Space Objects Classification, DMML CA2

Anurag Ratnaparkhe
Dept. of Computing
National College of Ireland
Dublin, Ireland
X19229992@student.ncirl.ie

***Abstract*—The history of astronomy is a history of receding horizons. - Edwin Powell Hubble. [1]**
Astronomy has made huge leaps of advancement in the recent years, due to advancement of modern technology and increasing computing power, it has become significantly easier to collect and maintain huge amount of data which comes up with space exploration. Machine learning has numerous applications in the field of astronomy, for reference SDSS (Sloan digital sky survey) project which is an astronomical telescope, charts the sky and collects approximately 175 GB of data each night. [2] This project deals with classification problems of different sorts for e.g. Binary classification, multi-class classification and classification on highly imbalanced dataset. For The scope of this project 3 distinct datasets are used which share a common theme which is 'Space exploration'. Rigorous data pre-processing, transformation and appropriate classification models and evaluation methods to validate the models have been applied on the datasets in order to properly classify the lights coming from far distance to us in the night. questions like whether light coming from the sky is exoplanet or not, whether it is a star, galaxy or a quasar and whether the asteroids which float around the earth are potentially hazardous or not are answered using Machine learning.

I. Introduction

As the mankind started developing electronic computers back in 1940s which could do high level computations very easily, accurately and quickly, the curiosity to explore the sky and even beyond has always been strong. As soon as the first ever person "Neil Armstrong" went to the moon and returned the curiosity has only increased exponentially. As a result, there are now thousands if not millions of satellites already orbiting the earth which provide a fast and reliable way to transfer data and communicate to distant places. Life as we know it today originated in Earth and we are not able to find any trace of living organism on any other planets that we know of, as it turns out to flourish life, a planet must adhere to some rules and balance of the nature in the perfect ratio for e.g. temperature of the planet should not be too high so as to evaporate any water nor it should be too low to freeze it away. To our surprise, it turns out Earth is not the only planet in the galaxy which adheres to these very sensitive rules, but there also other planets which could potentially support life, such planets are called 'exoplanets'. After using very powerful space telescopes such as Kepler space telescope we could find out the various properties of the light which are being reflected back from these planets to us. By applying machine learning on these calculated properties of light, it is now possible to classify these lights coming from these celestial bodies as an 'Exoplanet' or 'not an Exoplanet'. By doing years and years of research we now know that, our planet is not the only planet in the world but one of the many which we cannot even count, as it turns out our planet is part of a solar system which revolves around the Sun, which is one of the billions of "STARS" in the universe. The collection of these millions of solar systems is called an "GALAXY" and the center of these galaxies have an extremely luminous active nucleus which is called a "QUASAR". The SDSS (Sloan digital Sky Survey) is a astronomy project which surveys and charts the sky by calculating the properties of light we see in night coming from billions of different sources. It is the same as mapping a country or area in the world , but here its done using high computational machines. It is possible to classify the source of these lights coming from millions of kilometers away from us. The second dataset is a multiclass classification problem which classifies these lights into 'Stars', 'Galaxies' and 'Quasars'. These stars exoplanets etc. are on an average millions and billions kilometers away from the earth, these celestial bodies are that much far away from us that in order to properly calculate the distance a different unit called 'Light years' is used. A single light year is a distance which a photon travels in a year. The point is these bodies are too far away to be hazardous for us in any way, but there are also millions of asteroids floating in sky which are not that far away, and could seriously pose a threat if one asteroids collides with earth, it could end the life on earth as we know it. JPL labs which is a branch of NASA, keeps tack of these asteroids and the data is also available to public. Using machine learning classification models it is possible to classify these asteroids as either 'pha' or potentially hazardous asteroids or not.

The General research question which could be asked by looking at these 3 datasets is , whether the lights coming from the millions of sources can be classified as an Exoplanet, Star, Galaxy or a Quasar and also the asteroids which are much closer to earth is potentially hazardous or not.
Related Work-:
Some papers used very good data understanding approach and a number of papers had some limitations, but overall learning opportunity was good.
Methodology -:
The Crisp-DM methodology is used in order to follow a proper machine learning approach. Different data pre-processing and transformation methods are applied as per the nature of dataset.

Evaluation Methods-:
Accuracy is not always the best method to evaluate the models as it can be often misleading, so different evaluation methods are used as per the nature of the problem.
Conclusion and Future Work-:
overall results were very good and all the reasearch questions were answered.

II. Related Work

Exoplanet Classification with Data Mining: [3]

This article's aim is to do case study of data mining in astronomy where the author has tried to classify exoplanets, the dataset extracted is very different from this paper, and has a target variable of multi-class labels, Decision Tree and KNN models are used to achieve this goal, the choice of the methods are good as it covers both the decision based model and value based model, However some hyper-parameter tuning could have been done by the author so as to increase the performance metric of the models.

Machine Learning Pipeline for Exoplanet: [4]

This paper aims to classify whether the celestial body is an exoplanet or not, To achieve this objective Random Forest classifier was used. The model was trained using cross validation technique and gave mean accuracy of 98 percent during the training and 95 percent of accuracy on the test data, the strong point of this paper was that a good model was used, However unlike this paper, the author did not apply any other classifier model in order to validate these high scores on random forest, and hence it remained ambiguous whether the model performed well was it over fitted.

Identifying Exoplanets With Deep Learning II: Two New Super-Earths Uncovered by A Neural Network In K2 Data: [5]

The aim of this paper is to classify exoplanets, using deep learning, nueral network .ASTRONET-K2 neural network was trained on the K2 space telescope dataset and the overall accuracy of 98 percent achieved, However The K2 dataset has a very small number of rows in the dataset approximately 2000 or less and it could bring up the problem of model being under-fit. However the model trained in theory is very powerful and could bring good results in the machine learning.

Machine Learning Approach to Classify Transit Signals and Assessing the Exoplanets Probability for Habitability: [6]

This is a MSc. research paper, and tries to classify the exoplanets using deep learning methods in combination with SVM and KNN methods.The overall methodology used is very promising however cross validation methods could have been used in order to find optimal hyper-parameter tuning.

Rapid classification of TESS planet candidates with convolutional neural networks [7]

The aim of this paper is to classify exoplanets, based on the dataset collected by TESS space telescope Using convolutional nueral network, The model worked well on simulated data with 92 percent average precision and 97 percent accuracy The start to end methodology applied in the paper is very good and data Understanding plots are also very insightful from which this paper could learn to use more better data Understanding plots.

Star–galaxy classification using deep convolutional neural networks [8]

This paper aims to classify galaxies by different shapes based on SDSS dataset and Hawai telescopes dataset,to achieve this deep learning and convolutional nueral networks are applied, the strong point of this paper is that is tackled the problem of overfitting and used various classification matrics to validate the results.

