National College of Ireland

Machine Learning Approach to Classify Transit Signals and Assessing the Exoplanets Probability for Habitability

MSc Research Project
Data Analytics

Shreyas Shriram Baxi
x17110297

School of Computing
National College of Ireland

Supervisor: Dympna O'Sullivan Paul Stynes Pramod Pathak

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Shreyas Shriram Baxi
Student ID:	x17110297
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dympna O'Sullivan, Paul Stynes, Pramod Pathak
Submission Due Date:	13/08/2018
Project Title:	Machine Learning Approach to Classify Transit Signals and Assessing the Exoplanets Probability for Habitability
Word Count:	7508

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	17th September 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Machine Learning Approach to Classify Transit Signals and Assessing the Exoplanets Probability for Habitability

Shreyas Shriram Baxi
x17110297
MSc Research Project in Data Analytics

17th September 2018

Abstract

There have been thousands of exoplanets which have been confirmed by the scientists. These identification of the exoplanets which were carried out using the transit method involved human intervention for analyzing the signals related to the exoplanets for manually classifying the signals.There have been past work which helped in automating this classification task with the help of machine learning algorithms.This paper utilizes Deep Learning and tries to classify the detected transit signals to be related to exoplanets or non-exoplanets using signal pre-processing techniques in order to study the impact of different pre-processing tasks on the performance parameters.The findings reveal that using Savitzky Golay Filter for the filtering purpose, a higher accuracy of the model is achieved as compared to a case where no pre-processing steps are taken and also better than the case where Gaussian filtering was used. Whereas using Gaussian filtering in the pre-processing stage along with normalization and standardizing steps better recall for the model was obtained as compared to other pre-processing tasks. In addition to that,this paper also utilizes the planetary characteristics related to the exoplanets to assess the probability of an exoplanet being habitable based on the characteristics of an Exoplanet using Naive Bayes algorithm. The probabilities obtained revealed some interesting insights about the habitability which have been discussed in the evaluation section. This study also utilizes Random Forest, KNN and SVM models for the purpose of classification of Exoplanets into Mesoplanets,Psychroplanets and Non-habitable planets and all the models seem to fair similar in the performance aspects.

1 Introduction

The quest to find the new planets outside our solar system have been paced up in the last two decades as the space exploration technologies have developed. The Kepler mission proved to be a major milestone in the journey of finding the planets far away from our solar system. These planets discovered outside our solar system are termed to be Exoplanets. The curiosity of the people to know about any other habitation in the distant world have led to rapid research in a bid to identify habitable exoplanets Seager (2013a).The

exoplanets are detected with the help of numerous methods and one of them is the transit method which is utilized during the Kepler mission. The transit light signals which are captured by the telescope are used to determine the existence of an exoplanet around a star.

So, this study attempts to contribute towards the space explorations by classifying the transit signals related to Exoplanets and to improve accuracy and precision with the help of pre-processing techniques.Along with it a machine learning approach to explore the probability of an exoplanet being habitable based on the characteristics of the exoplanets and also stellar characteristics of its parent star. A Deep Learning Neural Network is used for classification of exoplanets transit light curves, whereas Naive's Bayes algorithm is used for determining the probability of an exoplanet being habitable. Random Forest, KNN and SVM are the machine learning algorithms which are being incorporated in order to find out a more accurate and precise classification model among-st them.

Transit Signals are basically the signals which are represented as flux or brightness of the star varying with the time. So, by identifying the drop in the brightness of the star at regular intervals the exoplanets can be detected. This method for detecting exoplanets is called as the transit method. There are few drawbacks of the transit method such as detecting false positives as there are a lot of stellar objects which are having similar radii as that of planets as mentioned by Rice (2014). So, a machine learning approach such a neural networks can be implemented in order to correctly classify the transit signals for exoplanets and false-positives. According to Johnson (2015) there have been more than 2500 exoplanets detected with help of Kepler telescope, and so the light curve data which consists of the transit information captured during the Kepler mission for the exoplanets have been utilized for this study.

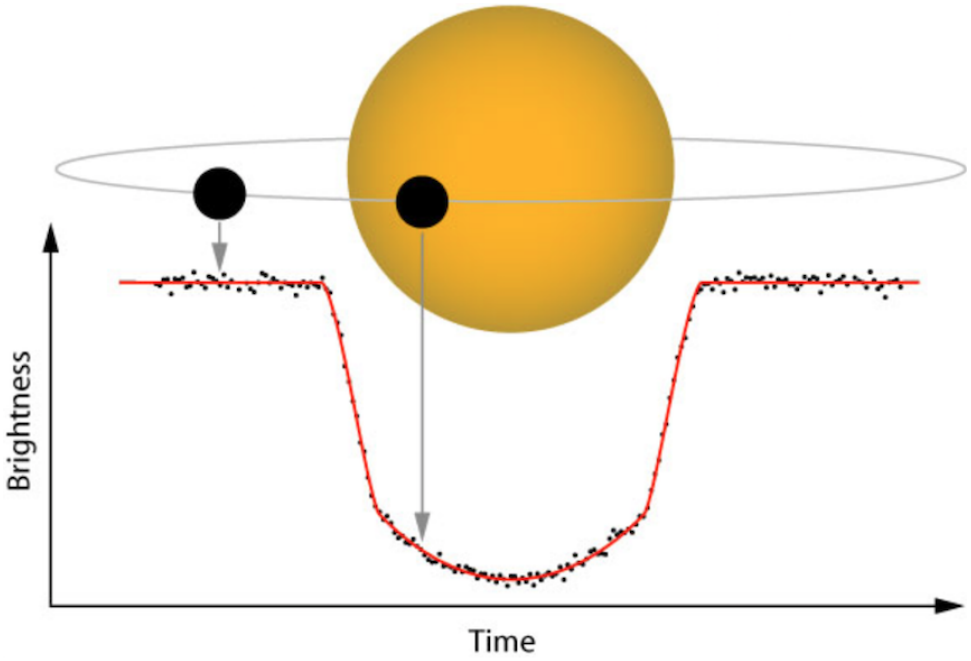

Figure 1: Transit Method

The further sections discuss in detail the different machine learning approaches and recent work carried out by the researchers in classifying the transit signals related to the exoplanets.The next section is all about the different findings obtained by the researchers

related to the habitability of an Exoplanet.Followed by that is the methodology section in which the approach to be followed for the implementation is described. The pre-processing tasks are carried out on the data prior to the model's being implemented and it is discussed discussed in the implementation section. The following section after implementation is the evaluation and discussion section wherein different results and their implications have been discussed briefly and to end a conclusion along-with future scope is also mentioned.

2 Related Work

This section discusses the state of art and the techniques carried out by the researchers in order to classify the exoplanets. The summary of this section would entail about how this study compares to that of the previous work in this particular domain.

2.1 Machine Learning Approaches for Detection of Exoplanets in Previous Studies

In the past decades the detection of the exoplanets have been carried out by vetting the light curves manually which is a time consuming activity as stated by McCauliff et al. (2015) and suggested that machine learning approaches could speed up this process to a great extent.The different approaches used by the researchers for carrying out the classification task includes the following : Random Forest, Dimensionality Reduction Technique, K-Nearest Neighbours(KNN), Self Organising Maps and Neural Networks.

So, discussing about different approaches starting with the work carried out by McCauliff et al. (2015) for classification of the stars that harbour exoplanets and the ones which didn't, based on the variability of the light intensities using Random Forest classification algorithm. This method is based on the Threshold Crossing Event (TCE) which is defined as a series of notable periodic features which resembles to a transiting exoplanet.The data used by McCauliff et al. (2015) is collected using the Kepler telescope during the Kepler mission. The main task of McCauliff et al. (2015) was to automating the process for classification of TCE's into three classes namely Planet Candidate(PC), Astronomical False Positive(AFP) and Non-transiting Planets(NTP). PC is basically the TCE's which are related to the transiting exoplanets, AFP is categorized to those TCE's which are having similar transit like features but actually aren't planets in real. The NTP is the instrumental noise which is falsely being considered to be related to that as a TCE as reported by McCauliff et al. (2015). In the experiments carried out by McCauliff et al. (2015) concluded that random forest was the best algorithm for classification as compared to that of KNN and Naive's Bayes which had a comparatively higher error rate.
In this study as well the data to be used is based on the Kepler Mission for the classification task but instead of three classes as mentioned in the work carried out by McCauliff et al. (2015),a deep learning classification model would be used in order to classify the transit signals into related to Exoplanets or False Positives.

