Identification and Classification of Exoplanets using Light Intensity.

MSc in Data Analytics - MSCDADSEP21AI

Research Project

Student Name: Aaditya Balkrishna Garude
Student ID: x20208596

Guide : Dr. Christian Horn

Abstract- The evolution of the astronomy field has been significantly impacted by science and technological innovation. Scientists have confirmed that there are more than a thousand exoplanets. The light curve, which is tiny and has uneven residual scattering, is defined by the brightness of the stars. The department of National Aeronautics and Space Administration (NASA) performed Kepler Mission and collected valuable insights in the form of data known as light curves which indicates brightness of stars. This data is in the form of Time series.Exoplanets were previously identified utilizing the transit methodology, which calls for human participation to analyze the signals associated to exoplanets. Therefore, automating a particular study is a crucial way for managing with huge Data that are generated by the most recent technology. It also helps to reduce human work. So, utilizing light intensity, we have presented a machine learning approach to finding exoplanets. Certain exploratory data analysis were performed to understand the data.The data then went through three Baseline Machine Learning models which gave undesirable results due to imbalance data. To overcome this imbalance nature of the Data SMOTE techniques was introduced which will help to balance the data and the identical Machine Learning models were applied again as a form of experiment two and desired outputs were achieved. To conclude, the results with and without the implementation of SMOTE technique are compared and it shows significant difference in getting better performance in terms of accuracy, confusion matrix, ROC and Area under the curve with the SMOTE technique.

Keywords- Exoplanets, Flux intensity, Machine Learning, SMOTE.

Contents

1 Introduction

Meteorites and radioactivity studies show that the solar system formed about 100 million years ago. This theory states that dust and gas are the main elements behind the formation of the entire solar system. Clouds formed by crystalline ice mixed with existing gravitational forces and later formed spinning disks containing most of the matter inside and rolled up to create planets and dwarf planets.Years of research and space exploration technology have also accelerated the urge to find extrasolar planets. These planets are called exoplanets. So what is an exoplanet? Exoplanets or exoplanets are defined as planets outside the solar system.The Kepler mission carried by NASA plays an vital role for increasing the amount of researches in knowing exoplanets and different factors affecting them. The planets outside or in the range of the sun are found by transit method. The methodology involved in transit method is calculating the drop of brightness and measuring the distance of the stars when the sun is at front of the planet which gives an fair idea of structure of the planet. Studies show that there are trillions of stars present in our galaxy. Around each of them are many solid surface planets. What fuels the study and discovery of these exoplanets is the urge to know what they look like?Are exoplanets circular? What kind of structure do these exoplanets have? Are exoplanets habitable? Imagine Earth in the same position as Pluto. After that, a small amount of sunlight would hit the surface, and the earth would freeze all the time.
Therefore, after an exoplanet is found to meet habitable conditions, further research will be conducted to know the physical conditions such as the star's temperature and atmospheric conditions. A second Kepler mission was conducted by NASA from 2014 to 2018. The mission explored hot, large planets to locate planets with habitable conditions. The time series Data gathered from the Kepler mission is suitable for carrying exploratory data analysis and applying certain Machine Learning and Deep Learning algorithms.

1.1 Motivation

Since many years, scholars have grappled with many curious questions such as: Is our planet Earth the only planet with life? Could life exist? These are questions that affect the deep research in the field of astronomy to find exoplanets. Astronomical studies show that the Milky Way galaxy contains hundreds of stars with at least one planet. The curiosity of finding same planet as earth is due to knowing the fact that there are number of stars mathematically there are trillion amounts of stars and also existence of trillions of galaxies in the solar system it is very much possible to have planet which has same environment as like earth. Hence by finding the exoplanets and studying about their types, size's and structure it will be easy to discover the habitual life of discovered planet. The idea of going to space and doing research to this extent is only possible because of advances in science and technology. There is also a wide range of technologies available for exploring this exoplanet. The main goal of discovering exoplanets is to find out if planets other than Earth are potentially habitable. While studying about the shape of exoplanet it is also assist in knowing Temperature, form of gases present, pressure in atmosphere, presence of Gravity which are important terms of research. Therefore, studying exoplanets indirectly brings us closer to having overall information of the planets including statistical factors and simultaneously knowing actual time occurrences.

1.2 Research Question and Objectives

The research work contains time series data.There appears to be 100x more stars without exoplanet. Any data gathered automatically will therefore be highly imbalanced,hence any pre-processing techniques would be must before applying models.Hence the idea is to use SMOTE as a pre-processing technique to compensate this imbalance nature.
Research Question:.

- **Can the performance of a system that uses only machine learning approach to find exoplanets using light intensity be outperformed by a system that combines SMOTE (Synthetic Minority Oversampling Technique) with machine learning approach?**

Objectives: Following are the objectives of the research:
The research is done by working on two Experiments.

- In the first experiment focus of the research is to work on imbalance data and note down the performance of each model.
- In the second experiment , Use of pre-processing technique is implemented called as SMOTE(Synthetic minority Oversampling Technique) to balance the imbalanced Data and avoid generating of duplicates.
- Comparing the Performance of the model with and without the use of SMOTE Technique.

1.3 Plan of Paper

In this research document, Starting with the discussion and motivation behind the research, In section 2- This section covers Literature review by studying previous research papers. Section 3- The main focus in this section is to propose and design new methodology and techniques. Section 4- Architectural Design of the project will be carried out right from the start to the outcome of the project.Section 5- This section includes implementation of the proposed methodology. Section 6- The implementation of research is done in the form of Two experiments, the first experiment will be carried by without SMOTE technique and second experiment includes pre-processing using SMOTE technique.Section 7- The results produced in both the experiments are discussed on the basis of evaluation parameters.Section 8- Limitation of applied methodology and future work will be studied.

2 Related Work

In this section, we will discuss about various research or Machine Learning methods executed for finding of exoplanets.The study of research papers will help in knowing the background of research and efforts needed to be taken for successfully carrying out research work.
The study of previous research is divided into four sections which are as follows:

1. Existing Research on Detection of Exoplanets.
2. Techniques for Extraction of Light curves.
3. Existing Pre-processing Techniques for Imbalance Data.
4. Previous Classification Techniques.

2.1 Existing Research on Detection of Exoplanets

(Malik and Obermeier, 2011) Time series analysis was used in the study. Light curve analysis has been done to complete the analysis. The TSFresh time series analysis package was used to extract features from light curves. Each light curve has 789 characteristics total that were retrieved. The attributes were eventually used to train the required Machine Learning tool, "lightGBM," a tree-based classifier. Simulated data was used to test this strategy, which turned out to be more effective than conventional Box least squares fitting (BLS). In comparison to the present state-of-the-art, this technique was effective in delivering results that were equivalent. It is also well arranged for powerful computers without the need for folding and secondary views of light curves. Shorter light curves make categorization more challenging, but their system was still able to do it withaccuracy of 98%.

(McCauliff et al., 2015) recommended that machine learning approaches can speed up the exoplanet detection process compared to traditional exoplanet detection methods that manually test light curves, but this is very time consuming. increase. In the study, an approach to star classification was developed using a random forest classification algorithm based on light curve variability. The method was based on the principle of a threshold-passing event, defined as a series of features in time-series analysis when a planet has crossed a target to a sufficient extent for further analysis. The main focus of the study was to classify processes into three classes: astronomical false positives (AFPs), planet candidates (PCs) and non-transiting

planets (NTPs). AFP is a TCE, not a real planet, but it has the same transit properties. PCs are the same TCEs that pass through exoplanets, and NTPs have been viewed as errors given by TCEs. An evaluation conducted by the study concluded that the Random Forest model was the best fit for the ML model compared to Naive Bayes.
(Armstrong, D. J., Pollacco, D. and Santerne, 2016) deployed the Self Organising Maps (SOM) unsupervised learning algorithm to find exoplanets. With this approach, the investigation was conducted using the Kepler mission's data. The input layer and output layer are two layers that make up the architecture of the SOM. It had employed competitive learning for SOMs, and weights were transmitted straight to the output layer. This approach is accomplished by giving each neuron in the SOM a weight vector, measuring the distance between each neuron in the output layer and input layer, and selecting the neuron with the smallest estimated distance as the best. This study showed that the SOMs are more rapid and precise in the process of removing false positives from exoplanets. This technique worked well and provided 90% accuracy.