The miniJPAS survey: star-galaxy classification using machine learning [9]

The aim of this paper is to classify stars and using miniJPAS survey dataset, The author applied varieties of machine learning models such as KNN, decision tree random forest , artificial neural network for this classification problem, However the author only compared the models based on the accuracy values and did not take any other classification metrics into consideraion.

Machine Learning Applied to Star–Galaxy–QSO Classification and Stellar Effective Temperature Regression [10]

This paper deals with multiclass classification problem where the data points are classified in either being stars, galaxies or quasars various machine learning methods are applied and accuracy is discussed as the evaluation matric, The dataset is extracted from different sources and the models are trained on different datasets.

Identifying Earth-impacting asteroids using an artificial neural network [11]

The aim of this paper is to identify earth impacting aesteroids using artificial neural network,The author was able to achieve accuracy of 95 percent however no other evaluation metric was discussed so as to get a better Understanding of the model.

Hazardous Asteroid Classification through Various Machine Learning [12] The Final aim of this paper was to classify hazardous asteroids from the non hazardous ones, A wide variety of classification models were selected and trained in this paper, and XG-boost classier was stated as the best model on this dataset, However no other evaluation method such as Kappa score or specitivity or sensitivity were used in the paper.

Galaxy Morphology Classification [13] In The paper author has used galaxy zoo dataset, and aims to classify galaxies in their respective shapes, It is a multiclass classification problem, The author has used both supervised and unsupervised machine learning methods to solve this problem.Accuracy of 67 percent and 94 percent were achieved by using Random Forest and Regression, K- means clustering was also used as unsupervised learning method.

Machine learning and image analysis for morphological galaxy classification [19] In this paper, author tries to implement machine learning classification and image classification analysis,neural networks and begging ensemble methods were used to achieve the final goal,ensemble method was the most accurate.

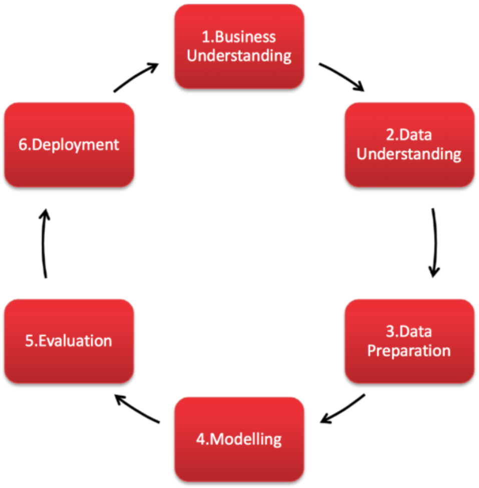


Fig. 1. CRSIP-DM Methodology

Star–galaxy classification using deep convolutional neural networks [20] In this paper the author tries to classify stars and galaxies based using deep convolutional neural networks, The SDSS has been used as the source dataset.The Nueral Network was then concluded as able to produce results as comparable to Random Forest.

Deep Learning for Star-Galaxy Classificatio [21] In this paper the author tries to classify stars and galaxies based on CNN based binary classifier. the main aim of author was to understand the working of CNN based classifier, and concluded that this classier, works well in unsupervised environment and thus human error can be reduced significantly

III. Methodology

To apply the proper start to end machine learning approach, the CRISPDM methodology is used, it is acronym for Cross industry standard process from data mining. CRISPDM is a generalized process of finding the meaningful information from the raw data. It is considered as the most popular methodology for data mining.Figure 1 explains the CRISP-DM methodlogy. [14]

A. DataSet 1 Exoplanet classification

Business Understanding :- Now a days Elon Musk is the most popular CEO, and his mission is to Colonize mars, which seems like a long way to go but not impossible, the need for search of new habitable planets is very strong as NASA and all space agencies are trying to find life or on the least meaningful resourced on the different planet. Machine learning can be applied to classify these exoplanets from the other celestial bodies.

Data Understanding :- The Dataset in order to carry out this project is extracted from NASA Archive. [15] Initially only 9564 rows were extracted from Kepler space telescope ,due to the project requirements another space telescope dataset which is K2 (a continuation of Kepler project) was extracted and merged using the common columns present in both the datasets. The new merged dataset now had 12146 rows and 12 columns of data, which fulfills the initial requirement of the project. The dataset, because of the merging process contained some null rows, but essentially were very less so were dropped. As the end goal was to classify whether the label is exoplanet or not, a bar plot was plotted in order to understand the balance of the dataset.

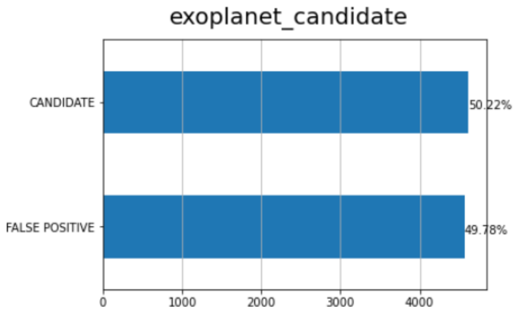


Fig. 2. Exoplanet class Balance

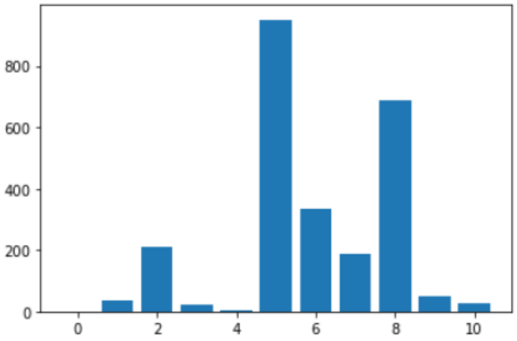


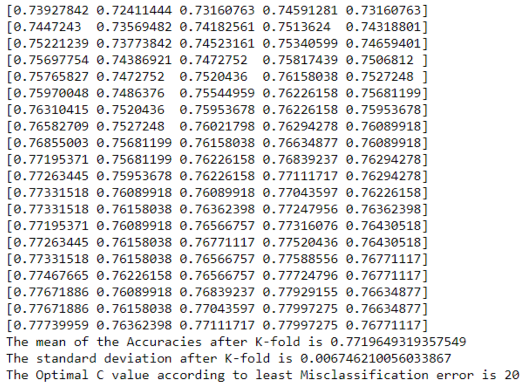
Fig. 3. ANOVA Test For Feature Selection

As the figure 2 illustrates the data is very balanced and is ready for pre-processing stage.

Data Preperation :- After Understanding the nature and distribution of the dataset, data pre-processing could be started. The first step was to select the independent variables which are going to be useful in the machine learning models, As the input data is numerical in nature and the output data is categorical in nature i.e. exoplanet or not ANOVA test was performed in order to find the important features. [16]

As the figure 3 illustrates, ’orbit id’ is not the important feature, it was droppped from the dataset, in this way the Feature selection process was done.