In a study carried out by Thompson et al. (2015), the classification of transit signals was done with the help of KNN but dimensionality reduction technique was the key in his work as the light curves could have been represented using fewer features as well so a technique called as Locality Preserving Projection (LPP) was used which is a similar to

that of Principal Component Analysis(PCA) as reported by Thompson et al. (2015). The study by Thompson et al. (2015) tries to differentiate the U-shaped from the V-shaped variations in the light curves. Thompson et al. (2015) concluded that with the help of LPP and the KNN combined was able to eliminate 90% of the non transiting TCE's.

Another study by Armstrong et al. (2016) used Self Organising Maps(SOM) which is another machine learning approach which has been utilized in a bid to classify the exoplanet transit signals. The data related to the exoplanets which were detected during the Kepler mission is used in the study by Armstrong et al. (2016). It also claims SOMs to be faster and little more accurate than the work carried out in the predecessor researches in differentiating the Exoplanets from the False Positives with an accuracy of about 90%. The SOMs have been successful till that time as it was also very effective while estimating the photometric redshifts in the galaxies as reported by Carrasco Kind and Brunner (2014).

Whereas, Deep neural networks have been used previously for various studies related to the planetary science such as multi-planet detection and atmospheric classification. Pearson et al. (2017) presented a new method for exoplanets detection with the help of convolutional deep neural nets. Pearson et al. (2017) claimed that utilization of deep neural nets for this task yielded a much more accurate and precise results as compared to other machine learning algorithms. In the study carried out by Pearson et al. (2017),CNN was implemented using a data generated using an algorithm which was similar to that of the transit signals detected during the Kepler mission. The aim of the study conducted by Pearson et al. (2017) was to detect exoplanets in a noisy environment.The data used was not the real data but a similar recreated data. Whereas, in this study the data represents a real data corresponding to the Exoplanets and Non-Exoplanets. Also, as suggested by Pearson et al. (2017) including a pre-processing step would lead to potential improvement in the performance of the model in detection. So, a pre-processing step is also been incorporated in the proposed study which consists of Normalizing and flattening the signal with the help of appropriate filters such as Savitzky Golay Filter and Gaussian Filter.

2.2 Habitability of Exoplanets

According to Seager (2013b) one of the important things for a planet to be habitable is the existence of water in liquid form. So, for water to exist in a liquid form on the exoplanet it should be at an appropriate distance from its parent star. This zone where the liquid water may exist on the surface of the exoplanet is known as the Habitable zone or the Goldilock's zone as stated by Seager (2013a). As per Kopparapu et al. (2013) habitable zone for an Exoplanet is considered to be between a region starting with a distance of 0.95 AU from the parent star that is the inner edge of this habitable zone and extends up to 1.97 AU distance from the parent star which is the outer edge of the habitable zone. AU stands for Astronomical Units, and 1 AU is defined as the distance between the earth and the parent star. Late in the 20th Century Kasting et al. (1993) presented a model and concept associated to the habitable zones of the exoplanets for the very first time. In that study it was suggested that Venus which is at a distance of 0.7 AU and Mars which is at 1.5 AU from sun should have been habitable as per the bounds stated for the habitable zone as stated by Kopparapu et al. (2013). Still they are non-habitable and Kasting et al. (1993) explains for this by reporting that Venus and Mars are not-habitable due to "run away greenhouse gases". So, it can be concluded

from the studies of Kopparapu et al. (2013) and Kasting et al. (1993) that atmospheric composition is also important along with necessity of being in the habitable zone.

So, in the proposed study as well while determining the probability of the Exoplanets of being habitable both features related to the habitable zone as well as atmosphere and planets composition have been taken into account while building the models.

As, Earth is the sole planet which is known to harbor life in the entire universe till date so it would make sense for comparing the different parameters related to the Exoplanets to that of the Earth. According to Laboratory (2018) besides surface liquid water other measures can also be accounted for determining the habitability of an exoplanet which includes size, radius, mass and orbit of the Exoplanet. As reported by Laboratory (2018) the mass of the exoplanet should lie between 0.5 to 5 times the Earth masses whereas radii should be in the range of 0.8 to 1.5 times the Earth radii must lie in the habitable zone while revolving around the star.

But, in the proposed study for evaluation purposes for determining the probability the Exoplanets with a radii as much as 2.5 times that of Earth and mass upto 10 times that of the Earth would be considered as supported by the literature provided by Kopparapu et al. (2013).As, it explains that even if the Exoplanet of interest is having a radii as large as 2.5 times of Earth and mass 10 times of Earth, it may also have a comparatively larger habitable zone so they could be considered for vetting for habitability as well.

There has been very little amount of work carried out till now, when it comes to using machine learning approaches for determining the habitability of Exoplanets. So,using the support of the above mentioned literature it would be meaningful to assess the probability for habitability of an Exoplanet using machine learning approach and build a classification model to classify the Exoplanets into Mesoplanets, Psychroplanets and Non-Habitable planets.

3 Methodology

A KDD approach is followed for the purpose of implementation. KDD basically involves five stages namely selection, pre-processing, transformation, Data Mining and Evaluation/Interpretation. The application of these above mentioned stages in this study have been mentioned in the implementation section in detail.

3.1 Methodology for Classification of Exoplanet Light Curves

The first part of the proposed study is implemented for classification of time series data related to exoplanets and non-exoplanets using Deep Learning.The programming language used for this task is Python. This classification is implemented along with the following pre-processing steps as mentioned in the figure 2. The steps carried out for the pre-processing consists of using oversampling technique known as SMOTE(Synthetic Minority Oversampling Technique) for handling the imbalance in the data,the Normalization and Standardization steps are carried out in order to get the signals in to a uniform scale for the purpose of comparison. A smoothing Filter known as Savitzky Golay Filter is used in this study with a purpose of boosting the performance of the classification technique .

Savitzky Golay Filter is incorporated in the pre-processing activities due its robustness in a noisy environment, and also the way it works that is by working out average for the neighbouring points and replacing the data points with it as mentioned by Press (1996).

The Evaluation is carried out by computing the accuracy and the creating training history plots for accuracy and the loss over the training epochs. The performance is computed in the following scenarios without filtering , with Gaussian Filter and then using Savitzky Golay Filter.