2.2 Techniques for Extraction of Light curves

Search for neutral hydrogen on rocky earth like planets is carried in a research by (Santos et, al 2019).The presence of water in a lower atmosphere has a covered evidence by knowing that the extended neutral hydrogen exosphere exists around tiny exoplanets.This research still does not suits for all the rocky exoplanets.The main objective of the research is to detect neutral hydrogen exosphere of a planet similar to that of earth by transiting M dwarf employing Lyman-alpha spectroscopy and discuss important strategy for future studies.The research concluded by obtaining excess absorption in Lyman-alpha by using LUVOIR/LUMOS in M dwarf inside distance of 15pc.Also the analysis indicates that there is an possibility of detecting Earth-like planet by transiting TRAPPIST-1 inside 20 transits.
(Martinazzo and Hirata,2004)conducted a study using astronomical photometric surveys. In this strategy, numerous ConvNet architectures for celestial objects are evaluated. The implementation findings for three separate astronomical challenges were based on five well-known systems and five datasets. The following three datasets were used: detection of merging galaxies, categorization of stars and galaxies, and galaxy morphology. The technique is based on categorization of images. With this approach, the study is described by demonstrating how altering the regularization, optimizer, and training setup parameters for different architectures and datasets affects accuracy.The research came to the conclusion from their investigation that VGG-style architectures that had been trained on ImageNet performed better even on smaller datasets.
(Spiegel and Fortney, 2013) has studied the exoplanets' structures and the circumstances that impact them. The primary goal of this text is to examine the fundamental conditions, such as temperature, gases, density, etc., that are different on exoplanets. The following are key exoplanet fundamentals that are covered in this research: Gas giants are studied using the specifications of Jupiter and Saturn, which are enormous spheres of hydrogen and helium. Terrestrial and oceanic planets are then studied using the chemical composition of the elements Si, Mg, Fe, O, and C. For this research, Venus and Earth are more suitable candidates. Oceanic planets are then studied with a focus on the existence of life on the planet since water is a necessary component for the emergence of life.Mars' characteristics were more suited for research on marine exoplanets. The conclusion of this paper states that the recently approved Transiting Exoplanet Survey Satellite will identify a number of additional exoplanets that are located near stars that are bright enough to be seen by the James Webb Space Telescope.

2.3 Pre-Processing Techniques for Imbalanced Data.

The need of pre-processing technique for imbalanced data is explained by(Tallo and Musdholifah,2018).The pre-processing technique studied in this research was SMOTE (Synthetic Minority Oversampling Technique).The implementation of SMOTE lead to over generating of same classes because of having same instances in spite of distribution of class.By applying this technique the data overcomed the imbalanced masses by making simulated instances of minority classes.The research obtained significant better results with the application of SMOTE.
The research based on Refining exoplanet detection was carried by (Margarita Bugueno et.al, 2018)using supervised learning and Feature Engineering.The research is focused on producing results of each case analysis of light curves.The research suggested that the automatic techniques of extracting information from light

curves wast not satisfied due to complex problem.The feature engineering techniques used has dedicated problem with the execution time.The study concluded by showing that the metadata can give better results.
(Linderholm and Dreborg, 2015)has considered conducting research on the detection of exoplanets using the Support Vector Machine (SVM) and Convolutional Neural Network machine learning methods (CNN). The unbalanced dataset used in this study, which contains more stars that are not exoplanets than exoplanets themselves, is the primary issue. Mirroring the curves of stars with a rotating exoplanet and adding them to the collection solves the issue of unbalanced data. CNN outperformed the other machine learning models by accurately predicting curves that were both favorably and negatively labeled. Python has been used for all of the analyses.
(Amerongen et al., 2018)The available data creates problems due to irregularity of data. Pre-Processing of such data is necessary which was caught in the research carried on 3,000 stars by Kepler spacecraft.The document provides insightful details on using convolutional neural networks to detect exoplanets. This research main goal was to lay a low bar basis for astronomers interested in neural network research. By giving open source code, this research gave other researchers a solid foundation to build upon. In this study, two objects with the KIC IDs 2163434 and 2854994 were categorized from an investigation of 3,000 stars found by the Kepler spacecraft. The study reveals that the stars exhibit erratic brightness dips that support the hypothesis that exoplanets exist. The study finds that reliable exoplanet detection using convolutional Neural networks is definitely possible by using data reduction procedures or approaches.
(Cameron et al., 2019) In another project WASP (Wide Angle Search For Planets) data are the foundation of the study. The data was chosen because it has been successful in removing the negative effects brought on by erroneous positives. The main goal of this approach was to assess how well machine learning performed the same function as the previous transit survey data processing procedure without the assistance of a person. Convolutional neural networks and machine learning techniques were combined in this study to distinguish between various signals. The study came to the conclusion that the WASP data was insufficient for identifying planets since the nature of the data was unfavourable and it was impractical for a human observer to scrutinize each one individually in order to select the best candidate.The feature of this technique was that the algorithms could be retrained as soon as fresh categorization information became available. Additionally, the final algorithm 90% of the time successfully discriminated between discovered exoplanets and those of false positives.

2.4 Classification Techniques.

(Mena and Araya, 2018) has carried research of exoplanet detection using supervised learning and feature engineering. In this method, the study of irregular light curves was carried out, and it was proved that the motion of a celestial body in front of another celestial body gives answers about the existence of exoplanets in other solar systems. The results show normal results compared to other solar systems.Manually encountered features evaluated by scientists on exoplanet astronomy. The best-fitting model is a random forest that provides confirmed candidates (exoplanets), and a support vector machine model was used to identify false positives. The study concluded with a proposal to use the fitted Mundell-Agol light curve.
(Zhang and Zhao , 2015) have listed various data mining algorithms and data mining software and tools that are applied or used in the field of astronomy. The motivation behind this research was how important the fields of astrophysics and cosmoinformatics are in dealing with the big data case in the field of astronomy. They outlined the factors that influenced the success of the Sloan Digital Sky Survey (SDSS) project, known as the most successful astronomical survey in the history of astronomy and a frequent reference in several astronomy projects.The research is divided into his three parts related to data mining, data visualization, and data exploration of astronomical data. The second is the type of software that can be used to apply algorithms, storage, and computation. The third is the optimal algorithm and its effectiveness. The study concluded by stating that the astronomical data provide extensive data containing multi-wavelength features and time-series analyze from various sources. It can be used to manage well-characterized data and implementation software such as MS SQLserver, Orcale, MongoDB and PostgreSQL. It can also be used to integrate data where speed and accuracy are critical and to create easy-to-understand visualization algorithms. The study also mentions that Sloan Digital Sky Servers (SDSS) is the most advanced study and will later be further developed into SDSS-1 and SDSS-2 to support future research projects.

(McGovern and Wagstaff, 2011) had decided to apply machine learning to the data obtained for space applications. The primary goal of this research was to determine whether machine learning could be used to conduct space missions in an affordable, reliable and accurate manner. For research purposes, this survey is divided into several different sections. Places where people decide to do manual research can be too dangerous. Distance was a major concern as all exploration had to be done remotely. These are just a few of the challenges posed as machine learning challenges in the first part of space operations research. The second is an overview of the literature on artificial intelligence and machine learning in space today. In this part, we looked at some research papers to refer and understand previous studies. The third part gives quick hints about possible applications of machine learning in space, such as image analysis, time series analysis, clustering, and reinforcement learning. The study ended with these useful suggestions to fellow researchers for future research.
(Manry and Sturrock, 2019) have performed the analysis to develop a machine learning model that will help classify Kepler cumulative object of interest data. To integrate the data, train, and test the models throughout implementation, a machine learning pipeline was created. After taking into consideration missing values, inconsistencies, correlations, and bias, K-Nearest Neighbor, Support Vector, Machine, and Random Forest classifiers were employed as the machine learning models to provide a collection of candidates. Random Forest classifier has been chosen as the most appropriate model among these applied models. Following this, two deployment strategies were investigated: An API was first developed in Python using the Flask framework, but owing to uncertainty and a lack of stability for accuracy, a more reliable technology was built up on the Microsoft Azure cloud with higher stability.