Also the columns vary in magnitude for e.g. the temperature column has the values in thousands as it represents effective temperature of the planet and the surface gravity is not greater than 10, there was a need to scale the dataset in equal magnitude, as not doing so can result in biased model training for this, the min max scaler function from sklearn library is used which scales all the data in equal magnitude. After this the dataset was divided in train test splits, so as to train the model on one portion and test the dataset on the other. The



```
[0.73927842 0.72411444 0.73160763 0.74591281 0.73160763]
[0.7447243  0.73569482 0.74182561 0.7513624  0.74318801]
[0.75221239 0.73773842 0.74523161 0.75340599 0.74659401]
[0.75697754 0.74386921 0.7472752  0.75817439 0.7506812 ]
[0.75765827 0.7472752  0.7520436  0.76158038 0.7527248 ]
[0.75970048 0.7486376  0.75544959 0.76226158 0.75681199]
[0.76310415 0.7520436  0.75953678 0.76226158 0.75953678]
[0.76582709 0.7527248  0.76021798 0.76294278 0.76089918]
[0.76855003 0.75681199 0.76158038 0.76634877 0.76089918]
[0.77195371 0.75681199 0.76226158 0.76839237 0.76294278]
[0.77263445 0.75953678 0.76226158 0.77111717 0.76294278]
[0.77331518 0.76089918 0.76089918 0.77043597 0.76226158]
[0.77331518 0.76158038 0.76362398 0.77247956 0.76362398]
[0.77195371 0.76089918 0.76566757 0.77316076 0.76430518]
[0.77263445 0.76158038 0.76771117 0.77520436 0.76430518]
[0.77331518 0.76158038 0.76566757 0.77588556 0.76771117]
[0.77467665 0.76226158 0.76566757 0.77724796 0.76771117]
[0.77671886 0.76089918 0.76839237 0.77929155 0.76634877]
[0.77671886 0.76158038 0.77043597 0.77997275 0.76634877]
[0.77739959 0.76362398 0.77111717 0.77997275 0.76771117]
The mean of the Accuracies after K-fold is 0.7719649319357549
The standard deviation after K-fold is 0.006746210056033867
The Optimal C value according to least Misclassification error is 20
```

Fig. 4. Mean Accuray and STD on K-fold SVM

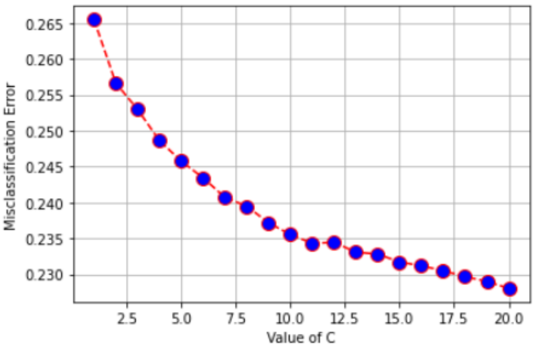


Fig. 5. SVM C value vs Misclassification error

dataset was split in 80:20 train test ratio.

Modelling :- SVM -: After all the data preprocessing dataset was ready to be trained, for the first classification model SVM(Support Vector machine) was choosen,because it works well when the data is not too noisy and after the preprocessing this dataset was already clean, and also it works well with high dimentionality space.

Hyper parameter tuning is also needed to be done on SVM like finding the optimal value for 'C=?' , which kernel should be used 'linear' or 'rbf', in order to do hyper parameter tuning K-Fold cross validation method is used which divides the train dataset in further splits or 'folds'. So as to find optimal values for these values mean Accuracy along with its standard deviation was used. First step was to test performance of 'linear' kernel on K folds, and different C values .i.e 1-20 were also fed in using a loop. The mean Accuracy on each c value was computed along with its standard deviation and the mean Accuracy was 74 percent on c value of 20 along with standard deviation of .0074, After same process was repeated with 'rbf' kernel and turns out it gave the mean Accuracy value of 77 percent on C value of 20, and thus kernel 'rbf' and C value of 20 was selected to train the model on the training dataset.

Figure 4 illustrates the mean Accuracy on C value of 1 - 20 and its standard deviation.

Figure 5 illustrates the mean Accuracy on C value of 1 - 20, and gives the C value of 20 as the best choice.

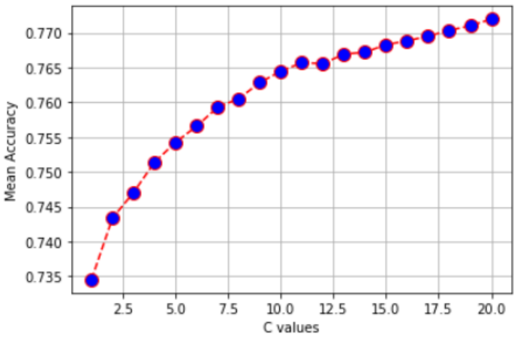


Fig. 6. SVM C value vs Mean Accuracy

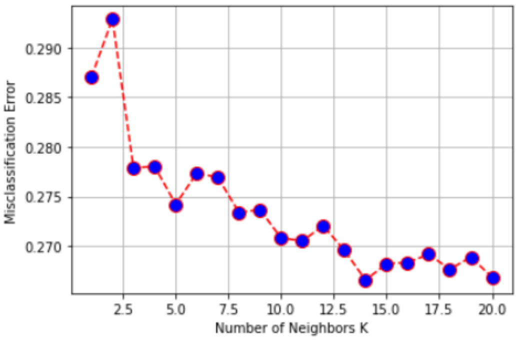


Fig. 7. K value VS Misclassification Error

Figure 6 illustrates the Misclassification error on C value of 1 - 20, and gives the c vlaue of 20 as the best choice.

KNN -: Another Model 'Knn' was used to classify the exoplanets. Knn was used in this dataset because, It is relatively easier to implement and works well with numeric input variables. The 'K' in the Knn model needs to be set manually which changes the perfomance of the overall model in order to find optimal value of K, K-fold Cross validation method was used and Figure is plotted for better Understanding. as the figure 7 and 8 illustrates as the value of K increased uptill 14 the mean Accuracy for the K folds was also increasing and misclassification error was decreasing, but after the value of 14 Accuracy started declining and misclassification error started increasing, hence value of K was chosen as 14.

B. *DataSet 2 Star,Galaxy or Quasar?*

Business Understanding :- The End goal of this dataset is a multiclass classification problem, which classifies whether the light coming from distant object is 'STAR', 'GALAXY' or a 'QUASAR'. It helps in charting or mapping the universe just as we map the countries or areas on earth.