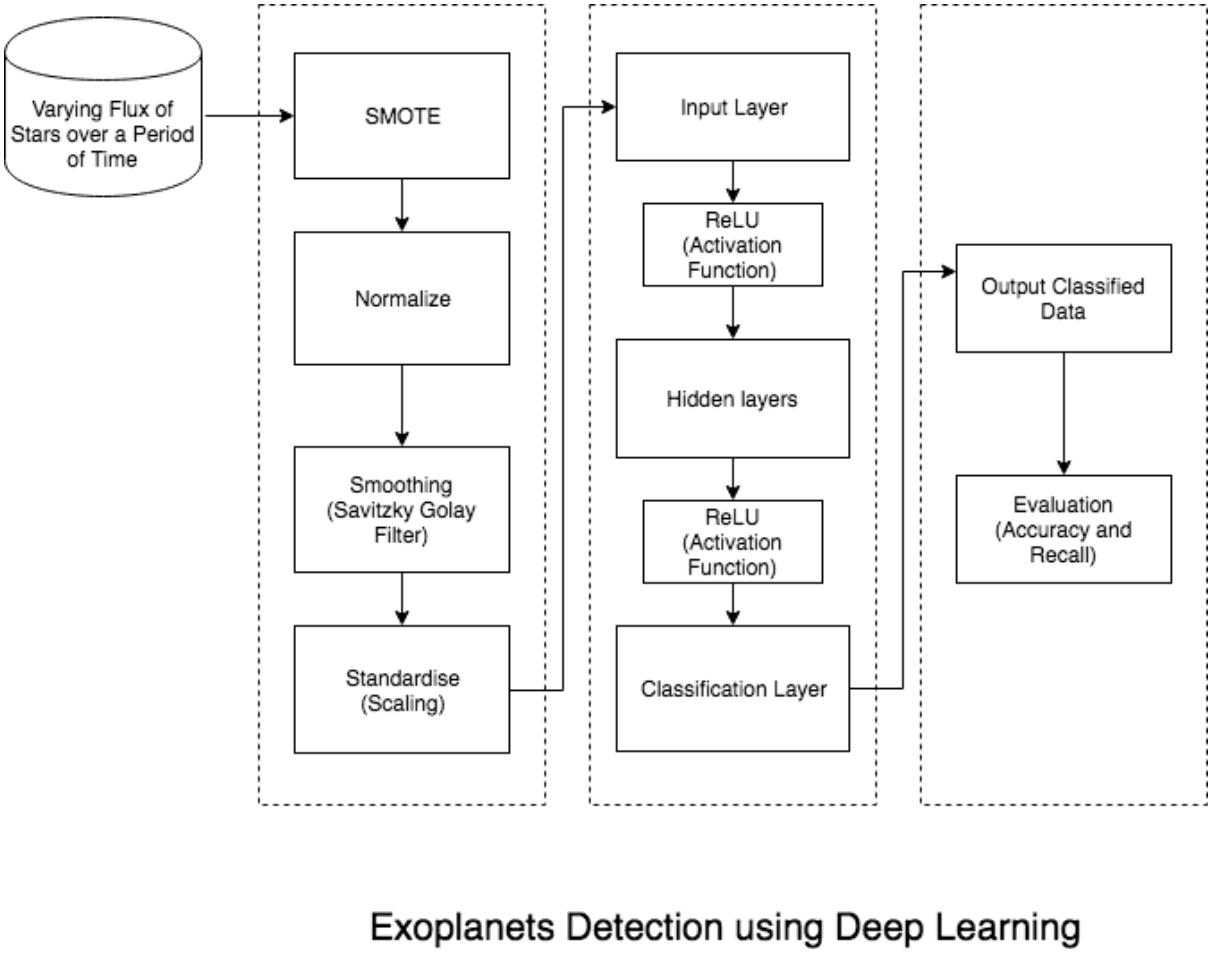

Figure 2: Block Diagram for Classification of Exoplanets Light Curves

3.2 Methodology for Probability of Habitability for Exoplanets

The later part of the study is implemented using programming language called R for getting the insights about the probability of an Exoplanet being habitable based on the various characteristics of the Exoplanets.The tasks carried out to achieve these are handling the missing values,handling the Imbalance in data using SMOTE, converting continuous variables into categorical variables in a meaningful way. All these mentioned tasks have been discussed briefly in the sub-sections of 4.2.

Whereas one of the task related to habitability is implemented for the classification into Mesoplanet,Psychroplanet,Non-habitable that is based on thermal temperatures of Exoplanets. The data for the task of classification of Exoplanets into habitable or not is obtained from Planetary Habitability Laboratory's(PHL) Exoplanet catalog.The task of finding the probabilities is implemented using predictor variable called P_Habitable which is coded 1 for habitable and 0 for non-habitable. In addition to it one more variable is

important in the data namely P_HabitableClass which is used as the predictor variable for building a classification model based on thermal temperatures of the Exoplanets, it consists of three classes Mesoplanet, Psychroplanet and Non-Habitable planet. So, classification models have been implemented to achieve this task using random forest, SVM and KNN.The classes for the classification are defined as follows:

1. Mesoplanets- These are the planets which are having a size smaller than Mercury but larger than the Ceres.According to *A Thermal Planetary Habitability Classification for Exoplanets* (n.d.) Mesoplanet term is derived from microbiological term Mesophile which refers to the organisms which can grow in a thermal conditions ranging from 10 to 45 degrees Celsius.So, same way Exoplanets which support life between 10 to 45 degrees Celsius are termed as Mesoplanets.

2. Psychroplanets- These are the planets which are having thermal temperatures ranging from -50 to 0 degrees Celsius. Pyschroplanets are the planets which harbour some simple lives even in such extreme temperatures as well as a study carried out by Price (2000) suggests that Psychrophiles are found to be habitat deep inside the ice of Antarctic . So, it would be sensible classifying the planets as Psychroplanets.

3. Non-Habitable Planets- These are basically the planets which do not belong to either Mesoplanets type or Psychroplanets type of Planets based on the thermal classification on planets.

3.3 Data Description

3.3.1 Data for Classification of Light Curves Related to Exoplanets and Non-Exoplanets using Deep Learning

The data to be used is a time series data which consists of flux values from the stars over certain time periods and a label associated with it corresponding to Exoplanets or Non-Exoplanets related light curves. 1 corresponds to a light curve related to Non-Exoplanets whereas 2 corresponds to the light curve related to Exoplanets. The data is the original Kepler data obtained from Kaggle.com, which was originally extracted from the Exoplanets archive hosted by MAST(Milkulski Archive for Space Telescopes). This data obtained through kaggle is a result of extraction of flux and time from the raw telescope data files that are in .fits format. The data is related to the campaign three of the Kepler Mission. The data consists of 3198 columns in which first column consists of labels whereas the remaining columns hold the values for flux over the time.

3.3.2 Data for Estimating the Probability of Exoplanets Being Habitable

The data for this task consists of a structured data which is obtained from the Planetary Habitability Laboratory(PHL) website. The steps to acquire the data is mentioned in the configuration manual. The data is used to achieve two tasks namely getting the probability of the exoplanets being habitable and the other is classifying the Exoplanets into Mesoplanets, Psychroplanets and Non-habitable planets.

4 Implementation

The selection, pre-processing and Transformation for all the mentioned tasks are carried out according to the KDD approach and mentioned in the further sections.

4.1 Implementation for Building Exoplanets classification model

4.1.1 Data Pre-processing

The pre-processing of the data consisting of light curves of Exoplanets are carried out using following steps.

- Normalizing
- Gaussian Filtering
- Filtering using Savitzky Golay Filter
- Standardizing

4.1.2 Model Building for Deep Learning

After the pre-processing of the data is completed the data is feed to the deep learning model. The model architecture consists of input layer with ReLU activation function, one hidden layer with ReLU activation function and one output classification layer with a Sigmoid activation function. The number of hidden layers and values for hyperparameters used are decided with the help of Trial and Error technique in order to maximize the precision, recall and accuracy for the testing data.
The hyperparameters used are as follows:

1. Learning rate = 0.001
2. Drop-out rate = 0
3. Epoch = 100
4. Batch-size =32

The optimizer used to find the set of optimal weights is Stochastic Gradient Descent(SGD) algorithm. So, of the various SGDs the algorithm called Adam is used. The SGD is dependent on the loss function, so a logarithmic loss function called binary crossentropy is used as the dependent variable is binary in nature.

4.2 Implementation for the Probability of an Exoplanet being Habitable

The data is downloaded from the Planetary Habitability Laboratory website. The data needs to be pre-processed prior before fitting to a model. In pre-processing of the data, the main challenges are imbalance in the predictor variable and the missing values.The count of these missing values is exorbitant and data imputation could not be used in this scenario as the originality of the data would be hampered to a greater extent.Some of the columns were dropped from the data due to following reasons: Missing values, large number of zeroes in particular columns, unimportant columns and inter-correlated columns.

4.2.1 Handling Missing Values

The basic exploration of data for missing values is carried out using both Microsoft Excel and R in order to know extent of missing values. According to the exploration of columns, some of the columns were deleted with the help of Microsoft Excel and some using R after importing the data into the work space.The plot for the missing values is presented in Figure 3. By looking at the plot it can be easily decided which of the columns to drop and which columns need to be considered for analysis.So, basically once we run the code the column names to be deleted from the data are displayed. The columns which are retained contain missing values as well so using R functions such as na.omit(), the missing values are removed accordingly.