Research Paper	Year	Model's Applied	Accuracy/Perfromance of Model	Description
Malik and Obermeier	2011	"lightGBM," a tree-based classifier	98%	The attributes were eventually used to train the required Machine Learning tool, "lightGBM," a tree-based classifier. Simulated data was used to test this strategy, which turned out to be more effective than conventional Box least squares fitting
McCauliff et al.	2015	Random Forest	90%	The method was based on the the principle of Threshold Crossing Event which is defined as sequence of feature in the time series analysis when the planet crosses the target at a particular degree which is sufficient for further analysis.
Santos et, al	2019	Survey Paper	Survey Paper	The main objective of the research is to detect neutral hydrogen exosphere of a planet similar to that of earth by transiting M dwarf employing Lyman-alpha spectroscopy and discuss important strategy for future studies
Tallo and Musdholifah	2018	Pre-Processing SMOTE Technique	Better Results	The implementation of SMOTE lead to over generating of same classes because of having same instances in spite of distribution of class.By applying this technique the data overcomed the imbalanced masses by making simulated instances of minority classes.The research obtained significant better results with the application of SMOTE.

Fig.1.Summary of Related Work

From the above literature review it is quite evident that there have been numerous efforts to identify or locate exoplanets based on the previous Related Work. Due to incomplete or unbalanced data, the study was not completed for several methodologies. The development of machine learning and deep learning algorithms has greatly aided most research. For the research of exoplanets, it has been crucial to take into account variables like speed, dependability, stability, and precision. When deciding whether to use machine learning and deep learning techniques to identify exoplanets, the data from NASA's Kepler spacecraft is crucially important.

3 Research Methodology

The data has to process through number of different phase to classify the data into exoplanets.The methodology includes two types standard medium called as KDD and CRISP-DM.The standard procedure followed in this research is Knowledge Discovery Database (KDD) which is best suitable methodology to be used with Machine Learning technique.The sections below provide a detailed explanation of the KDD Process's phases.

3.1 Data Selection

The data gathered from the source comprises of 5087 rows and 3198 columns. The Data is obtained in the form of Train data and Test data.The variables of Row defines the flux intensity produced from the specific time. According to the inverse square law the flux from an source present in space direct depends upon the intensity of light coming from the object. According to the inverse square law the flux of an astronomical origin depends upon the luminosity of an object and its distance from the planet earth. Mathematically Flux can be defined as watts per square metre.

The data was gathered from Kaggle under the topic of "Hunting of Deep space".
Data: https://www.kaggle.com/keplersmachines/kepler-labelled-time-series-data

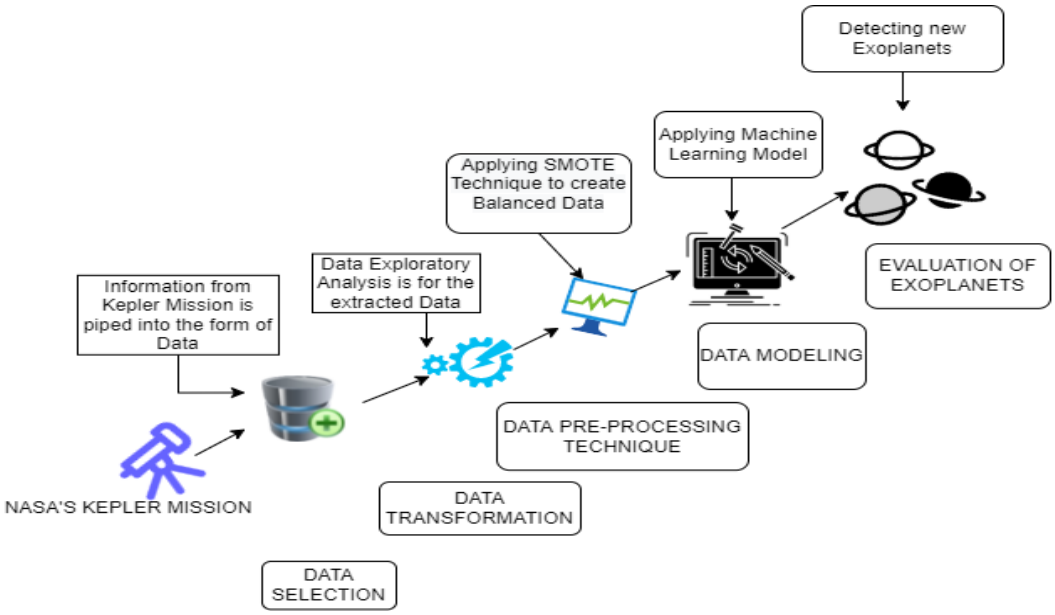

Fig.2. Finding Exoplanets using KDD Process

3.2 Transformation of Data

The Data Contains LABEL (2) as the presence of exoplanets and LABEL (1) as the non-presence of exoplanets.These values are converted into the terms of binary for easy understanding. As the data contains only 37 recordings of exoplanets and 5050 records of non-exoplanets which also hints that the data is highly imbalanced.Next is identification of null values which is visualized by heat map by using seaborn function.The heat map clearly states that there are no null values present in the data.The correlation matrix is implemented to understand the relationship between the straight line and the distance of data points which shows the frequency distribution of the Flux.As the data consists of time series it has lot of regular and irregular patterns.The Gaussian distribution is introduced which provides the normal distribution of curves.The outliers are studied using the boxplot and exploratory Data analysis is performed by using matplolib.pyplot library of python.The relationship between the variables is understood by implementing Scatter plot and pair plot.In Scatter plot the relationship between two FLUX are studied here which show scattered Data points of

both non-exoplanet and exoplanets, Where as in Pair plot it is possible to observe the distribution of single variable and also between the two variables.The pair plot distribution of five Flux intensities are studied. The relationship between two variables helps in observing the presence of non-exoplanets and exoplanets in a Data to understand the features in a much finer way.

3.3 Pre-Processing Techniques

While the transformation of Data was performed it showed that the data available was highly imbalanced which would definitely produce undesirable and irregularity results after the application of models. Hence to achieve the objective of this research it is important to pre-processed the data which is to be performed by using SMOTE (Synthetic minority over sampling method). By this technique it will be possible to make the data balanced by randomizing the data of training set equal or near to the test data. This will possibly increase the chances of getting desirable results and correct accuracy of the applied model. Putting a model together will make the data separated and can be utilized for prediction while keeping certain portions of the data in its original form. The original data and the anticipated data will then be compared. Performing hold out cross validation is another name for this technique. Because it gets rid of the over-fitting issue, validation is crucial. When training the model, validation is also utilized to evaluate it. Data After dividing the model, the penultimate step is pre-processing, at which point regression modeling techniques are applied.

3.4 Data Modeling

The data is further carried for model training, where Machine Learning models are applied on the split dataset.To compare the results and knowing the advantage of using SMOTE as a pre-processing technique, the Machine Learning models are applied by doing two experiments.In the First experiment the Machine Learning Algorithms will be applied directly without applying SMOTE Technique.
Following are the models discussed for implementation.

- Naive Bayes- (McCauliff et al., 2015) Naive Bayes algorithm is based on the Bayes theorem is a supervised learning method.The reason behind using Naive Bayes was its ability to work in real-world situation especially for document classification and spam filtering.
- Logistic Regression-Logistic Regression is one most simple and effective classification model which makes it best suited to apply on the data having Binary values.Logistic regression model function is give definite relation between continuous and discrete variables.(Manry and sturrock, 2019)
- Decision Tree-The cause of selecting Decision Tree algorithm is that it produces optimum results which proves to be successful for the classification models(Camporeale et al., 2019).
  Decision Tree model is imported by using sklearn.linear_model.

Followed by that in second experiment the same Machine Learning models are applied on the data which has been pre-processed by using SMOTE technique and the results will be noted for all the evaluation terms

3.5 Evaluation and Results

All the evaluation results are tested and noted for 1st experiment by observing the terms like Accuracy of the model, Recall value, Precision,F1 score ,significance of ROC curve and area under the curve.The Experiment two contains outputs of Models applied on the data with the pre-processing SMOTE technique.The objective of the research is thus achieved by comparing the results produce by the Machine Learning Algorithms with and without SMOTE technique.The evaluation parameters will provide details about how the models are performed on data in the Experiment one and two.

