 |
| Accuracy | 0.998256 | 0.998256 | 0.998256 | 0.998256 |
| Macro avg. | 0.957152 | 0.977189 | 0.966949 | 3441.0 |
| Wt. avg. | 0.998305 | 0.998256 | 0.998275 | 3441.0 |

Table 3: Classification Report of K Nearest Neighbors

| Logistic Regression | | | | |
|---|---|---|---|---|
| Metric | Prec. | Recall | F1 | Support |
| 0 | 0.999411 | 0.999411 | 0.999411 | 3394.0 |
| 1 | 0.957447 | 0.957447 | 0.957447 | 47.0 |
| Accuracy | 0.998838 | 0.998838 | 0.998838 | 0.998838 |
| Macro avg. | 0.978429 | 0.978429 | 0.978429 | 3441.0 |
| Wt. avg. | 0.998838 | 0.998838 | 0.998838 | 3441.0 |

Table 4: Classification Report of Logistic Regression

| XGBoost | | | | |
|---|---|---|---|---|
| Metric | Prec. | Recall | F1 | Support |
| 0 | 0.997938 | 0.995884 | 0.996909 | 3401.0 |
| 1 | 0.702128 | 0.825000 | 0.758621 | 40.0 |
| Accuracy | 0.993897 | 0.993897 | 0.993897 | 0.993897 |
| Macro avg. | 0.850033 | 0.910442 | 0.877765 | 3441.0 |
| Wt. avg. | 0.994499 | 0.993897 | 0.994139 | 3411.0 |

Table 5: Classification Report of XGBoost

| Gaussian Naive Bayes | | | | |
|---|---|---|---|---|
| Metric | Prec. | Recall | F1 | Support |
| 0 | 1.000000 | 0.999117 | 0.999558 | 3397.0 |
| 1 | 0.936170 | 1.000000 | 0.967033 | 44.0 |
| Accuracy | 0.999128 | 0.999128 | 0.999128 | 0.999128 |
| Macro avg. | 0.968085 | 0.999558 | 0.983296 | 3441.0 |
| Wt. avg. | 0.999184 | 0.999128 | 0.999142 | 3441.0 |

Table 6: Classification Report of Gaussian Naive Bayes

total of 112 features. After removing irrelevant and NaN dominant features (where more than half of the rows were NaN), finally we have 61 features. The categorical features dealt with in 2 ways. The ordinal features were encoded using Label Encoding while features which had no specific order were then encoded to one hot encoding which proved to be better than doing just one way. We scaled the data to prevent features with huge values from overshadowing the smaller ones and thus make every feature comparable. Classifiers work best when the output classes are equally distributed, and of all balancing techniques, SMOTE Tomek worked the best.

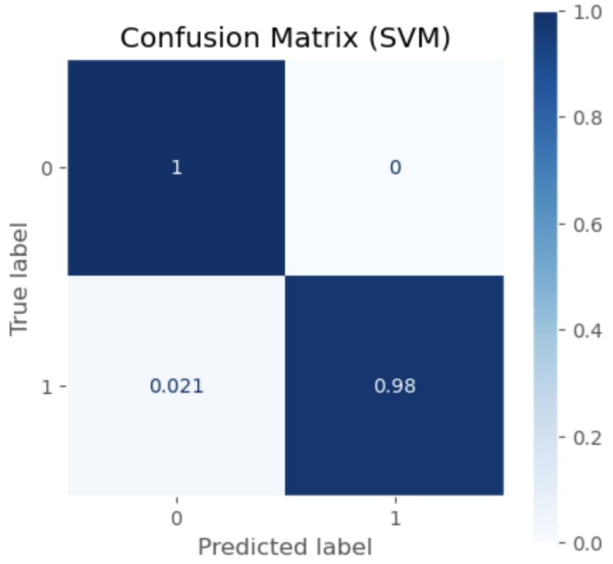


Figure 6: Confusion Matrix

A train test split of 15:85 was made to test our model to their limits. With such low training data and large testing set, the SVM could achieve an f1 score of 98.9% with just 1 planet misclassified. Tree based classifiers and ensemble techniques on the other hand did not perform so well. Reasons could be such as they assume the decision boundaries to be parallel to the axes. This may not be the case in real world and results could change a bit in real world but we have tried our best to simulate real world scenario by keeping test set size as high as possible. It does prove that machine learning could do wonders in the field of exoplanet classification.

## 7. Conclusion

Exoplanet Classification is a problem which will always remain a highly imbalanced classification problem with majority of the planets being discovered being uninhabitable. Hence dealing with Imbalanced classes like using Synthetic Minority Oversampling, Stratified splits etc is a must to get better predictions, and scaling is necessary since certain features can have astronomic values while some can have microscopic values.

Almost all the models which were tested have 90%+ accuracy with Earth Similarity Index as the most important feature. After doing Grid search over a lot of hyper parameters for a lot of models we concluded that support vector machines perform the best with f1 scores close to 98.9%, whereas tree based algorithms did not fare that well with f1 scores just below 90% which is not that bad and we made sure that at least all habitable planets get classified rightly.

This shows that this problem statement can easily be solved by machine learning methods and can give astronomers a huge insight into how to classify the habitable exoplanets comfortably.

## 8. Contributions

- **Shaunak Pal:** Exploratory Data Analysis, Data Preprocessing, Data Visualisation, Training and testing models
- **Abhinav S.E.:** Literature Review, Data Preprocessing, Exploratory Data Analysis, Training and testing models
- **Rajat Prakash:** Literature Review, Exploratory Data Analysis, Report writing

## References

[1] S. R. George Clayton Sturrock, Brychan Manry. Machine Learning Pipeline for Exoplanet Classification, 2019.

[2] S. S. A. J. T. K. B. G. D. J. M. Suryoday Basak, Surbhi Agrawal. Habitability Classification of Exoplanets: A Machine Learning Insight, 2018.

[3] N. University Of Puerto Rico. PHL's Exoplanet Catalog of the Planetary Habitability Laboratory. 2020.