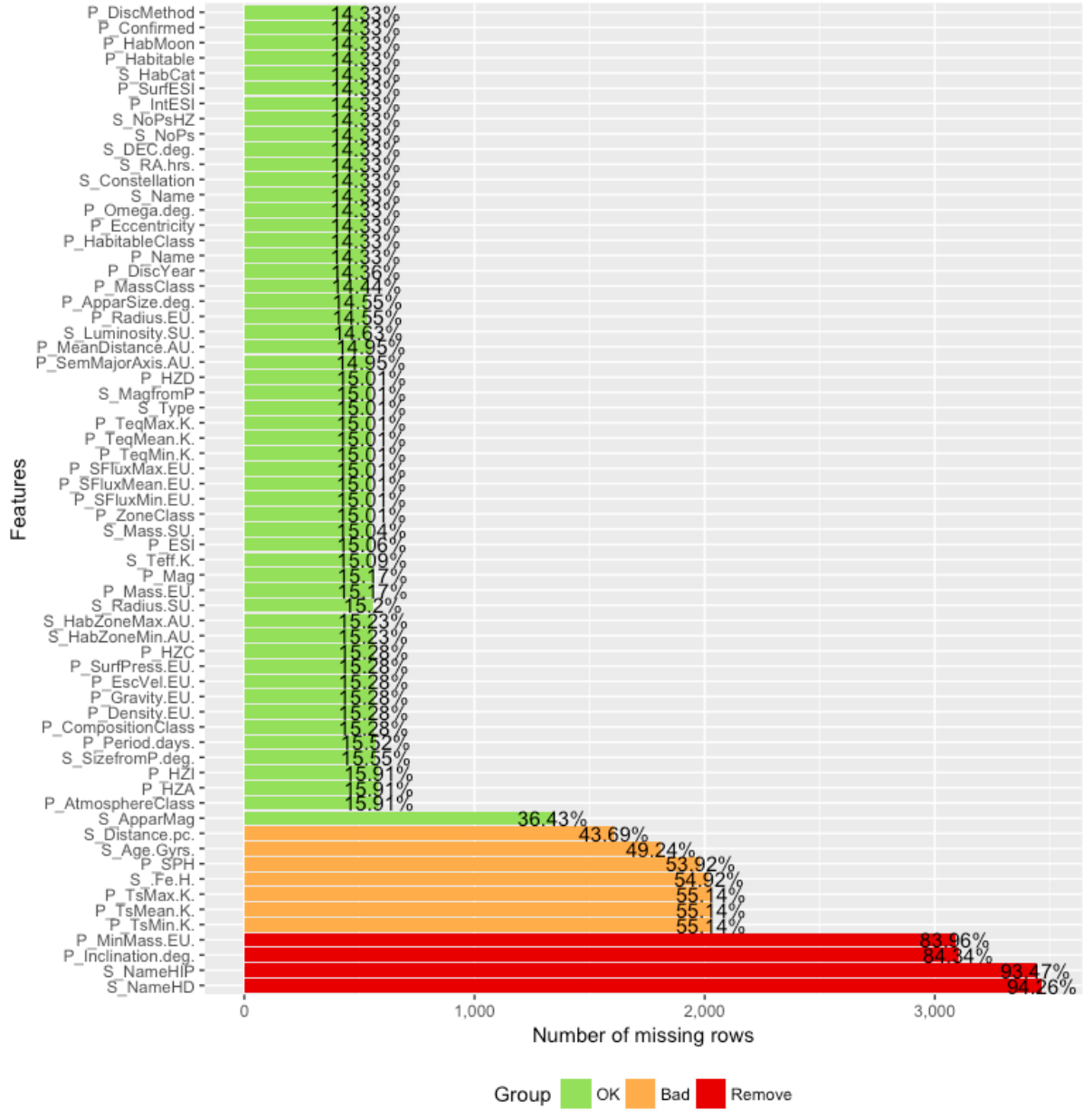

Figure 3: Plot Showcasing Missing Values

4.2.2 Handling Imbalance in the data

The imbalance in the predictor variable is handled by using an oversampling technique known as Synthetic Minority Oversampling Technique(SMOTE) with chosen values for perc.over as 2400 and the value for K as 3 for selecting the nearest three data points while creating new samples. The next step is to get the variables into appropriate data types for example, coding the categorical variables as factors for implementation purposes.

4.2.3 Converting Continuous Variables into Categorical Variables using Data Exploration Techniques

The Naive's Bayes algorithm is known to be effective while estimating the probability of an event. A categorical variables in a data are expected to be provided as input for calculation of different probabilities and in turn build a classification model using Naive's Bayes algorithm.The tables generated with the help of Naive's Bayes classification model gives a meaningful insight by providing information regarding parameters which make an exoplanet more probable to be habitable.Some of the variables consisted in the data-set are continuous in nature, like mass of the planet, radius, density, Effective temperature,and few more. So, prior to passing these continuous variables to the model they are converted into categorical variables. The continuous variables are converted into categorical by examining the distribution of data using tools such as histogram and table() function in R to inspect the number of values below or above a certain number before deciding the splits accordingly to the distribution of data. In some scenarios the literature mentioned in the section 2.2 regarding different values for physical characteristics such as mass or radius of Exoplanets and its implication on habitability is also considered while deciding the splits for the continuous variables.

4.2.4 Creating Train and Test Sets

Now, once the data is ready to be pushed into the Naive's Bayes model, the data frame consisting of the pre-processed data is now randomly split into train and test data with proportions as 75% and 25% respectively. The model is trained on the training data with the predictor as P_Habitable which consists of 0 and 1 corresponding to Non-Habitable and Habitable respectively. The model is than tested with the help of test data for evaluation purposes. The relevant insights obtained through this model and the performance metrics are discussed in the further sections in brief.

4.3 Implementation for Classification of Exoplanets into Mesoplanet,Pyscroplanet,Non-habitable Planet

4.3.1 Classification using Random Forest

The Random Forest is basically an ensemble of decision trees which is implemented in this study with the sole purpose of getting the variable importance plot in order to get the information regarding the variables which are having a higher importance with respect to the dependent variable as compared to the others. The planets need to be classified into three classes that is Mesoplanets,Psychroplanets and the Non-Habitable planets. Before building a random forest model for classification the data is over-sampled using SMOTE to reduce the imbalance in the data.In the next step training and testing sets are created by randomly sampling the data. Now, the model is ready to be built. The dependent

variable is passed to the model along with the training set. The performance metrics are then calculated with the help of test data set and confusion matrix.The variable importance is calculated with the help of Random Forest model which is then used as a reference for variable selection while implementing KNN and SVM models. The Variable Importance Plot is visualized with the help of Tableau. The performance metrics are presented in the evaluation section and a variable importance plot is presented in Figure 4.