4 Design Specification

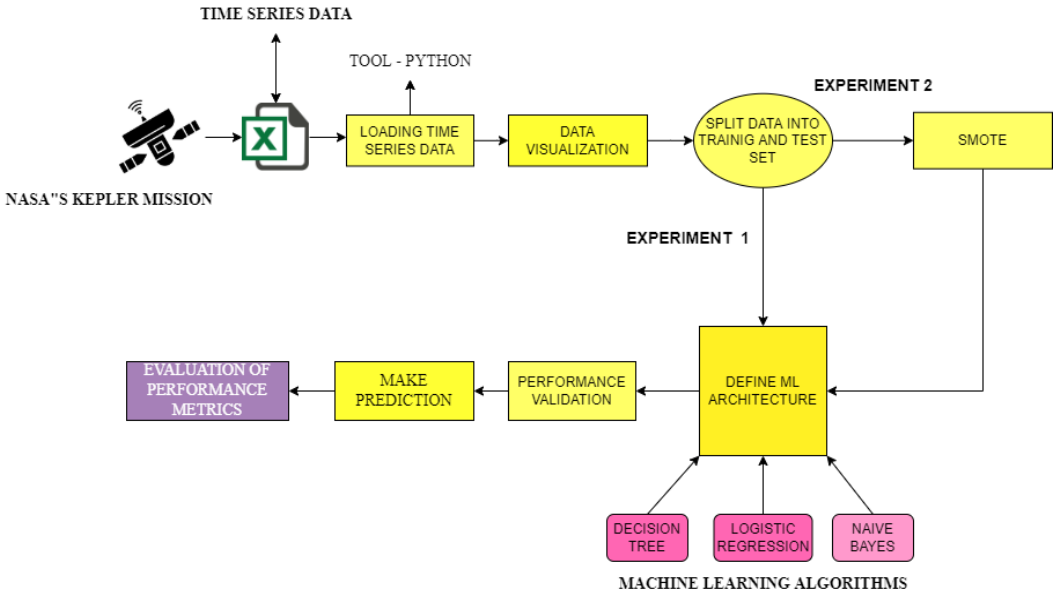

Fig.3. Architecture of classification of Exoplanets

The classification of exoplanet is a whole long process which is why the designing of the entire process is essential before implementation step. The data extracted is originally available in the form of Raw data.Firstly the data is taken in the mix form of exoplanet and non-exoplanet from NASA's Kepler mission.The next is converting it into .Csv format and data cleansing.The data containing large patterns makes it favourable for Data Exploratory analysis, Gaussian histogram is plotted for both non-exoplanet and exoplanet. The histogram is very unevenly distributed for both the data.The next step is running a Boxplot to visualize and find outliers in the data. To understand the potential relationship of data points scatter plot is plotted.The correlation between the two variables shows how much the data points are close to the imaginary straight line plotted.

The final steps includes the split of Data set into Train and Test Data.The Data Normalization is crucial technique operated before applying the model for data preparation.This step is useful in making the data into consistent state without misshaping the values.Feature scaling is implemented for standardizing of independent data points into a specified ranges.This is important step as skipping this step might cause the problem as algorithms be likely to weigh the higher and lower values different from the value unit.

Once the Machine Learning models are implemented the performance of the model is validated.Predictions are done while evaluating and differentiating using certain evaluate terms available such as Accuracy ,F1 Score, Confusion matrix, ROC and AUC curve.

5 Implementation

The implementation step of a research project is based on application of specific methods and algorithms to achieve the final result.In the present research project the implementation is research through definite steps such as Collection of data, Data Transformation,Pre-processing and reaching the implemention step,In the final modeling computation tool called open source python language of version 3.8.8 is used .Before application of models encoding specific important libraries like Keras, Tensorflow, matplotlib, Sklearn, Numpy, Seaborn,Pandas are imported.

5.1 Data Cleansing and Exploratory Data Analysis.

The research is carried by taking Data from open public source called Kaggle.The data is in the form of Time series containing numerical values which are referred as flux intensities captured at a particular phase of time.These Data was then imported into Jupyter notebook as a .Csv format and following exploratory Data analysis was done which are evaluated by visualizations.

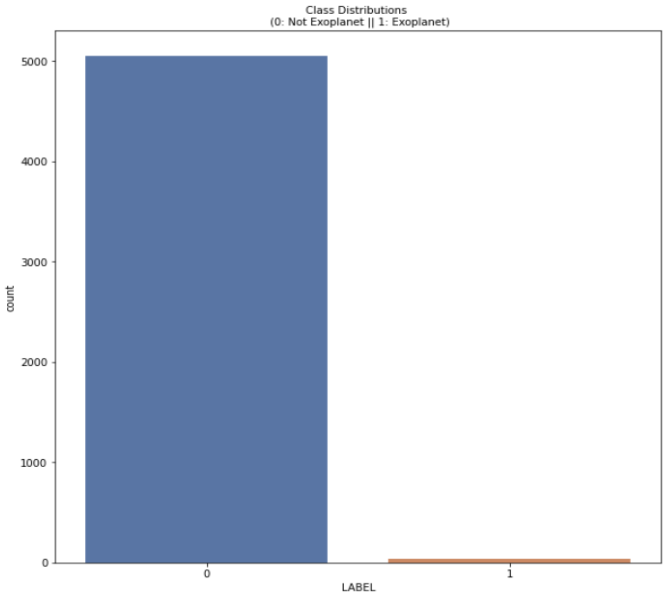

Fig.4. Distribution of Class

From the Fig.4 its quite evident that the Dataset is in the form of imbalance.

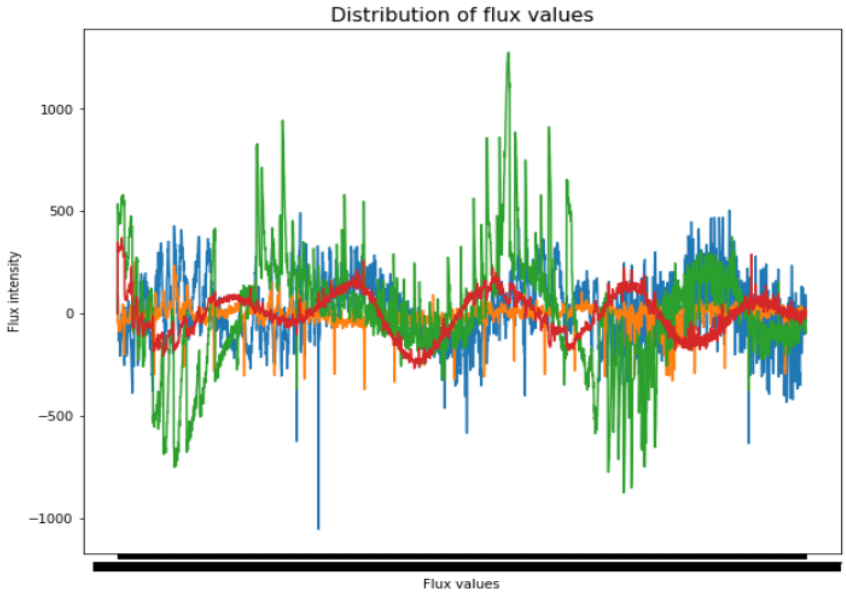

Fig.5. Distribution of Flux

The Data containing patterns of Flux intensity which are understood by visualization for first few rows.Now,the next step is learning about whether there are any null values present in the Data.From the observation it is evident that there is no presence of null values,hence no process of removing the any rows or column is necessary from the data.

