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.

1. Introduction

Wondering 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])

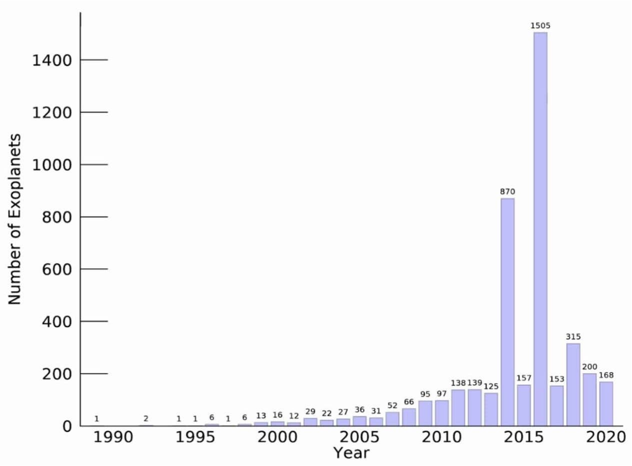

Figure 1: Planet Discovery(years) Vs Count

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

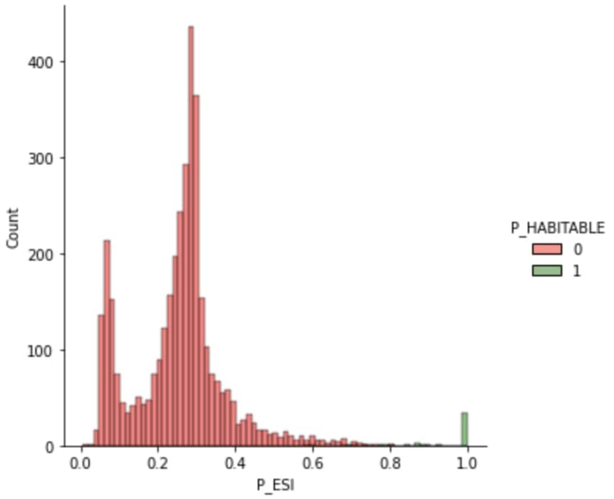

Figure 2: Importance of Earth Similarity Index

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

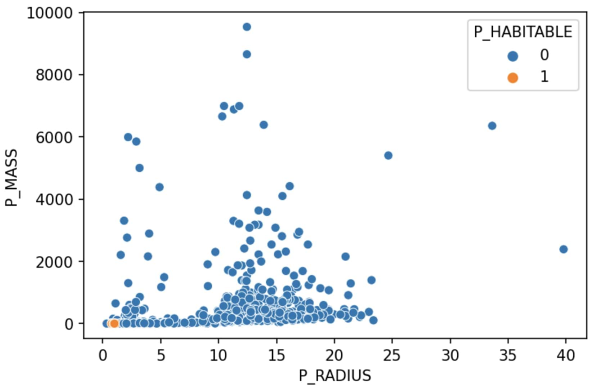

Figure 3: Planet Radius vs Planet Mass

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 remove 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 column 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 applied one-hot encoding as well as the categorical columns are low in number. The dataset is highly imbalanced as well, and to combat that we used SMOTE for artificially over-sampling 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.

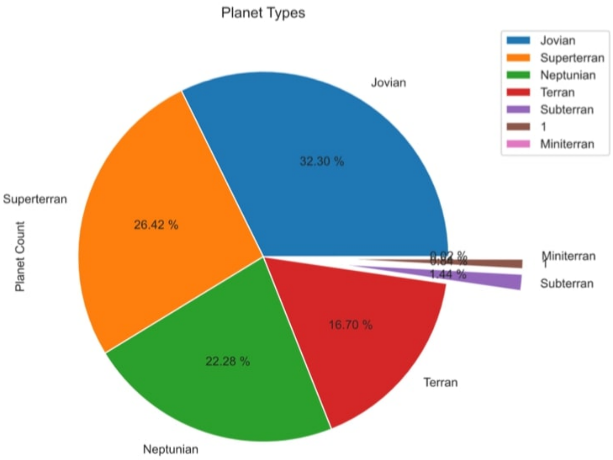

Figure 4: Planet Types

5. Methodology

Before training the models, a stratified train-test split was done in the ratio 70:30 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.

For KNN the parameters were (algorithm='auto', leaf_size=10, p=1, weights='uniform'). 'auto' will obviously be the most preferred since it chooses the algorithm 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=1.0, gamma=0.0001, kernel='rbf'). C=1.0 shows regularization is not needed for the model and the Radial Basis Function is the most appropriate for this dataset. For random forest the optimal parameters were quite complicated as follows: (n_estimators=500, bootstrap=True, oob_score=True, n_jobs=-1, criterion='gini', max_depth=8, min_samples_leaf=1, min_samples_split=2).