Data Understanding :- The Dataset is extracted from SDDS website [17]

The dataset Initially contained 100000 rows and 18 column, along with the 'class' which is the target column in the dataset, which contains, values as either Star,Galaxy or a quasar. This represented a multiclass classification problem, in order to

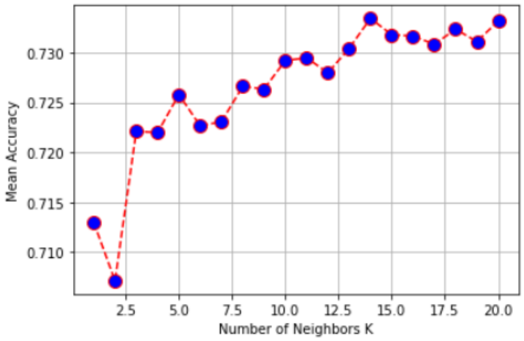


Fig. 8. K value vs Mean Accuracy

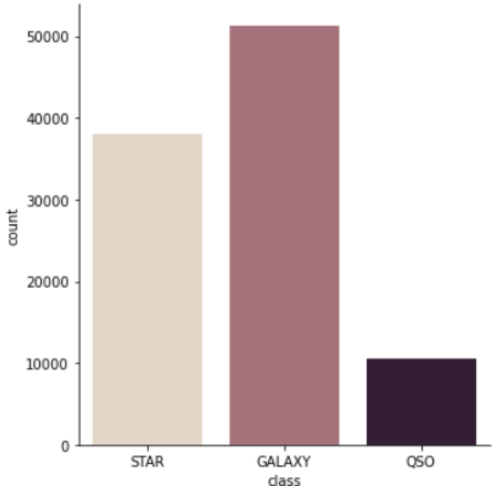


Fig. 9. CLASS Imbalance

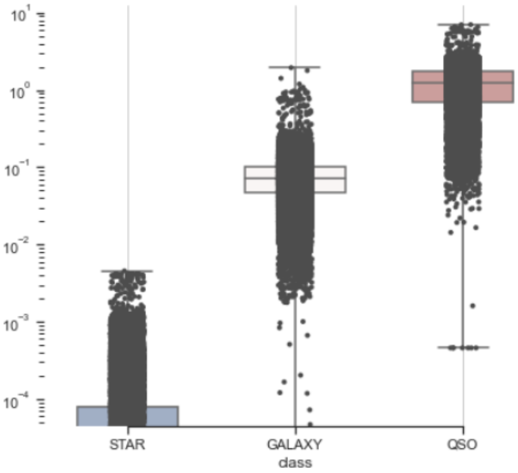


Fig. 10. Redshift of light representing distance

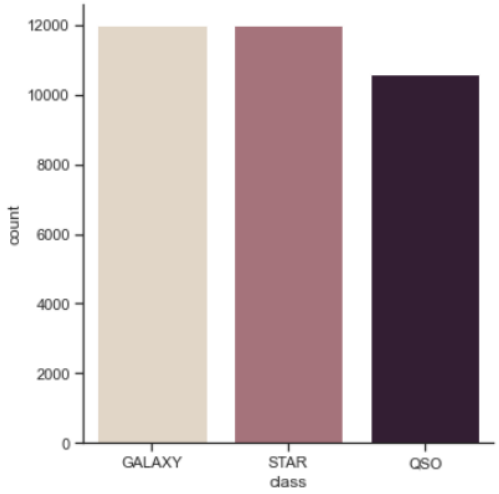


Fig. 11. Class balanced

better understand the balance of the data a bargraph was plotted.

Figure 9 illustrates that the data is imbalanced with high number of stars and galaxies but low number of quasars which needs to be addressed in the data processing phase.

Another Figure 10 illustrates that by calculating the redshift of light we can calculate the distance to the object. It represents that the stars are closest to us after then galaxies and the quasars are the most distant objects.

Data Preperation :- According to the data definition provided by sdss website the columns can be categorized in two categories that is a group of columns represents the light bands and other properties which are actually helpful while classifying the dataset, and the other group represents the camera columns and nave points which are in no way useful to us,they are only used to keep track of camera locations, hence only group 1 of data i.e. which represents the light properties are included and the other group of columns are removed. These included columns were then renamed in order to easily identify their nature. As we saw from figure 9 ,the data is imbalanced as the number of quasars are very less as compared to stars and galaxies, if the models to be applied on this imbalanced data the models could become biased towards the classes with more frequencies. In order to overcome this problem resampling methods were used such that, the classes with more count were downsampled randomly so as to balance the data.

Figure 11 illustrates that after resampling the data is now balanced After this data was split into X and y, where X represented feature columns and y label. The data was also scaled using min max scaler. The data was now ready to be split into train and test groups and 80:20 ratio was used to split the data.

Modelling :- Gaussian Naive Bayes :- As the data was downsampled randomly so as to balance it, the number of rows left to train the model also decreased, this is why Gaussian Naive Bayes was used as it works with less amount of data another reason is that Gaussian Naive Bayes works well with multi-class classification probelms , The Gaussian Naive Bayes model was trained with K-fold cross validation,The mean Accuracy of all 10 folds was 96 percent along with

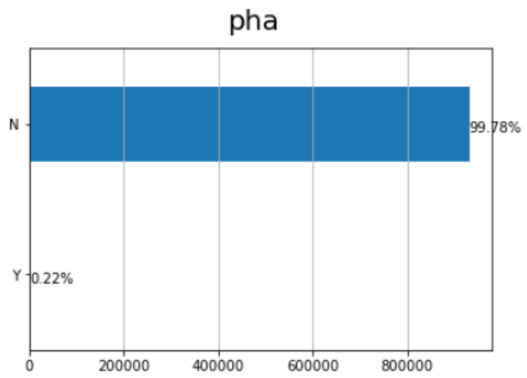


Fig. 12. Asteroid PHA class imbalance

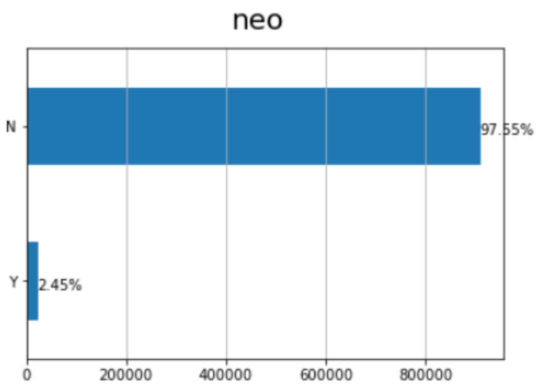


Fig. 13. Asteroid Near Earth Objects

standard deviations of .0046 which shows that the model works well in generalized environment and not just in any particular sample of data.

One vs Rest:- One vs rest classifier was also trained on the same dataset as it along with one vs one classifier also works well on multi-class classification problems. The model was imported and trained using sklearn package.

One vs One:- one vs one is also used with the same reason as one vs rest model and also to compare the Gaussian Naive Bayes model.

C. DataSet 3 Potentially Hazardous Asteroid or Not!

Business Understanding :-

The End goal of this dataset is a binary classification problem, which answers the question whether the asteroids which are near to earth are potentially hazardous or not.

Data Understanding :-

The dataset was extracted from [18], The dataset initially contained 958524 rows and 45 columns, along with label column 'pha' which had the 'Y' or 'N' values, bases on whether the asteroid of potentially hazardous or not.