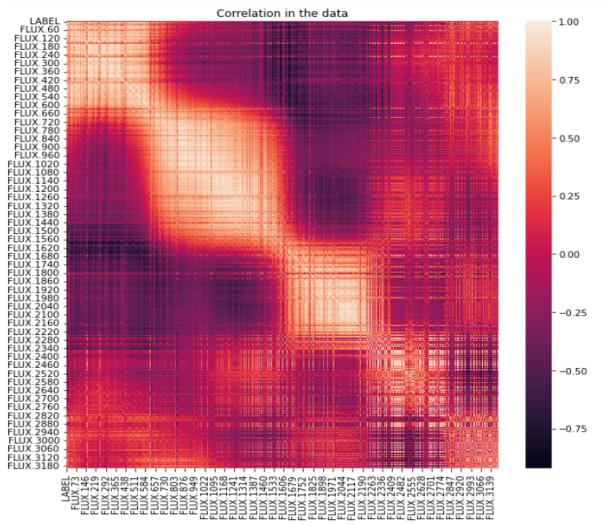

Fig.6.construction of Correlation Matrix

To understand the relationship between the straight line and the distance between data points, which displays the frequency distribution of the Flux, the correlation matrix is used.The frequency and direction of the linear interaction (straight line) between two quantitative variables are summed up by the correlation. Values are denoted by r and range from -1 to +1. A positive link results from a high value of r, and a negative link results from a low value of r. The more closely the data points are to a straight line and the closer r is to 1, the stronger the linear relationship. The linear relationship would be less strong the closer r gets to 0.
In order to understand the normal distribution of curves, the Gaussian distribution is introduced.
Here, the number of rows is defined by labels 1=[200,400,600]. The non-exoplanet graphs are shown below.

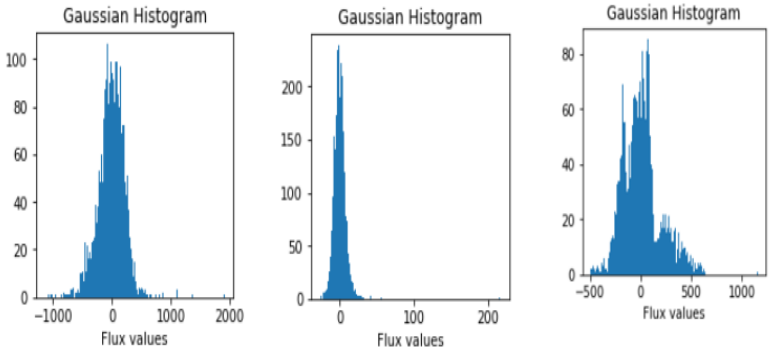

Fig.7.Gaussian Histogram for Non-Exoplanet

Here, the number of rows is defined by labels 1=[15,30,45].

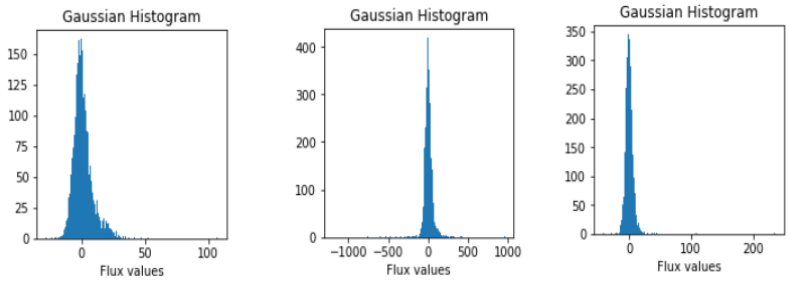

Fig.8.Gaussian Histogram for Exoplanet

5.1.1 Relation between variables

For comparing the relationship between two variables scatter plot is implemented.To compare the variables the Data points of random column FLUX 1 and FLUX 6 has beeb selected

```
sns.scatterplot(data=exo_train, x='FLUX.1', y='FLUX.6', hue='LABEL', palette=['b','r'])
plt.title('Relation of FLUX1 and FLUX6')
plt.show()
```

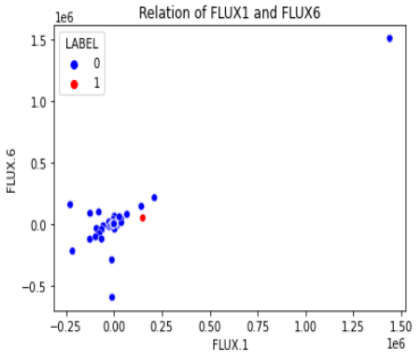

Fig.9.Construction of Scatter plot

From the Fig 9. It is evident to say that the if any imaginary straight line is plotted the data points are positioned very close to the line of both FLUX 1 and FLUX 6.The blue data points shows the presence of non-exoplanets and red data points shows that there is a exo-planet present in the data.

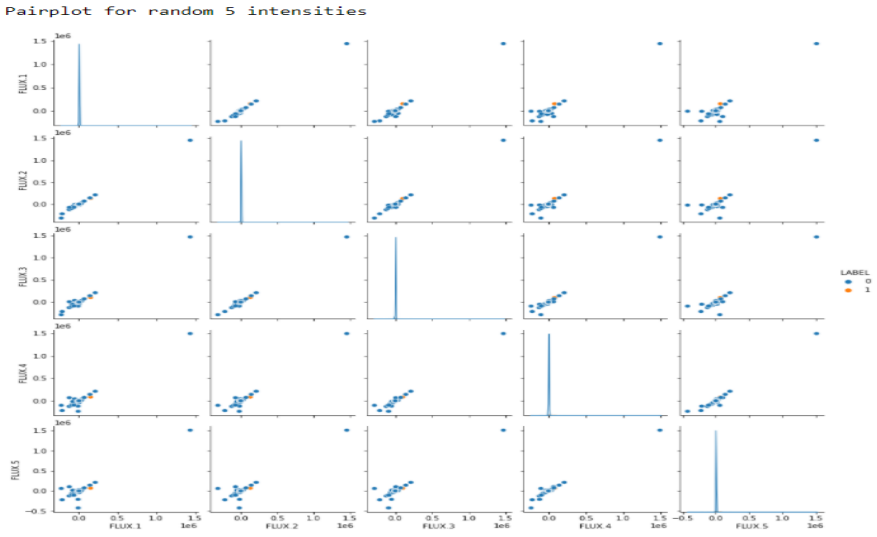

Fig.10.Construction of Pairplot

Pairplot is again a best way to observed relationship between two variables. The pairplot of first five rows has been executed.The Data points of non-exoplanet and exoplanet are marked closely.

5.1.2 Evaluation of Outliers

To know the range of the values present in the data,execution of outliers is necessary. Constructing Boxplot for detecting outliers for first three columns for FLUX 1 ,FLUX 2, FLUX 3.The insights from the constructed outlier for three columns is very indefinite to study.Also it does not provide suitable information about the value range.

```
#Detecting outliers using boxplot
fig, axes = plt.subplots(1, 3,figsize=(15, 6), sharey=True)
fig.suptitle('Distribution of FLUX')

sns.boxplot(ax=axes[0], data=exo_train, x='LABEL', y='FLUX.1',palette="Set2")
sns.boxplot(ax=axes[1], data=exo_train, x='LABEL', y='FLUX.2',palette="Set2")
sns.boxplot(ax=axes[2], data=exo_train, x='LABEL', y='FLUX.3',palette="Set2")
```

```
<AxesSubplot:xlabel='LABEL', ylabel='FLUX.3'>
```

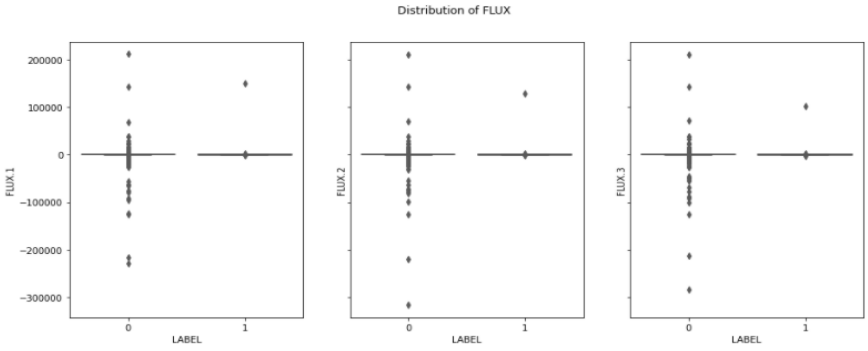

Fig.11 Evaluation of Outliers

Thus, for getting suitable result about the values in the data the solution can be done for studying values of just one row. The values of the data points are close and distinct enough and suitable for gathering information of the value range.

```
plt.boxplot(list(exo_train.loc[1]))
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x17e07a546a0>,
  <matplotlib.lines.Line2D at 0x17e03070220>],
 'caps': [<matplotlib.lines.Line2D at 0x17e19dff9d0>,
  <matplotlib.lines.Line2D at 0x17e19e19f10>],
 'boxes': [<matplotlib.lines.Line2D at 0x17e07a543d0>],
 'medians': [<matplotlib.lines.Line2D at 0x17e19e06310>],
 'fliers': [<matplotlib.lines.Line2D at 0x17e02f6b910>],
 'means': []}
```

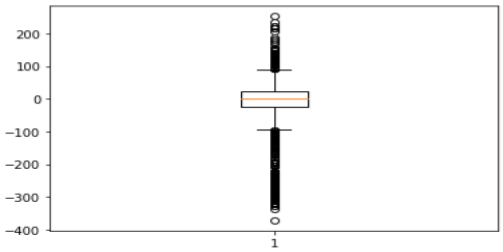

Fig.12 Outliers for single 'FLUX'

From the Boxplot constructed in Fig.13 we select value which ranges below 250000 because values are very close and distinct in this range.Hence the range by dropping value above 250000.