Finally for the XGBoost Classifier, the optimal parameters were (learning_rate=0.1, max_depth=4). This shows the XGBoost model is very simple as the max_depth required is only 4.

Logisitic regression on the other hand performed the worst when using all the features available, with f1 score of 0.0, in fact it did not predict any 1s, which shows that it was overfit, and reason is curse of dimensionality. When the top 2 most important features given by random forest were used, there was a big spike in performance with an f1 score of around 0.62, with around half of 1s being predicted correctly, which shows how bad curse of dimensionality could effect logistic regression. Regardless, it's still bad for the task of exoplanets classification.

The Feature importance measure for all the models show that Earth Similarity Index is the most contributing attribute.This is intuitive since a planet that is very similar to Earth according to its mass and distance from the star might have factor into whether it can sustain life and be habitable to classify it as a habitable exoplanet.

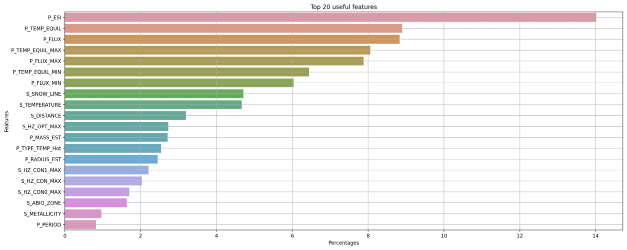

Figure 5: Feature Importance

6. Results

Random Forest				
Metric	Prec.	Recall	F1	Support
0	1.0	1.0	1.0	1198.0
1	1.0	1.0	1.0	17.0
Accuracy	1.0	1.0	1.0	1.0
Macro avg.	1.0	1.0	1.0	1215.0
Wt. avg.	1.0	1.0	1.0	1215.0

Table 1: Classification Report of Random Forest

SVM				
Metric	Prec.	Recall	F1	Support
0	1.000000	0.996672	0.998333	1202.000
1	0.764706	1.000000	0.866667	13.00000
Accuracy	0.996708	0.996708	0.996708	0.996708
Macro avg.	0.882353	0.998336	0.932500	1215.000
Wt. avg.	0.997482	0.996708	0.996925	1215.000

Table 2: Classification Report of SVM

KNN				
Metric	Prec.	Recall	F1	Support
0	0.998331	0.999165	0.998747	1197.000
1	0.941176	0.888889	0.914286	18.00000
Accuracy	0.997531	0.997531	0.997531	0.997531
Macro avg.	0.969754	0.944027	0.956517	1215.000
Wt. avg.	0.997484	0.997531	0.997496	1215.000

Table 3: Classification Report of K Nearest Neighbors

Logistic Regression				
Metric	Prec.	Recall	F1	Support
0	0.997496	0.993350	0.995419	1203.000
1	0.529412	0.750000	0.620690	12.00000
Accuracy	0.990947	0.990947	0.990947	0.990947
Macro avg.	0.763454	0.871675	0.808054	1215.000
Wt. avg.	0.992873	0.990947	0.991718	1215.000

Table 4: Classification Report of Logistic Regression

The missing values, imbalanced data and outliers were dealt with as described previously. Initially there were a total of 112 features. After removing irrelevant and NaN dominant features (where more than half of the rows were NaN), finally we have 49 features. The categorical features were then encoded to one hot encoding which proved to be better than Label Encoding. Classifiers work best when the output classes are equally distributed, and of all balancing techniques, SMOTE Tomek worked the best.
A number of classifiers were experimented and tuned using Grid Search CV. Random Forest and XGBoost gave the best results with almost a perfect f1 score, precision, recall etc.

XGBoost				
Metric	Prec.	Recall	F1	Support
0	1.0	1.0	1.0	1198.0
1	1.0	1.0	1.0	17.0
Accuracy	1.0	1.0	1.0	1.0
Macro avg.	1.0	1.0	1.0	1215.0
Wt. avg.	1.0	1.0	1.0	1215.0

Table 5: Classification Report of XGBoost

The better model would be XGBoost since it was overall a less complicated model and as Occam's Razor suggests, if 2 models score similar, then the less complicated model must be preferred.

7. Conclusion

As of now, the XGBoost model looks most promising. Even though it is not necessary for it to have a 100% accuracy in real world scenarios, a perfect score on the test set is a good indication on how the model performs on unseen data. Almost all the models which were tested have 95%+ accuracy with Earth Similarity Index as the most important feature, with random forest giving it an importance of 14% and XGBoost giving as high as 70% to it alone and achieving scores close to 100%. 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.

Further fine-tuning may be required and more models can be tried like a probabilistic classifier such as the Gaussian Naive Bayes. This will help in the final comparison and choosing the best model. Final cleaning of the code along with the documentation and the report is remaining and we have already started working on it.

8. Contributions

- **Rajat Prakash:** Literature Review, Exploratory Data Analysis, Data Visualisation, Report writing

- **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

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.