The figure 12 illustrates that this dataset is highly imbalanced and needs resampling in the data processing phase in order to train a model on it.A quick glance at the data and it was understood that the dataset was a combination of numeric as well as categorical values which needed to be dealt with in data processing phase.

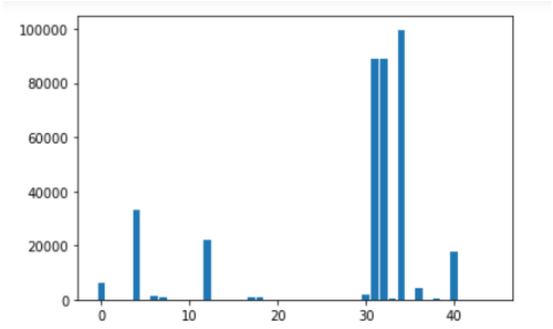


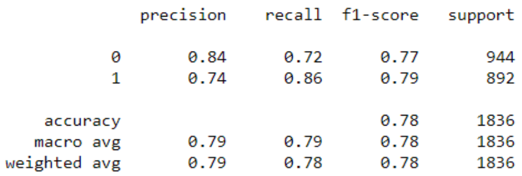
Fig. 14. ANOVA Feature Selection

Data Preperation :- Some of the columns were only representing the name of the asteroid and also had a lot of null values so they can be dropped from the further analysis. initially dropping the useless column null values were checked in the dataset and removed. As the Figure 12 illustrates the dataset is highly imbalanced with 99 percent of the tagert classes belonging to the 'N' Class, random downsampling of the majority class was done in order to somewhat balance the dataset, After this Dataset was split in X and y where X represented the features and y represented lable , also one hot encoding of y label was done and the other column with N being 0 or 1 was dropped, as it can be done in one hot encoding without the loss of any information. Different features were also on different scales so, a subset dataframe which contained only the numeric columns was extracted from original dataframe and then was scaled using min-max scaler of sklearn package, After the feature scaling procedure the subset dataframe was merged back with original dataframe,(which also contained the categorical variables). The categorical features in the original dataset were now one-hot encoded so as to prepare them to be fed in the model training. After the one hot encoding procedure the column count increased to 45, As after the one-hot encoding all the input featured were now numeric in nature and the output variable is categorical, ANOVA test was most appropriate to do the feature selecton. ANOVA test feature selection was applied on the X dataframe so as to eliminate the least important features.

As the Figure 14 illustrates the ANOVA score of features with barplot, the features with least ANOVA scores can be dropped from the the further analysis. After feature selecton the column count went down to 26.After that the data was split in train and test groups with 65:35 ratio. The rationale for doing this was as the dataset was already large enough, the testing could be done with more data.

Modelling :-

Random Forest :- The major reason for selecting random forest as the first model in this dataset was Random Forest model works well with imbalanced dataset, and also it does



```
              precision    recall  f1-score   support

           0       0.84      0.72      0.77       944
           1       0.74      0.86      0.79       892

    accuracy                           0.78      1836
   macro avg       0.79      0.79      0.78      1836
weighted avg       0.79      0.78      0.78      1836
```

Fig. 15. SVM Classification Report

not require feature scaling, But the major reason is that is is very robust and works well with both numeric and categorical data. For the hyper parameter tuning of n-estimators in random forest, different values were experimented but not much difference was noticeable.

Decision tree :- Same as Random Forest, Decision Tree classifiers also works well with imbalanced dataset, along with accepting the numeric and categorical dataset both. K-fold cross validation was also used in order to check whether the training model works well on all the samples of data or not. At all the 5 folds the Accuracy was stable, which represented that the model does work on all the random samples of data.

Gaussian Naive Bayes :-

In order to cross validate the peformane of Decision Tree and random Forest, another model .i.e. 'Gaussian Naive Bayes' was used so as to compare the base performance of Decision based models (Decision Tree and Random Forest) and probabilistic model.

IV. EVALUATION

A. DataSet 1 Exoplanet classification

CRISP DM methodology was used right from the start in order to follow the systematic approach for the machine learning model building process,The major goal was to classify whether the light coming from the distance is exoplanet or not, which both the models were able to do classify with the base Accuracies of, as the Figure 15 and Figure 17 illustrates, 78 percent and 75 percent for SVM and KNN respectively. In order to properly evaluate the models, classification matrix(Figure 16 and Figure 18) was also used along with Accuracy, The SVM model was able to classify exoplanet 74 percent of the time according to the 'precision metric' and false positives 84 percent of the time. The Classification matrix also shows the 'TRUE POSITIVES' and 'TRUE NEGATIVES' were predicted more number of times than false positive and false NEGATIVES. The precision and recall is also a bit on the higher side for SVM than KNN.

B. DataSet 2 Star,Galaxy or Quasar?

The precision/specitivity for the GNB model was 94 ,95 and 99 percent for galaxy, quasar and star respectively as shown in Figure 20. Also the overall Accuracy of the models turns out to be 96 percent, However to validate these results cohen's Kappa was also used which turns out to be 93 percent for this model. The confusion matrix in Figure 19, also shows that the model performed very well, as the diagonal shows all the true predicted values.

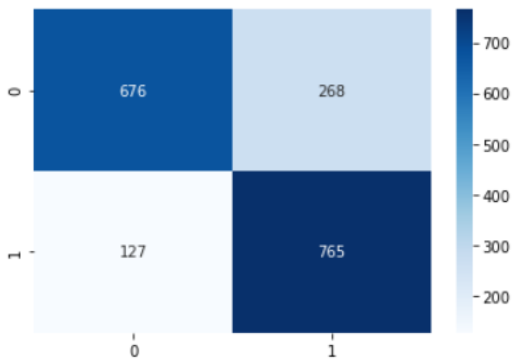
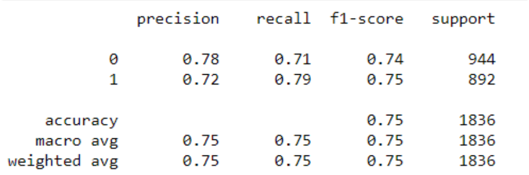


Fig. 16. SVM Confusion Matrix



```
              precision    recall  f1-score   support

           0       0.78      0.71      0.74       944
           1       0.72      0.79      0.75       892

    accuracy                           0.75      1836
   macro avg       0.75      0.75      0.75      1836
weighted avg       0.75      0.75      0.75      1836
```

Fig. 17. KNN Classification Report

OnevsRest, The evaluation of one vs rest model is also done using precision and kappa score, which is valued as 87,99,85 for galaxy, quasar and star respectively,along with 84,94,93 as the recall/sensitivity score where as the kappa score was calculated to be 85 percent as shown in Figure 21 and Figure 22.