5.2 Data Normalization

```
#Data Normalization
x_train = normalized = normalize(x_train)
x_test = normalize(x_test)
```

```
#standardization of the Data for consistent values.
std_scaler = StandardScaler()
x_train = scaled = std_scaler.fit_transform(x_train)
x_test = std_scaler.fit_transform(x_test)
```

Fig.13.Normalization of Data

Before applying the Machine Learning model,the data is split and followed by the step of data preparation by normalization of the data used for normalization of columns for converting it into consistent scale within the data.The significant feature of continuous normal distribution for random sample is very much efficient. To keep the variables in the same range Feature Scaling is done.By inclusion of Feature scaling it is possible to standardized independent characteristics in specific range of data which helps model to give desirable result in contrast if not then the model would take irregular range of values that is higher and lower value.

5.3 Pre-processing Technique

Synthetic Minority Over-Sampling Method (SMOTE)
Classifiers are analyze by imbalanced datasets using the Synthetic Minority Oversampling Technique (SMOTE). This is a novel way to eliminate data that are unbalanced.Theoretically, N is constructed for total no of samplings such that the binary class holds 1:1 distribution.The process begin by selecting Positive class instances randomly.The KNNs for that instance are then obtained (there are 5 by default). Finally, N is picked from among these K instances to interpolate fresh synthetic instances. To do that, the distance between the feature vector and its neighbors is determined using any distance metric. Now, the preceding feature vector is increased by multiplying this difference by any random value in the range of (0,1).As the efficiency of the model is reduced and the overfitting issue caused by random oversampling is resolved, it is feasible to improve the model's accuracy by applying the SMOTE algorithm and the boosting technique. Highly unbalanced data was gathered from Kaggle. 37 of the false positives that were found in the data (5050 total) were exoplanets. A issue of overfitting and decreased accuracy would result as a result. So that the data is balanced and we can get more accurate results.

6 Results and Evaluation

This is the final step in the research where the execution of Machine Learning Algorithms which were discussed earlier are applied.The results produced by the Models are hence observed by certain parameters called as evaluation of the results.This are standard evaluation terms by which it is feasible to discuss performance and knowing what has been produced by the applied model. Following are the Evaluation terms which will used to discuss the performance.

1. **Accuracy of the Model**:Accuracy refers to acquired percentage of accurate prediction of the test data.Mathematically accuracy is defined as ratio of Correct prediction to the total number of prediction.ie:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Where

- TP : True Positives
- FP : False Positives
- TN : True Negatives
- FN : False Negatives

2. **Precision**:Precision is defined as ratio of suitable examples(True positive) to the sum of all predicted examples of the same class.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

3. **Recall**:Recall refers to the fraction of positive cases identified to the sum of all the correct prediction

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

4. **F1 Score**:F1 score perform as balance between Recall and precision.It provides accuracy of the model by taking the average values from the Precision and Recall as it appraises both values from truth table.Mathematically it is given by fraction of two times multiple of Recall and precision to the sum of Precision and Recall.

5. **ROC and Area under the curve**:The ROC (Receiver Operator Characteristic)curve act as a evaluation term for classification problems.It is visualized as a graph of False positive rate vs True positive rate showing the performance of the model.Whereas Area under the curve provides the summary of the ROC curve. AUC works best under 0.5 <AUC <1 as it is able to distinguish Positive and Negative class.

6.1 Experiment 1:Execution and Evaluation of Model without SMOTE.

In Experiment 1, The execution of Three Machine Learning algorithms is done which are 1)Naive Bayes,2) Logistic regression and 3) Decision Tree.

1)Naive Bayes-The first model is supervised Machine Learning model called Naive Bayes .

```
Accuracy of Naive Bayes is 0.9719298245614035

Classification report :
              precision    recall  f1-score   support

           0       1.00      0.98      0.99       565
           1       0.18      0.60      0.27         5

    accuracy                           0.97       570
   macro avg       0.59      0.79      0.63       570
weighted avg       0.99      0.97      0.98       570

Text(0.5, 1.0, 'ROC - CURVE & AREA UNDER CURVE')
```

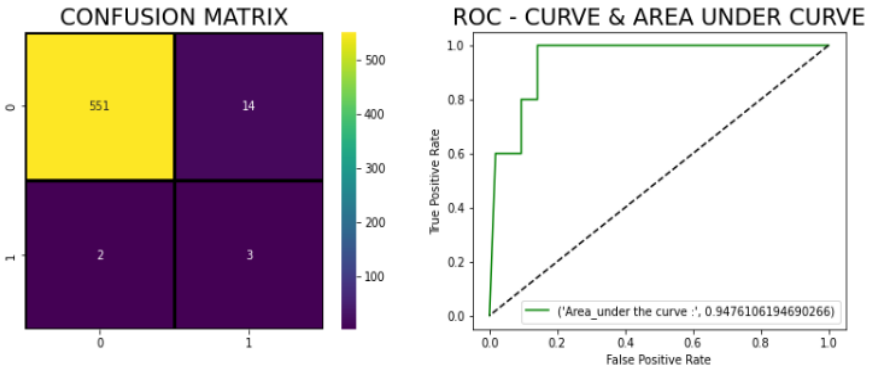

Fig.14.Naive Bayes Confusion Matrix and Classification Report

2)Logistic Regression-The second model applied is supervised Machine Learning model called KNN.

```
Accuracy of Logistic Regression is 0.9912280701754386

Classification report :
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       565
           1       0.00      0.00      0.00         5

    accuracy                           0.99       570
   macro avg       0.50      0.50      0.50       570
weighted avg       0.98      0.99      0.99       570

Text(0.5, 1.0, 'ROC - CURVE & AREA UNDER CURVE')
```

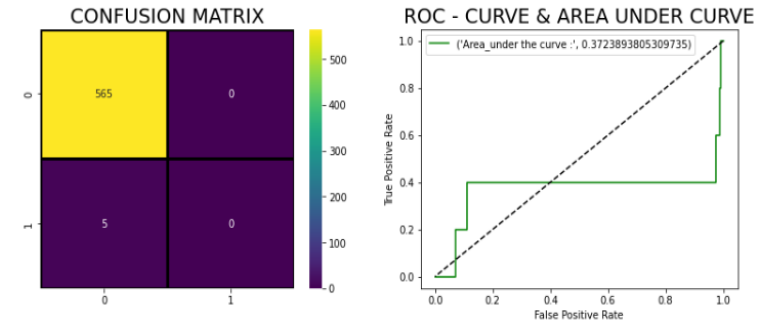

Fig.15.KNN Confusion Matrix and Classification Report

3)Decision Tree-The Third and the last model applied is supervised Machine Learning model called Decision Tree.

```
Validation accuracy of Decision Tree is 0.9894736842105263

Classification report :
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       565
           1       0.33      0.20      0.25         5

    accuracy                           0.99       570
   macro avg       0.66      0.60      0.62       570
weighted avg       0.99      0.99      0.99       570

Text(0.5, 1.0, 'ROC - CURVE & AREA UNDER CURVE')
```

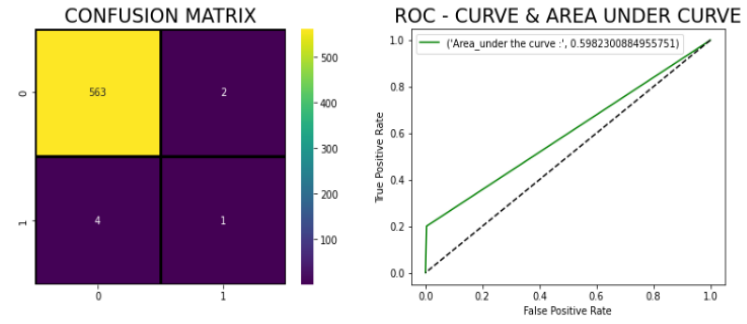

Fig.16.Decision Tree Confusion Matrix and Classification Report

The execution of all three decided models has been done and above results were produced by each model.Here the Models have been performed before the pre-processing technique of SMOTE.The results produced by each models is undesirable and not accurate enough to provide conclusion for the result.The accuracy of the Naive Bayes is 99%,whereas the accuracy of the Logistic Regression is 98% and that of Decision treee is 96%. All the three applied models does not produce satisfactory output this is due to imbalance nature of the data. The model seems to be overfitted with close to 100% accuracy.Thus there must be some application of technique to balance the Data,hence we introduced a technique called Synthetic Minority Oversampling Technique(SMOTE).The execution and results of the models after SMOTE will be carried in Experiment 2.