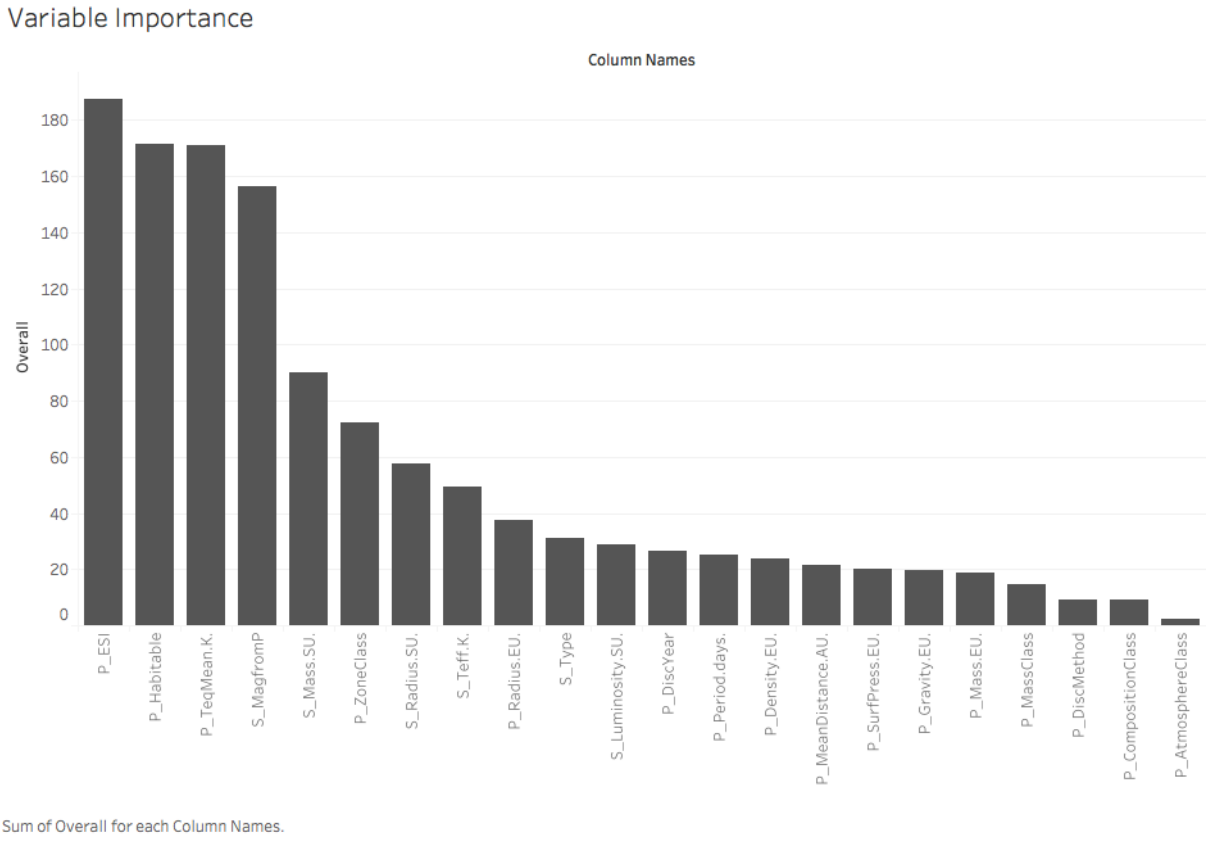

Figure 4: Variable Importance Plot

4.4 Classification using KNN and SVM

The first step remains the same in this task as well that is handling the imbalance in the data with the help of SMOTE algorithm. In the later stage creating the training and testing data sets. For, both KNN and SVM one additional processing needs to be carried out on the data which is known as normalization. The normalization of the data is essential for the KNN and SVM as both rely on the euclidean distances which is computed in a feature space so in order to compare between different variables they must lie in the same range or the scale. The KNN and SVM models are then implemented with the help of their respective library functions. SVM is used with default control settings and Kernel type whereas the value of K is chosen after trying out different values of K and finally narrowing down to K=9 for the KNN classification model.In both KNN and SVM a selected number of variables are used for building the model. These variables are selected with the help of the variable importance plot obtained from

random forest.The columns selected for implementation are P_MassEU,P_Density.EU, P_Gravity,P_TeqMean.K,P_SurfPress,P_MeanDistance.AU, S_Type, S_MassSU,S_Radius.SU, S_Teff.K, S_Luminosity.SU,S_MagfromP,P_ESI. The performance metrics are presented in the next section that is Evaluation.

5 Evaluation

5.1 Evaluation of Deep Learning Model

The Evaluation for this classification task of light curves related to exoplanets or not using deep learning is carried out using three cases. In the Case 1 the performance parameters are evaluated for a model without pre-processing performed on the light curves, in the Case 2 pre-processing is carried out by normalizing and standardizing along with applying Gaussian Filter to the training and testing data, and Case 3 a Savitzky Golay Filter is used in place of the Gaussian Filter in the pre-processing task rest is similar to that of case two.

5.1.1 Case 1 : Without Pre-Processing

The model accuracy and loss for the 100 epoch for the case 1 is presented in the figures 5 and 6 respectively.

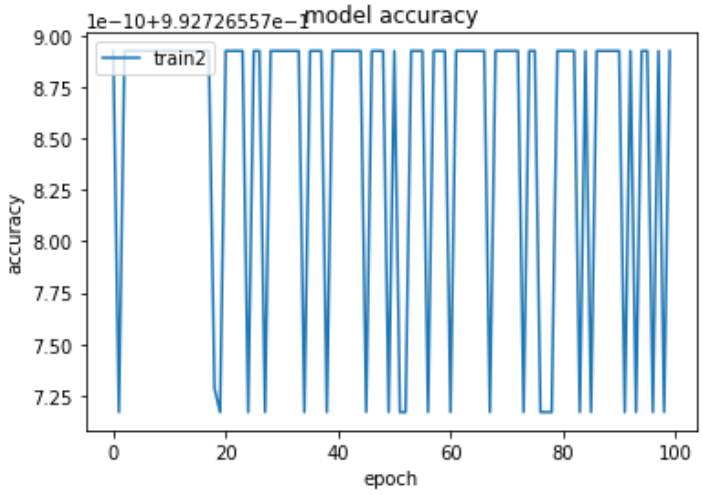

Figure 5: Model Accuracy with No Pre-processing

The different performance metrics are mentioned as below for case 1:

1. Training Accuracy = 0.9927
2. Testing Accuracy = 0.9912
3. Training Set Error = 0.0072
4. Testing Set Error = 0.0087
5. Precision Train Set = 0.0

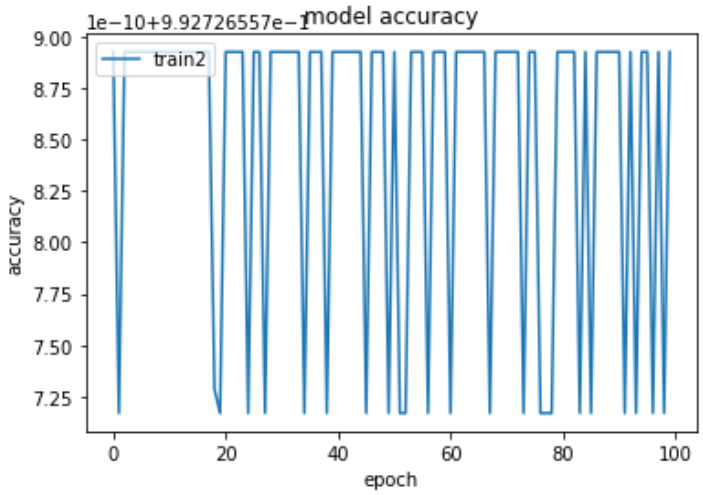

Figure 6: Model Loss with No Pre-processing

Table 1: Confusion Matrix Train Set for case 1

3328	1722
0	37

6. Precision Test Set = 0.0
7. Recall Train Set = 0.0
8. Recall Test Set = 0.0

Table 2: Confusion Matrix Test Set for case 1

336	229
0	5

5.1.2 Case 2 : With Gaussian Filter

The model accuracy and loss for the 100 epoch for the case 2 is presented in the figures 7 and 8 respectively.

The different performance metrics are mentioned as below for case 2:

1. Training Accuracy = 0.6614
2. Testing Accuracy = 0.5982
3. Training Set Error = 0.3385
4. Testing Set Error = 0.4017
5. Precision Train Set = 0.021
6. Precision Test Set = 0.0213

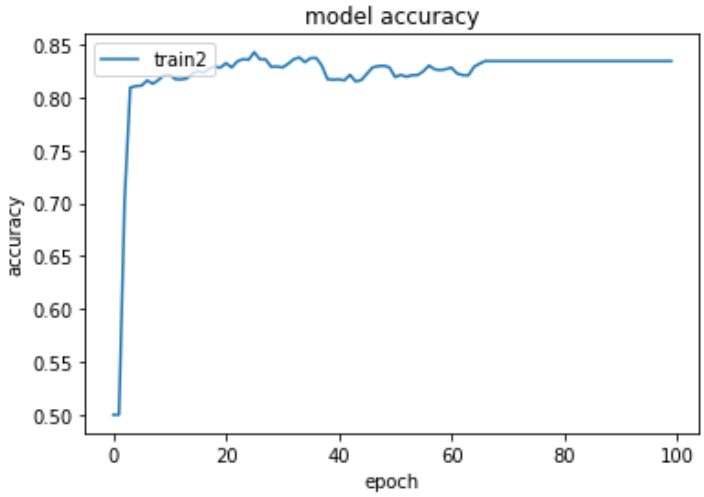

Figure 7: Model Accuracy with Gaussian Filter

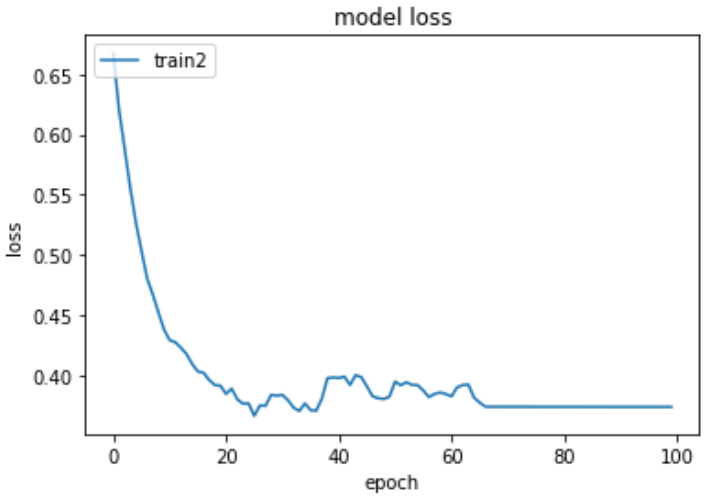

Figure 8: Model Loss with Gaussian Filter

Table 3: Confusion Matrix Train Set for case 2

3328	1722
0	37

Table 4: Confusion Matrix Test Set for case 2

336	229
0	5

7. Recall Train Set = 1
8. Recall Test Set = 1

5.1.3 Case 3 : With Savitzky Golay Filter

The model accuracy and loss for the 100 epoch for the case 3 is presented in the figures 9 and 10 respectively.