OnevsOne, The same evaluation methods were used in this model as well so as to better compare all the models, and the precision/specitivity was 93 ,99 and 89 along with recall scores as 88,95,97 as shown in Figure 23 and Kappa score was calculated to be 90 percent as showin in Figure 25, Also Figure 24 illustrates the confusion matrix for OvO model. Here a direct comparison can be made among the three models which

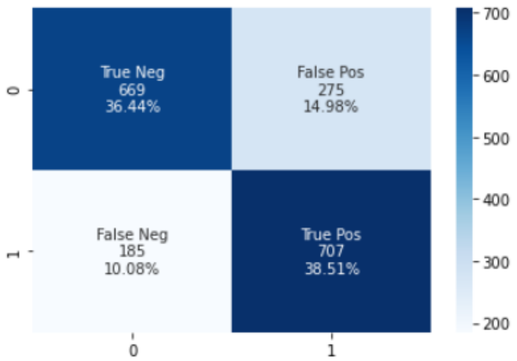


Fig. 18. KNN Confusion Matrix

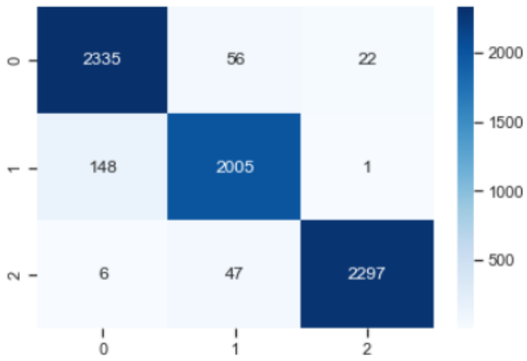
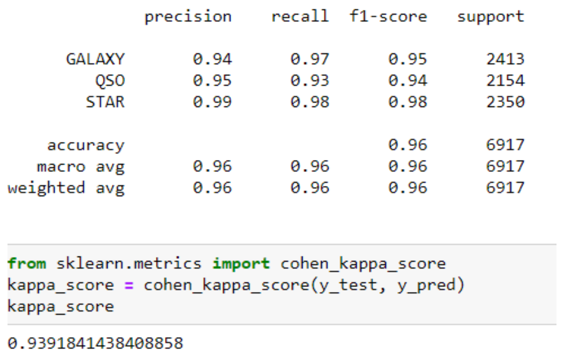


Fig. 19. GNB Confusion Matrix



```
              precision    recall  f1-score   support

      GALAXY       0.94      0.97      0.95      2413
         QSO       0.95      0.93      0.94      2154
        STAR       0.99      0.98      0.98      2350

    accuracy                           0.96      6917
   macro avg       0.96      0.96      0.96      6917
weighted avg       0.96      0.96      0.96      6917
```

```
from sklearn.metrics import cohen_kappa_score
kappa_score = cohen_kappa_score(y_test, y_pred)
kappa_score
```

```
0.9391841438408858
```

Fig. 20. GNB classification report and Kappa Score

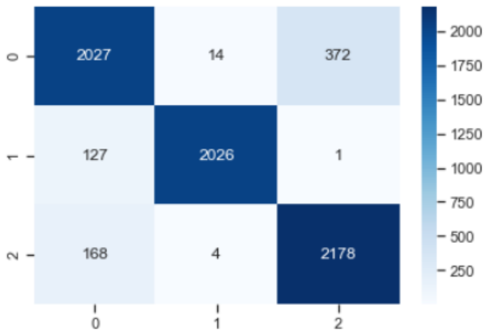
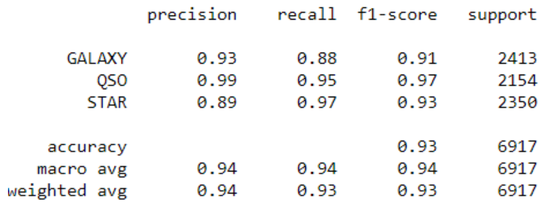


Fig. 22. OVR confusion matrix



```
              precision    recall  f1-score   support

      GALAXY       0.93      0.88      0.91      2413
         QSO       0.99      0.95      0.97      2154
        STAR       0.89      0.97      0.93      2350

    accuracy                           0.93      6917
   macro avg       0.94      0.94      0.94      6917
weighted avg       0.94      0.93      0.93      6917
```

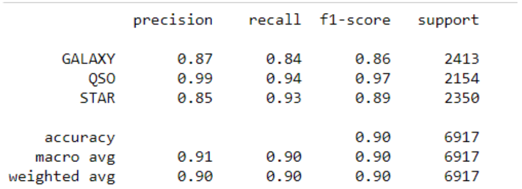
Fig. 23. OVO classification report

were applied here, with kappa score of 93 percent GNB was best fit model followed by ovo and then ovr models.

Among these 3 model based on above evaluation methods GNB model performed best, and was able to answer the initial research question.

C. DataSet 3 Potentially Hazardous Asteroid or Not!

Random Forest :- After evaluating the Random Forest model on this particular dataset, the Accuracy was 100 percent as shown in Figure 26, However Accuracy is often misleading while dealing with the highly imbalanced data and hence is not a valid evaluation matric, because for e.g. out of 10 values 9 are of one class and 1 is of other even if we dont train the model for that one minority class, it will predict 9 of the majority of the classes and Accuracy will be 90 percent, but which is useless to the end result.After evaluating based on recall and precision or specitivity and sensitivity the model is still giving the values as 100 percent and 99 percent(Figure 26) ,The confusion matrix in Figure 27 also shows the same story out of 721 total positives 717 were accurately predicted and 4 were missclassified. To validate these scores Kappa scoring was used, which is best suited to evaluate models trained on imbalanced datasets, and kappa score for the random forest model was calculated to be 99 percent as illustrated in Figure 28. Here either the model really worked well, as the three



```
              precision    recall  f1-score   support

      GALAXY       0.87      0.84      0.86      2413
         QSO       0.99      0.94      0.97      2154
        STAR       0.85      0.93      0.89      2350

    accuracy                           0.90      6917
   macro avg       0.91      0.90      0.90      6917
weighted avg       0.90      0.90      0.90      6917
```

Fig. 21. OVR classification report

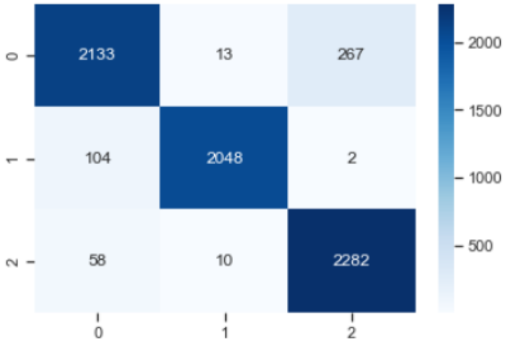
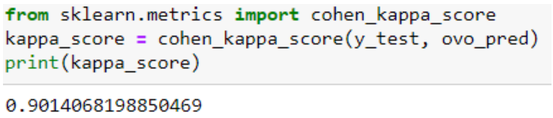