Pre-processing using SMOTE As discussed earlier, the experiment two will be executed by making the data balanced by using SMOTE as an pre-processing Technique. SMOTE's unique characteristic is that it avoids producing duplicate data points, instead focusing on creation of Synthetic Data.
In the data we have 5050 number of Non-exoplanets and just minimum amount that is 37 of Exoplanets. Following is the implementation of SMOTE algorithm on train data.

```
: #SMOTE techinque for Balancing of the imbalanced Data.
from imblearn.over_sampling import SMOTE
model = SMOTE()
ov_train_x,ov_train_y = model.fit_resample(exo_train.drop('LABEL',axis=1), exo_train['LABEL'])
ov_train_y = ov_train_y.astype('int')
ov_train_y.value_counts().reset_index().plot(kind='bar', x='index', y='LABEL')
```

Fig.17.Applying SMOTE Technique

.
The SMOTE application is successful which can be evaluated via graphical form of visualization.

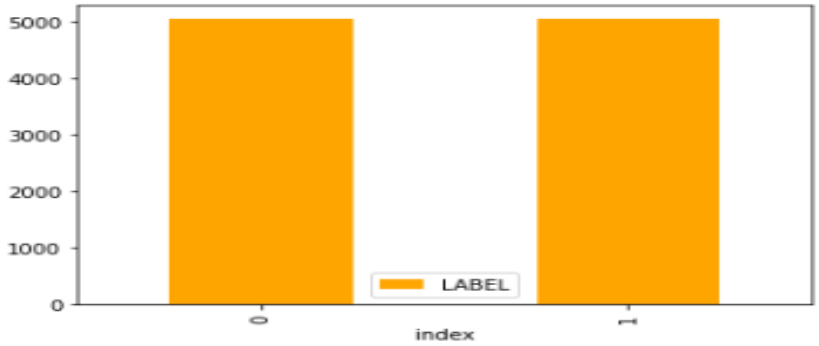

Fig.18.Visualization of applied SMOTE technique

6.2 Experiment 2:Execution and Evaluation of Model After SMOTE.

In Experiment 2, Synthesization of new examples is done by using SMOTE by randomly selecting examples.
1)Naive Bayes-

```
Accuracy of Naive Bayes is 0.48934893489348935

Classification report :
              precision    recall  f1-score   support

           0       0.53      0.04      0.07      1709
           1       0.49      0.97      0.65      1624

    accuracy                           0.49      3333
   macro avg       0.51      0.50      0.36      3333
weighted avg       0.51      0.49      0.35      3333

: Text(0.5, 1.0, 'ROC - CURVE & AREA UNDER CURVE')
```

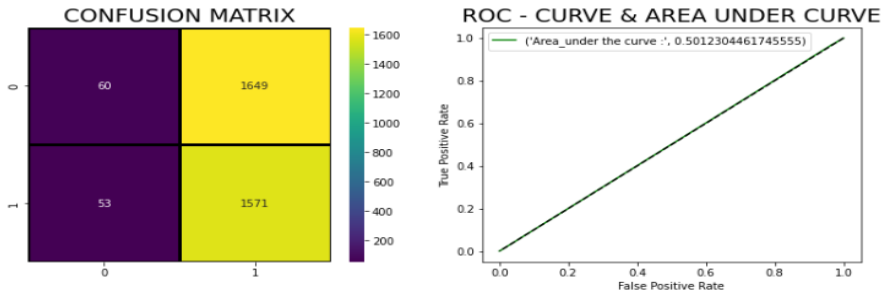

Fig.19.Naive Bayes Confusion Matrix and Classification Report after applying SMOTE

2)Logistic Regression-

```
Accuracy of Logistic Regression is 0.6945694569456946

Classification report :
              precision    recall  f1-score   support

           0       0.72      0.66      0.69      1709
           1       0.67      0.73      0.70      1624

    accuracy                           0.69      3333
   macro avg       0.70      0.70      0.69      3333
weighted avg       0.70      0.69      0.69      3333

Text(0.5, 1.0, 'ROC - CURVE & AREA UNDER CURVE')
```

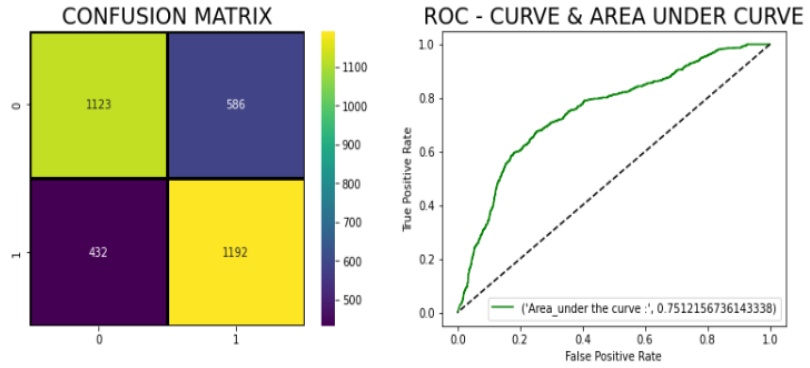

Fig.20.Logistic Regression Confusion Matrix and Classification Report after applying SMOTE

3)Decision Tree-

```
Validation accuracy of Decision Tree is 0.9366936693669367

Classification report :
              precision    recall  f1-score   support

           0       0.98      0.89      0.94      1709
           1       0.90      0.99      0.94      1624

    accuracy                           0.94      3333
   macro avg       0.94      0.94      0.94      3333
weighted avg       0.94      0.94      0.94      3333

: Text(0.5, 1.0, 'ROC - CURVE & AREA UNDER CURVE')
```

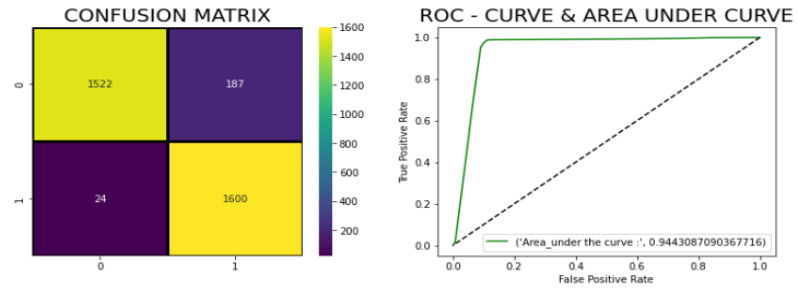

Fig.21.Decision Tree Confusion Matrix and Classification Report after applying SMOTE

The implementation process of previously executed model is carried out again by using Balanced data.The data went under the pre-processing step this time by using SMOTE technique.The results produced by models with and without SMOTE is discussed in the next section.Here all the values produced by models are observed and compared on the basis of evaluation parameters.

7 Discussion

The comparison of models with and without the SMOTE in terms of Evaluation parameters is done below which gives better understanding about the performance of the model.The parameters are Accuracy of the model, Precision value, Recall value, F1 Score and ROC AUC curve explanation.