The different performance metrics are mentioned as below:

1. Training Accuracy = 0.6846
2. Testing Accuracy = 0.7614
3. Training Set Error = 0.3153
4. Testing Set Error = 0.2385
5. Precision Train Set = 0.0225
6. Precision Test Set = 0.0148
7. Recall Train Set = 1
8. Recall Test Set = 0.4

Table 5: Confusion Matrix Train Set for case 3

3346	1604
0	37

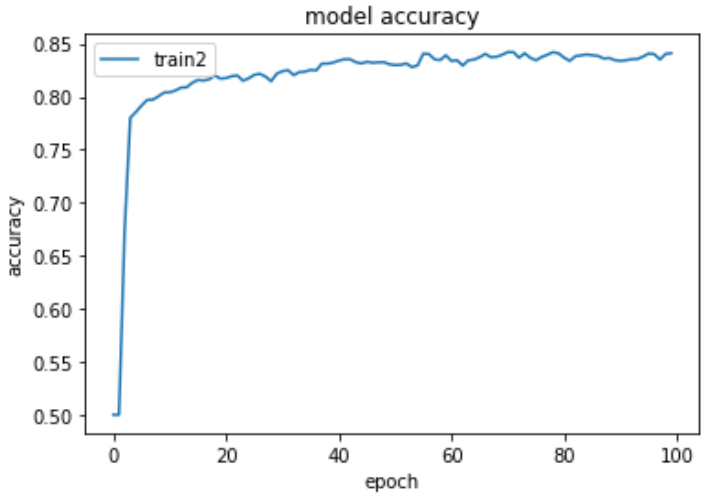

Figure 9: Model Accuracy with Savitzky Golay Filter

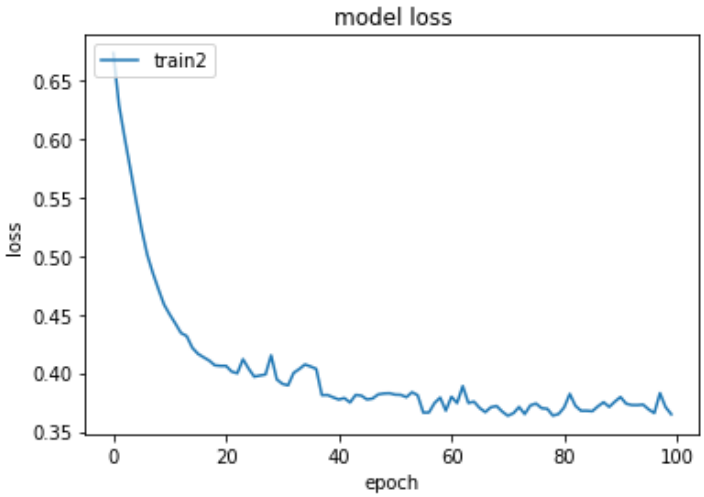

Figure 10: Model Loss with Savitzky Golay Filter

Table 6: Confusion Matrix Test Set for case 3

432	133
3	2

5.2 Probability of an Exoplanet being Habitable

Naive's Bayes Algorithm is used for this particular case study and results are outlined in this particular section. The Findings for probabilities of the Exoplanet being habitable based on various characteristics of Exoplanets itself and the parent star as well are presented as below.

```
$P_ZoneClass
   P_ZoneClass
Y             Cold            Hot           Warm
  0 0.005102040816 0.962018140590 0.032879818594
  1 0.000000000000 0.000000000000 1.000000000000

$P_MassClass
   P_MassClass
Y           Jovian      Neptunian      Subterran    Superterran         Terran
  0 0.218253968254 0.260770975057 0.022108843537 0.295351473923 0.203514739229
  1 0.000000000000 0.000000000000 0.009793253536 0.290533188248 0.699673558215

$P_CompositionClass
   P_CompositionClass
Y              gas           iron     rocky iron    rocky water      water gas
  0 0.483560090703 0.003968253968 0.488662131519 0.015306122449 0.008503401361
  1 0.000000000000 0.000000000000 0.989118607182 0.010881392818 0.000000000000

$P_AtmosphereClass
   P_AtmosphereClass
Y   hydrogen rich   metals rich no atmosphere
  0 0.14795918367 0.75907029478 0.09297052154
  1 0.15669205658 0.84330794342 0.00000000000

$P_Mass.EU.
   P_Mass.EU.
Y          0 to 1        1 to 5       5 to 10      10 to 20      Above 20
  0 0.05045351474 0.30272108844 0.17290249433 0.16893424036 0.30498866213
  1 0.08269858542 0.63329706202 0.17736670294 0.09140369967 0.01523394995

$P_Radius.EU.
   P_Radius.EU.
Y          0 to 1      1 to 2.5     Above 2.5
  0 0.04421768707 0.46088435374 0.49489795918
  1 0.06964091404 0.93035908596 0.00000000000
```

Figure 11: Probabilities for Habitability

Based on the Figure 11 which is obtained from implementation of Naive's Bayes algorithm following discussions can be inferred

- The P_ZoneClass is the Zone around the star. This suggests that the probability of an exoplanet of being habitable is very high if it lies in the Warm zone whereas it is almost certain that in would be non-habitable if it lies in other zones that is Cold and Hot.

- P_MassClass is the type of mass the Exoplanet resembles. The probability of an Exoplanet to be habitable is highest for the Exoplanets belonging to the Terran class. SuperTerran class have a subsequently lower probability of about 29% and the exoplanets belonging to the class Jovian and Neptunian are almost certainly non-habitable.

- P_CompositionClass is the composition of the Exoplanets. So, the probability of an Exoplanet to be habitable is most for rocky iron type of composition of exoplanets whereas as very less for other types of composition types.

- P_AtmosphereClass is the variable which holds information whether an Exoplanet is having hydrogen rich atmosphere, metal rich atmosphere or absolutely no atmosphere. So, the probability of an Exoplanet to be habitable is highest for the Exoplanets belonging to the metals rich type of atmospheric class.

- P_Mass is the mass of the Exoplanet measured in Earth Units(EU), that is 1EU is equivalent to the mass of the Earth. So, the probability of the Exoplanet of being habitable is the most for the Exoplanets with Mass ranging between 1 to 5 Earth Units(EU), while very less probability of being habitable if an Exoplanet has a mass in the range 0 to 1 EU and also for the mass above 10 EU.

- P_Radius is the radius of the Exoplanet, so the Exoplanets with the radius between 1 to 2.5 are having a higher probability of being classified as habitable.

```
$P_Density.EU.
   P_Density.EU.
Y        0 to 0.5      0.5 to 1        1 to 2       Above 2
  0 0.47278911565 0.17800453515 0.32936507937 0.01984126984
  1 0.00000000000 0.56474428727 0.42219804135 0.01305767138

$P_Gravity.EU.
   P_Gravity.EU.
Y         0 to 1       1 to 2      Above 2
  0 0.3509070295 0.3004535147 0.3486394558
  1 0.1447225245 0.6474428727 0.2078346028

$P_TeqMean.K.
   P_TeqMean.K.
Y        0 to 200    200 to 300    300 to 600    600 to 900   Above 900 K
  0 0.00566893424 0.03968253968 0.29988662132 0.29875283447 0.35600907029
  1 0.01088139282 0.98911860718 0.00000000000 0.00000000000 0.00000000000

$P_SurfPress.EU.
   P_SurfPress.EU.
Y         0 to 1       1 to 2       2 to 4       4 to 8      Above 8
  0 0.1678004535 0.1394557823 0.1451247166 0.1213151927 0.4263038549
  1 0.1001088139 0.2415669206 0.2437431991 0.1980413493 0.2165397171

$P_Period.days.
   P_Period.days.
Y          0 to 10       10 to 50      50 to 300     300 to 400      Above 400
  0 0.470521541950 0.377551020408 0.120748299320 0.019274376417 0.011904761905
  1 0.092491838955 0.584330794342 0.316648531012 0.006528835691 0.000000000000

$P_MeanDistance.AU.
   P_MeanDistance.AU.
Y         0 to 0.1    0.1 to 0.25      0.25 to 1        Above 1
  0 0.556122448980 0.292517006803 0.132086167800 0.019274376417
  1 0.335146898803 0.426550598477 0.236126224157 0.002176278564
```

Figure 12: Probabilities for Habitability

The below mentioned insights is implied with the help of figure 12.