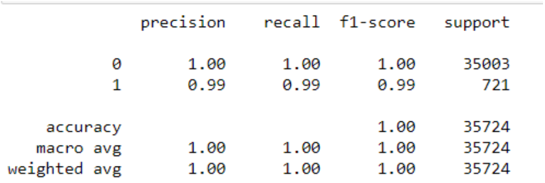
Fig. 24. OVO confusion matrix



```
from sklearn.metrics import cohen_kappa_score
kappa_score = cohen_kappa_score(y_test, ovo_pred)
print(kappa_score)
```

```
0.9014068198850469
```

Fig. 25. OVO Kappa score



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     35003
           1       0.99      0.99      0.99       721

    accuracy                           1.00     35724
   macro avg       1.00      1.00      1.00     35724
weighted avg       1.00      1.00      1.00     35724
```

Fig. 26. Random Forest Classification Report

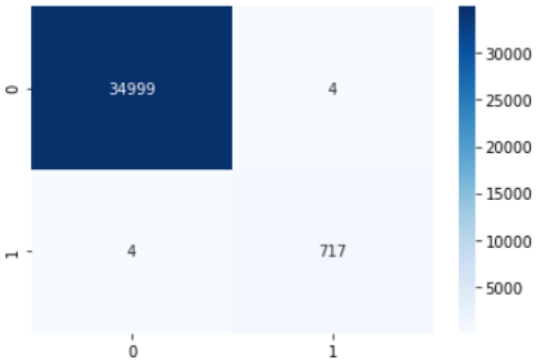
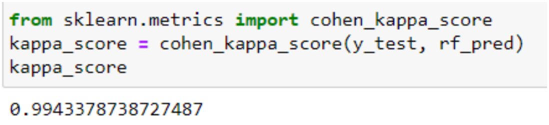


Fig. 27. Random Forest Confusion Matrix

different evaluation matrics shown, or the model was overfit and hence that's why these scores are very high.

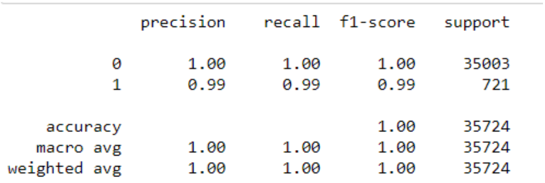
Decision Tree :- As being the model which works very similarly to random forest and have almost similar properties for e.g. both are Decision based models the results for the Decision tree were almost comparable to random forest, here also accuracy is not the correct evaluation metric. The precision and recall were also 99 and 100 for 0 and 1 respectively as shown in Figure 29.Classification matrix(Figure 30) also evaluated that model predicted 715 true positives out of 721 test samples.To validate these scores KAPPA score was evaluated and its value was also 99 percent. Here Also similar to random forest either the model worked really well or it was ”Over-Fitted”.



```
from sklearn.metrics import cohen_kappa_score
kappa_score = cohen_kappa_score(y_test, rf_pred)
kappa_score
```

```
0.9943378738727487
```

Fig. 28. Random Forest KAPPA Score



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     35003
           1       0.99      0.99      0.99       721

    accuracy                           1.00     35724
   macro avg       1.00      1.00      1.00     35724
weighted avg       1.00      1.00      1.00     35724
```

Fig. 29. Decision Tree Classification Report

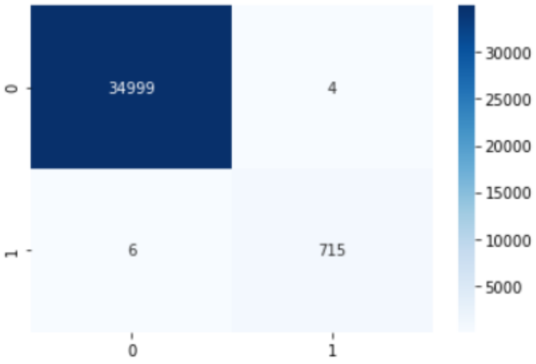
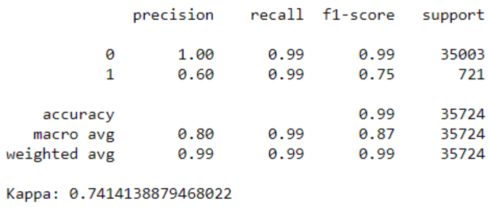


Fig. 30. Decision Tree Confusion Matrix



```
              precision    recall  f1-score   support

           0       1.00      0.99      0.99     35003
           1       0.60      0.99      0.75       721

    accuracy                           0.99     35724
   macro avg       0.80      0.99      0.87     35724
weighted avg       0.99      0.99      0.99     35724

Kappa: 0.7414138879468022
```

Fig. 31. GNB Classification Report and KAPPA Score

GNB :- To compare these scores of Decision based models a probabilistic model 'Gaussian Naive Bayes' was also trained and tested on the dataset, which gave the accuracy as 99 percent which is not useful(Figure 31), However the major difference was seen in precision and F1 score, where the precision/specitivity of the model fall down to 60 percent to classify '1' or 'pha' asteroid, along with F1 score being 75 percent. Classification matrix(Figure 32) also showed more false positives and false negatives. To validate these scores Kappa score was calculated on the model and it gave the value of 74 percent. Here It is clearly seen that the model performance fell down drastically but shows more real/believable results as compared to decision based models which could be ”Over-Fitted”.

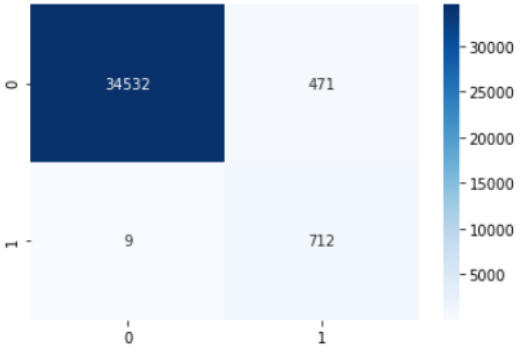


Fig. 32. GNB Confusion Matrix

V. Conclusion and Future Work

In conclusion it was observed during this project that probabilistic models works very well if the data fed in training is well scaled, and even if the decision based models are good at working with imbalanced dataset, they are very prone to ”Over-Fitting”. In the first Dataset where the aim was to classify whether the celestial body is an exoplanet or not based on the light properties reflecting from the body, The two models .i.e. KNN and SVM worked well with dataset, some hyper-parameter tuning was also done using K-fold cross validation method, in order to increase performance such as finding optimal K value for KNN and optimal C value for SVM, along with a best fit kernel type.

In Second dataset, the problem was a multiclass classification one where, aim was to classify the celestial object as 'Star', 'Galaxy' or a 'Quasar' and among the models applied i.e. GNB, One vs Rest classifier and One vs One classifier, The GNB outperformed the other two models by using probabilistic methods which fitted very well on the dataset provided.