MODELS	Model without SMOTE	Non-exoplanet(0)	Exo-planet (1)	Model with SMOTE	Non-exoplanet(0)	Exoplanet(1)
Naive Bayes	Balanced Accuracy - 97%			Balanced Accuracy - 48%		
	Precision Value-	0.99	0.18	Precision Value	0.53	0.49
	Recall Value-	0.98	0.67	Recall Value-	0.3	0.97
	F1 Score-	0.99	0.27	F1 Score-	0.7	0.65
	ROC & AUC Curve	No Difference		ROC & AUC Curve	Straight Line with balance True positive and False Positive	
Logistic Regression	Balanced Accuracy - 99%			Balanced Accuracy - 69%		
	Precision Value-	0.99	0	Precision Value	0.72	0.67
	Recall Value-	1	0	Recall Value-	0.66	0.73
	F1 Score-	1	0	F1 Score-	0.69	0.7
	ROC & AUC Curve	No Difference		ROC & AUC Curve	Curve shows Exopanets and non-exoplanets are predicted correctly	
Decision Tree	Balanced Accuracy - 98%			Balanced Accuracy - 91%		
	Precision Value-	0.99	0.3	Precision Value	0.98	0.9
	Recall Value-	1	0.2	Recall Value-	0.89	0.99
	F1 Score-	0.99	0.2	F1 Score-	0.94	0.94
	ROC & AUC Curve	No Difference		ROC & AUC Curve	Curve shows Exopanets and non-exoplanets are predicted correctly	

Fig.22.Comparing the performance of the model with and without SMOTE Technique

From the Fig.22 it is very much evident that the performance of all the models have been improved after the addition of SMOTE technique.The results are very desirable and all the evaluation parameters are satisfied.There is no sign of overfittiing of the model has been occurred.The balanced accuracy of Naive Bayes was 97% in experiment one which was done without SMOTE Technique.The performance of the model was fairly good with 48% of balanced accuracy with the SMOTE technique, Whereas by applying Logistic Regression the accuracy of the model becomes overfitted with 99% accuracy and moderately good after SMOTE which is 69%.The third model applied was Decision Tree which proves same behaviour as Logistic Regression before smote but after SMOTE technique it has came as a best suited model with balanced accuracy of 91%.The ROC and AUC curve shows that the non-exoplanets and exoplanets are differentiated at least to the mid level of the observation where the curve falling more towards the true positive rate in most of the times.Also we can conclude that the Decision Tree is best suited model for finding exoplanets as compared to other two models.

8 Conclusion and Future Work

In today's technological world curiosity of future prediction by using present data has became integral element of regular working exercises.This has given rise to building of Machine Learning and Deep Learning Algorithms which are suitable for prediction and analysis by using Data.Going back to the research question for finding exoplanets by Machine Learning approach using SMOTE or without SMOTE has definitely a answer that a system that uses SMOTE completely outperformed the system that do not uses SMOTE Technique as a pre-processing technique.The evaluation terms are evident of this research that is successfully done by using SMOTE technique and applying Three Machine Learning models.The Limitation of this research was use of small data after applying SMOTE technique and applying models.So basically by using the huge data would increase the Performance of the models.The astrobiology is big concept having huge data so it is important to have a system that would carry the whole data for the research.In Future work more visualization can been carried out for Time series data.Experiments can be done by using Deep Learning models by using FFT as a pre-processing technique.

Acknowledgement.

I want to express my gratitude to Prof. Christian Horn for his constant support, guidance, and assistance with the research project. I also want to convey my appreciation to all journalists and activists who have worked for the enhancement of Astrobiology and Exobiology.

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. Mcgovern and K. Wagstaff, "Machine learning in space: Extending our reach," Machine Learning, vol. 84, pp. 335–340, 09 2011.

Asif Amin, R.M., Talha Khan, A., Raisa, Z.T., Chisty, N., SamihaKhan, S., Khaja, M.S. and Rahman, R.M. (2018). Detection of Exoplanet Systems In Kepler Light Curves Using Adaptive Neuro-Fuzzy System. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8710502 [Accessed 10 Aug. 2022].

Baron, D. (2019). Machine Learning in Astronomy: a practical overview. [online] NASA ADS. Available at: https://ui.adsabs.harvard.edu/abs/2019arXiv190407248B/abstract [Accessed 10 Aug. 2022].

Bugueno, M., Mena, F. and Araya, M. (2018). Refining Exoplanet Detection Using Supervised Learning and Feature Engineering. 2018 XLIV Latin American Computer Conference (CLEI).

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. Space Weather, 17(8), pp.1166–1207.

Carrasco-Davis, R., Cabrera-Vives, G., F¨orster, F., Est´evez, P.A., Huijse, P., Protopapas, P., Reyes, I., Mart´ınez-Palomera, J. and Donoso, C. (2019). Deep Learning for Image Sequence Classification of Astronomical Events. Publications of the Astronomical Society of the Pacific, [online] 131(1004), p.108006. Available at: https://arxiv.org/abs/1807.03869 [Accessed 10 Aug. 2022].

Hendriks, L. and Aerts, C. (2019). Deep Learning Applied to the Asteroseismic Modeling of Stars with Coherent Oscillation Modes. Publications of the Astronomical Society of the Pacific, 131(1004), p.108001.

Kothari, V., Liberis, E. and Lane, N.D. (2020). The Final Frontier: Deep Learning in Space. arXiv:2001.10362 [cs, eess]. [online] Available at: https://arxiv.org/abs/2001.10362 [Accessed 10 Aug. 2022].

Malik, A., Moster, B.P. and Obermeier, C. (2011). Exoplanet detection using machine learning. Monthly Notices of the Royal Astronomical Society.

Martinazzo, A., Espadoto, M. and Hirata, N. (n.d.). Self-supervised Learning for Astronomical Image Classification. [online] Available at: https://arxiv.org/pdf/2004.11336.pdf [Accessed 10 Aug. 2022].

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 OFKEPLERPLANETARY TRANSIT CANDIDATES. The Astrophysical Journal, 806(1), p.6.

Santos, L.A. dos, Bourrier, V., Ehrenreich, D. and Kameda, S. (2019). Observability of hydrogen-rich exospheres in Earth-like exoplanets. Astronomy Astrophysics, [online] 622, p.A46. doi:10.1051/0004-6361/201833392.

Schanche, N., Cameron, A.C., H´ebrard, G., Nielsen, L., Triaud, A.H.M.J., Almenara, J.M., Alsubai, K.A., Anderson, D.R., Armstrong, D.J., Barros, S.C.C., Bouchy, F., Boumis, P., Brown, D.J.A., Faedi, F., Hay, K., Hebb, L., Kiefer, F., Mancini, L., Maxted, P.F.L. and Palle, E. (2019). Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys. Monthly Notices of the Royal Astronomical Society, [online] 483(4), pp.5534–5547. Available at: https://academic.oup.com/mnras/article/483/ [Accessed 10 Aug 2020].

Shallue, C.J. and Vanderburg, A. (2018). Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90. The Astronomical Journal, 155(2), p.94.

Sofia Dreborg,Maja Linderholm,Jacob Tiensuu,Fredrik Orn.(2019).Detecting exo- ¨ planets with machine learning. (n.d.). [online] Available at: http://www.diva-portal.org/smash/get/diva2:1325376/FULLTEXT01.pdf [Accessed 10 Aug. 2022].

Spiegel, D.S., Fortney, J.J. and Sotin, C. (2013). Structure of exoplanets. Proceedings of the National Academy of Sciences, 111(35), pp.12622–12627.

Tallo, T.E. and Musdholifah, A. (2018). The Implementation of Genetic Algorithm in Smote (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem. [online] IEEE Xplore. doi:10.1109/ICSTC.2018.8528591.

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.

Timpe, M.L., Han Veiga, M., Knabenhans, M., Stadel, J. and Marelli, S. (2020). Machine learning applied to simulations of collisions between rotating, differentiated planets. Computational Astrophysics and Cosmology, 7(1).

Van Amerongen, P. (2018). Discovering exoplanets using Convolutional Neural Networks. [online] fse.studenttheses.ub.rug.nl. Available at: https://fse.studenttheses.ub.rug.nl/17826/ [Accessed 10 Aug. 2022].

Zhang, Y. and Zhao, Y. (2015). Astronomy in the Big Data Era. Data Science Journal, 14(0), p.11.