- P_Density is the density of the Exoplanet. The probability of being habitable is quite descent for the Exoplanets with Density in the range of 0.5 to 1 and 1 to 2 which is around 56% and 42% respectively. Density for Exoplanets is also measured in Earth Units(EU)

- P_Gravity is the gravity on the particular Exoplanet. This is measured in Earth Units(EU) and probability for exoplanets to be habitable is higher for the ones with Gravity in the range of 1 EU to 2 EU.

- P_TeqMean is the exoplanets Mean Equilibrium Temperature which is measured in Kelvins. So, the Exoplanets with a temperature range from 200 to 300K are having very high probability of being classified as habitable. Even Earth's Equilibrium Temperature is in the same range as well.

- P_SurfPress is the pressure on the particular Exoplanets surface which is directly proportional to the mass air at that surface. It is measured in EU. So, the findings for this particular variable are quite confusing as surface pressure over the ranges 1 to 2, 2 to 4, 4 to 8 and above 8 are all having a similar probability of being classified as a habitable Exoplanet.

- P_PeriodDays is the period for an Exoplanet which is measured in days. So, the probability of the Exoplanets to be classified into habitable is high for the exoplanets with period in the range of 50 to 300 and 10 to 50. This might seem a bit absurd as compared to Earth's period which 365 days it is far more less. But, until now the Exoplanets which have been discovered are mostly with the help of transit method and the transit method. So, it is one of the drawbacks of this method as to confirm trnsists from exoplanets with this long periods the wait for second TCE event is quite long and so there is always a higher chance of missing it as compared to those with smaller periods.

- P_MeanDistance is the Planet's mean distance from the star. This is also measured in Earth Units. So, the probability of exoplanet is higher for the exoplanets at a distance between 0.1 to 0.25 and quite decent in range between 0.1 to 0.25 EU.

```
$S_Type
   S_Type
Y                A              B              D              F              G              K              M              T
  0 0.005668934240 0.001700680272 0.000000000000 0.200113378685 0.467120181406 0.272108843537 0.052154195011 0.000000000000
  1 0.000000000000 0.000000000000 0.000000000000 0.000000000000 0.054406964091 0.180631120783 0.764961915125 0.000000000000
   S_Type
Y                k            sdB
  0 0.001133786848 0.000000000000
  1 0.000000000000 0.000000000000

$S_Mass.SU.
   S_Mass.SU.
Y         0 to 0.5       0.5 to 1        Above 1
  0 0.027210884354 0.576530612245 0.396258503401
  1 0.783460282916 0.214363438520 0.002176278564

$S_Radius.SU.
   S_Radius.SU.
Y         0 to 0.5       0.5 to 1        Above 1
  0 0.043083900227 0.512471655329 0.444444444444
  1 0.808487486398 0.186071817193 0.005440696409

$S_Teff.K.
   S_Teff.K.
Y       0 to 4000  4000 to 5000  5000 to 6000    Above 6000
  0 0.05895691610 0.14285714286 0.59240362812 0.20578231293
  1 0.80522306855 0.16648531012 0.02829162133 0.00000000000

$S_Luminosity.SU.
   S_Luminosity.SU.
Y         0 to 0.5       0.5 to 1         1 to 2        Above 2
  0 0.312925170068 0.252267573696 0.240929705215 0.193877551020
  1 0.968443960827 0.026115342764 0.005440696409 0.000000000000
```

Figure 13: Probabilities for Habitability

The below insights can be implied from figure 13.

- S_Type is the type of the star that harbors a particular Exoplanet. So, considering the sun type the Exoplanet revolves around, it can be implied that the habitability for the exoplanets with the sun type as M has a 76% probability of harboring a habitable Exoplanet, whereas exoplanets having stars belonging to the type K and M have particularly lower chance of being habitable in nature.

- S_Mass is the Star's Mass which is measured in Solar Units(SU). 1 SU corresponds to the mass of our Sun. The probability of Exoplanets being habitable is higher for the ones's which have a star with a mass in the range of 0 to 0.5 SU and quite decent at 21% for the range in between 0.5 and 1 SU.

- S_Teff is the Effective temperature of the star measured in Kelvins. The Exoplanets probability for the habitability is higher if the parent star's Effective Temperature is below 4000 Kelvins.

- S_radius is the Radius of the parent star around which the exoplanet revolves around. So, the probability of habitability is good at 80% for the ones within ranges of 0 to 0.5 SU.

- S_Luminosity is the luminosity of the parent star around which the Exoplanet revolves. This is also measured in SU. So, the probability of an exoplanet of being habitable is 96% for the range of values from 0 to 0.5 SU.

```
$S_MagfromP
   S_MagfromP
Y          0 to 25       25 to 27       27 to 32       Above 32
  0 0.003968253968 0.027777777778 0.581065759637 0.387188208617
  1 0.000000000000 0.963003264418 0.036996735582 0.000000000000

$P_ESI
   P_ESI
Y       0 to 0.25  0.25 to 0.50  0.50 to 0.75         0.75+
  0 0.37131519274 0.58560090703 0.04308390023 0.00000000000
  1 0.00000000000 0.00000000000 0.18171926007 0.81828073993

$P_DiscYear
   P_DiscYear
Y     Before 2009  2009 to 2015  2015 to 2016    Above 2016
  0 0.06802721088 0.45181405896 0.42970521542 0.05045351474
  1 0.00000000000 0.21871599565 0.30794341676 0.47334058760
```

Figure 14: Probabilities for Habitability

The insights described below can be implied from figure 14.

- S_MagfromP is the magnitude of the star as observed from the planet. The exoplanet is having a higher probability of being habitable if the magnitude of the star as observed from the planet is between 25 to 27.

- P_ESI is the Earth Similarity Index of the Exoplanet which is measured in a scale of 0 to 1, where 1 being most likely identical to earth in terms of physical attributes.The Exoplanet is having a higher probability of about 81% of being habitable when the ESI is above 0.75 for a particular Exoplanet,whereas the probability of an Exoplanet being habitable is around 18% which is quite low for the exoplanets having ESI in the range of 0.5 to 0.75.

5.3 Classification Models for Exoplanets Habitability

The performance metrics for the Random Forest, KNN and SVM classification models are presented in this section.

5.3.1 Random Forest

The confusion matrix and the statistics are computed for this Random Forest Model and are presented in figure 15. From this it can be inferred that, no information rate is the

```
Confusion Matrix and Statistics

              Reference
Prediction      mesoplanet non habitable psychroplanet
  mesoplanet            47             0            43
  non habitable          0           710             1
  psychroplanet         69             0           202

Overall Statistics

               Accuracy : 0.8945896
                 95% CI : (0.8746451, 0.9123356)
    No Information Rate : 0.6623134
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 0.7864572
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: mesoplanet Class: non habitable Class: psychroplanet
Sensitivity                 0.40517241            1.0000000            0.8211382
Specificity                 0.95502092            0.9972376            0.9164649
Pos Pred Value              0.52222222            0.9985935            0.7453875
Neg Pred Value              0.92973523            1.0000000            0.9450687
Prevalence                  0.10820896            0.6623134            0.2294776
Detection Rate              0.04384328            0.6623134            0.1884328
Detection Prevalence        0.08395522            0.6632463            0.2527985
Balanced Accuracy           0.68009667            0.9986188            0.8688016
```

Figure 15: Confusion matrix for random forest

percent of predictors belonging to the majority class. It is never easy to get the accuracy higher than that of no information rate in such scenarios were the data is unbalanced.The accuracy for this particular classification model is around 89%. Also the p-value seem to indicate the model in statistically significant as the p-value is less than 0.05 and the value of Kappa is reported to be 0.7864 The other various performance metrics can be obtained by looking at the figure.

5.3.2 KNN

The confusion matrix for the KNN's classification model is also presented in the figure 16 below.

```
              Reference
Prediction      mesoplanet non habitable psychroplanet
  mesoplanet            19             0            13
  non habitable          0           583             1
  psychroplanet         84             8           185

Overall Statistics

               Accuracy : 0.881299
                 95% CI : (0.8582585, 0.9017868)
    No Information Rate : 0.6618141
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 0.7596817
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: mesoplanet Class: non habitable Class: psychroplanet
Sensitivity                 0.18446602            0.9864636            0.9296482
Specificity                 0.98354430            0.9966887            0.8674352
Pos Pred Value              0.59375000            0.9982877            0.6678700
Neg Pred Value              0.90243902            0.9741100            0.9772727
Prevalence                  0.11534155            0.6618141            0.2228443
Detection Rate              0.02127660            0.6528555            0.2071669
Detection Prevalence        0.03583427            0.6539754            0.3101904
Balanced Accuracy           0.58400516            0.9915762            0.8985417
>
```

Figure 16: Confusion matrix for KNN

The accuracy for this classification model is obtained to be 88.1%. The Kappa is reported to be 75.96%.The model is significant as the p value is less than 0.05 and accuracy is also greater than the no information rate.