In the third dataset, the aim was to classify whether the asteroid is potentially hazardous to earth or not, and as a logical sense, there were many values as 'N' as compared to 'Y', which signified that the dataset was highly imbalanced, in order to work well with this imbalanced data the best models which in theory works well with imbalanced data were applied, However the models became too good to be true, and can be considered as overfitted, even though three different evaluation metrics, i.e. Classification report, confusion matrix and KAPPA score the values were very high. So it is very difficult to understand whether the deciosion based models really worked well or were they overfitted as compared to probabilistic model of GNB, which showed less, but realistic and beleivable results.

In Future this project can be extended in such a way that , it becomes clear that whether the decision based models were ”Over-Fitted” or not, and also If the time permitted some more hyper-parameter tuning could be done on decision based models.

References

[1] J. D. Kelleher, M. B. Namee, and A. D'Arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Massachusetts: The MIT Press, 2015.

[2] J. D. Kelleher, M. B. Namee, and A. D'Arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Massachusetts: The MIT Press, 2015.

[3] I. dos Santos, A. B. M. Valio, N. Omar, and A. H. F. Guimaraes, “(PDF) Exoplanet Classification with Data Mining,” ResearchGate, 01-Jan-2016. [Online]. [Accessed: 01-May-2021].

[4] G. C. Sturrock, B. Manry, and S. Rafiqi, “Machine Learning Pipeline for Exoplanet Classification,” SMU Scholar. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol2/iss1/9/. [Accessed: 01-May-2021]. .

[5] Anne Dattilo1, Andrew Vanderburg1, Christopher J. Shallue, Andrew W. Mayo3, Perry Berlind, Allyson Bieryla, Michael L. Calkins, Gilbert A. Esquerdo, Mark E. Everett, Steve B. Howell, David W. Latham, Nicholas J. Scott , Liang Yu ,”Identifying Exoplanets With Deep Learning II: Two New Super-Earths Uncovered by A Neural Network In K2 Data” [Online]. Available: https://lweb.cfa.harvard.edu/ avanderb/Deep_Learning_2.pdf [Accessed: 01-May-2021].

[6] Baxi, Shreyas Shriram ”Machine Learning Approach to Classify Transit Signals and Assessing the Exoplanets Probability for Habitability” [Online]. Available: http://norma.ncirl.ie/3436/ [Accessed: 01-May-2021].

[7] H. P. Osborn, M. Ansdell, Y. Ioannou, M. Sasdelli, D. Angerhausen D. Caldwell, J. M. Jenkins, C. Räissi, and J. C. Smith ”Rapid classification of TESS planet candidates with convolutional neural networks” [Online]. Available:https://www.aanda.org/articles/aa/pdf/2020/01/aa35345-19.pdf [Accessed: 01-May-2021].

[8] E. J. Kim and R. J. Brunner, “Star–galaxy classification using deep convolutional neural networks,” OUP Academic, 17-Oct-2016. [Online]. Available: https://academic.oup.com/mnras/article/464/4/4463/2417400. [Accessed: 01-May-2021].

[9] P. O. Baqui, V. Marra, L. Casarini, R. Angulo, L. A. Díaz-García, C. Hernández-Monteagudo, P. A. Lopes, C. López-Sanjuan, D. Muniesa, V. M. Placco, M. Quartin, C. Queiroz, D. Sobral, E. Solano, E. Tempel, J. Varela, J. M. Vílchez, R. Abramo, J. Alcaniz, N. Benitez, S. Bonoli, S. Carneiro, A. J. Cenarro, D. Cristóbal-Hornillos, A. L. de Amorim, C. M. de Oliveira, R. Dupke, A. Ederoclite, R. M. González Delgado, A. Marín-Franch, M. Moles, H. Vázquez Ramió, L. Sodré, and K. Taylor, “The miniJPAS survey: star-galaxy classification using machine learning,” Astronomy &amp; Astrophysics, vol. 645, 2021.

[10] Y. Bai1, J. F. Liu1, S. Wang1, F. Yang2, Y. B. https://orcid.org/0000-0002-4740-3857, and S. W. https://orcid.org/0000-0003-3116-5038, “IOPscience,” The Astronomical Journal, 14-Dec-2018. [Online]. Available: https://iopscience.iop.org/article/10.3847/1538-3881/aaf009. [Accessed: 01-May-2021].

[11] J. D. Hefele, F. Bortolussi, and S. P. Zwart, “Identifying Earth-impacting asteroids using an artificial neural network,” Astronomy &amp; Astrophysics, vol. 634, 2020.

[12] Anish Si,”Hazardous Asteroid Classification through Various Machine Learning Techniques” Available: https://www.irjet.net/archives/V7/i3/IRJET-V7I31084.pdf [Accessed: 01-May-2021].

[13] Alexandre Gauthier, Archa Jain, and Emil Noordeh, ”Galaxy Morphology Classification” Available: http://cs229.stanford.edu/proj2016/report/GauthierJainNoordeh-GalaxyMorphology-report.pdf [Accessed: 01-May-2021].

[14] “Four Problems in Using CRISP-DM and How To Fix Them,” KDnuggets. [Online]. Available: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html. [Accessed: 01-May-2021].

[15] NASA Exoplanet Archive. [Online]. Available: https://exoplanetarchive.ipac.caltech.edu/. [Accessed: 01-May-2021].

[16] J. Brownlee, “How to Choose a Feature Selection Method For Machine Learning,” Machine Learning Mastery, 20-Aug-2020. [Online]. Available: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/#: :text=Feature%20selection%20is%20the%20process, the%20performance%20of%20the%20model. [Accessed: 01-May-2021].

[17] “Data Access for SDSS DR16 Overview,” SDSS. [Online]. Available: https://www.sdss.org/dr16/data_access/. [Accessed: 01-May-2021].

[18] M. S. Hossain, “Asteroid Dataset,” Kaggle, 30-Apr-2021. [Online]. Available: https://www.kaggle.com/sakhawat18/asteroid-dataset. [Accessed: 01-May-2021].

[19] Jorge De La Calleja, Olac Fuentes, ”Machine learning and image analysis for morphological galaxy classification” Available: https://academic.oup.com/mnras/article/349/1/87/3101624?login=true [Accessed: 01-May-2021].

[20] E. J. Kim and R. J. Brunner, “Star–galaxy classification using deep convolutional neural networks,” OUP Academic, 17-Oct-2016. [Online]. Available: https://academic.oup.com/mnras/article/464/4/4463/2417400?login=true. [Accessed: 01-May-2021].

[21] Ganesh Ranganath Chandrasekar, Iyer Krishna Chaithanya Vastare ”Deep Learning for Star-Galaxy Classification”[Online]. Available:http://noiselab.ucsd.edu/ECE285/FinalProjects/Group9.pdf [Accessed: 01-May-2021]