5.4 SVM

The confusion matrix for the classification task using SVM is reported in the figure 17 below.

The accuracy achieved for this classification model is 87.79% whereas the No information rate is 0.6618

```
Confusion Matrix and Statistics

              Reference
Prediction      mesoplanet non habitable psychroplanet
  mesoplanet             1             0             0
  non habitable          0           585             1
  psychroplanet        102             6           198

Overall Statistics

               Accuracy : 0.8779395
                 95% CI : (0.8546547, 0.8986952)
    No Information Rate : 0.6618141
    P-Value [Acc > NIR] : < 0.00000000000000022204

                  Kappa : 0.7504986
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: mesoplanet Class: non habitable Class: psychroplanet
Sensitivity                0.009708738            0.9898477            0.9949749
Specificity                1.000000000            0.9966887            0.8443804
Pos Pred Value             1.000000000            0.9982935            0.6470588
Neg Pred Value             0.885650224            0.9804560            0.9982964
Prevalence                 0.115341545            0.6618141            0.2228443
Detection Rate             0.001119821            0.6550952            0.2217245
Detection Prevalence       0.001119821            0.6562150            0.3426652
Balanced Accuracy          0.504854369            0.9932682            0.9196776
```

Figure 17: Confusion matrix for SVM

5.5 Discussion

The results obtained by the implementation of different machine learning approaches have been able presented in the evaluation section.As, suggested by Pearson et al. (2017), the pre-processing of the light curves by using different signal processing techniques may lead to subsequent improvement in the performance of the model and this seems to be true as evident from the results obtained from this study. The performance of the deep learner model evaluated in three different scenarios yielded different results.When the model was initiated to run without using any filtering technique in the pre-processing phase, the results seemed absurd with the model accuracy resonating vigorously over the 100 epoch and the resulting recall and accuracy as well was very poor. In the other two scenarios filters were used along with the normalizing and standardizing functions in order to pre-process the signal. This yielded better results. The recall was 1 on the test set for the Gaussian filter whereas for Savitzky Golay filter had a comparatively lesser precision of about 0.4 on test set. But, the accuracy of the model was higher when Savitzky Golay filter was used as compared to the model with pre-processing involving Gaussian filter. The accuracy was expected to improve over other scenarios with the help of the Savitzky Golay filter as mentioned in the related work section due to its robustness to the noisy environment.

In gauging the probability for habitability of the exoplanets using Naive Bayes unfolded some insights as discussed in evaluation section and most of the probabilities computed made sense and it could be supported by literature as well.

The classification of exoplanets based on the habitability is not been explored much so this was a great opportunity to build classification models using random forest,KNN

and SVM.The performance of all the three models was nearly similar. The common thing observed during evaluation was that the non-habitable planets were classified more precisely as compared to the mesoplanets and psychroplanet and the reason for this is likely due to the imbalance in the data which is around 68% even after using SMOTE. The classification model for habitability would be of great help for the researchers in future studies. As, this area is not explored extensively so this gives an opportunity to build models in order to leverage the researchers by this small part of the study.

6 Conclusion and Future Work

In this research the motive was to present a complete study by building models for exoplanets light curve classification with the help of pre-processing techniques and particularly using Savitzky Golay Filter and provide insights on the habitability of the exoplanets and the probabilities associated with it of being habitable. In addition to that classification of exoplanets, using machine learning approach was adopted as very less work was done when it comes to habitability using machine learning approaches.

The classification of exoplanets was achieved with the help of deep learning neural network and the usefulness of pre-processing of the transit light curves was demonstrated successfully by evaluating three different case scenarios. The proposed use of Savitzky Golay Filter seemed to improve the accuracy of the model as compared to the Gaussian filter. The accuracy of the model on the test data set with Savitzky Golay Filter was recorded as 76.14% whereas it was lower in other two scenarios. Whereas the recall for the test set was 1 for the model filtered using Gaussian filter and that with Savitzky Golay filter was 0.4 and precision was extremely low. So, this suggests that to achieve better precision and performance large amount of data must be required with comparatively lesser class imbalance.

The probability for the exoplanets being habitable based on different planetary and stellar characteristics have been presented and explained in the evaluation section which indicated some interesting findings. There have been several missions which have been planned just for exploring the exoplanets, with the latest mission called TESS being launched in April 2018 itself. This is just a stepping stone in the journey of exploring habitable exoplanets. As, more and more data gets available in due course of time the classification models would evolve with it accordingly.The data was not so complex but lots of variables seemed to be highly inter-correlated as many of the parameters are not measured directly but are derived from other parameters. So,a more robust algorithm for proper feature selection could have improved the model performance as a whole.

In future a LSTM or CNN based approach in detecting the exoplanets not just with the help of light curve patterns but also with the detected TCE's along with the physical attributes of the exoplanet and the parent star could be implemented for a more complete case study. The habitability front of the exoplanets can be explored more if data regarding water present and the atmospheric composition of gases could be estimated as the research is going on in the same direction under the leadership of Sara Seager.

References

Armstrong, D. J., Pollacco, D. and Santerne, A. (2016). Transit shapes and self organising maps as a tool for ranking planetary candidates: Application to kepler and k2, *Monthly Notices of the Royal Astronomical Society* p. stw2881.

A Thermal Planetary Habitability Classification for Exoplanets (n.d.).
URL: *http://phl.upr.edu/library/notes/athermalplanetaryhabitabilityclassificationforexoplanets*

Carrasco Kind, M. and Brunner, R. J. (2014). Som z: photometric redshift pdfs with self-organizing maps and random atlas, *Monthly Notices of the Royal Astronomical Society* **438**(4): 3409–3421.

Johnson, M. (2015). How many exoplanets has kepler discovered?
URL: *https://www.nasa.gov/kepler/discoveries*

Kasting, J. F., Whitmire, D. P. and Reynolds, R. T. (1993). Habitable zones around main sequence stars, *Icarus* **101**(1): 108–128.

Kopparapu, R. K., Ramirez, R., Kasting, J. F., Eymet, V., Robinson, T. D., Mahadevan, S., Terrien, R. C., Domagal-Goldman, S., Meadows, V. and Deshpande, R. (2013). Habitable zones around main-sequence stars: new estimates, *The Astrophysical Journal* **765**(2): 131.

Laboratory, P. H. (2018). Hec: Description of methods used in the catalog.
URL: *http://phl.upr.edu/projects/habitable-exoplanets-catalog/methods*

McCauliff, S. D., Jenkins, J. M., Catanzarite, J., Burke, C. J., Coughlin, J. L., Twicken, J. D., Tenenbaum, P., Seader, S., Li, J. and Cote, M. (2015). Automatic classification of kepler planetary transit candidates, *The Astrophysical Journal* **806**(1): 6.

Pearson, K. A., Palafox, L. and Griffith, C. A. (2017). Searching for exoplanets using artificial intelligence, *Monthly Notices of the Royal Astronomical Society* **474**(1): 478–491.

Press, W. H. (ed.) (1996). *FORTRAN numerical recipes*, 2nd ed edn, Cambridge University Press, Cambridge [England] ; New York.

Price, P. B. (2000). A habitat for psychrophiles in deep antarctic ice, *Proceedings of the National Academy of Sciences* **97**(3): 1247–1251.
URL: *http://www.pnas.org/content/97/3/1247*

Rice, K. (2014). The detection and characterization of extrasolar planets, *Challenges* **5**(2): 296–323.

Seager, S. (2013a). Exoplanet habitability, *Science* **340**(6132): 577–581.
URL: *http://science.sciencemag.org/content/340/6132/577*

Seager, S. (2013b). Exoplanet habitability, *Science* **340**(6132): 577–581.

Thompson, S. E., Mullally, F., Coughlin, J., Christiansen, J. L., Henze, C. E., Haas, M. R. and Burke, C. J. (2015). A machine learning technique to identify transit shaped signals, *The Astrophysical Journal* **812**(1): 46.