Analyzing the Properties and Relations of Exoplanets: NASA Kepler Mission

Kazım Emre Yüksel
Mathematics Engineering Department
Istanbul Technical University
Istanbul, Turkey
emreeyukseel@gmail.com

Beyza Nur Kebeli
Mathematics Engineering Department
Istanbul Technical University
Istanbul, Turkey
beyzanurkebeli@gmail.com

İpek Korkmaz
Mathematics Engineering Department
Istanbul Technical University
Istanbul, Turkey
ipekkorkmz@gmail.com

***Abstract*—Different studies have been conducted for many years for the discovery of exoplanets. The Kepler Mission, which used Kepler telescope, was started with the idea of searching habitable zones inside the Milky Way and was successful at finding many stars and various exoplanets. In this study, we analyzed the KOI and luminous flux changing data obtained from the Kepler telescope to determine the characteristics of stars of different statuses. We studied the features of candidate and false positive exoplanets to see the similarities and differences between the two. Using these features, we developed a model that predicts confirmed and unconfirmed planets.**

I. Introduction

Any planet that is not in our solar system is called an exoplanet. Some of them are free-floating exoplanets which are called rogue planets, but most of them orbit other star or stars. Discovered exoplanets are mostly in a relatively small region of the Milky Way, but from NASA's Kepler Space Telescope, it is known that number of planets are more than stars in our galaxy [1].

There are five methods are commonly used for finding an exoplanet:

- Radial velocity: A planet which orbits a star, exerts gravitational force to its star and this causes the star to wobble and due to wobbling, scientists observe changing in the length of the waves emitted by the star. With this method, 826 exoplanets are discovered.
- Transit: When observed star's light dims due to transition of an exoplanet, this slight changing gives astronomers an information whether there is an exoplanet orbits the star. Amount of blocked light informs us about exoplanet's size. Also, period of transit can tell us distance between the exoplanet and its star. With this method, 3294 exoplanets are discovered.
- Direct Imaging: Planets can be photographed directly, when its star's light blocks by several techniques. With this method, 51 exoplanets are discovered.
- Gravitational Microlensing: When a planet passes by, because of the gravitational force of the planet, the light that the star radiates gets refracted. With this method, 106 exoplanets are discovered.
- Astrometry: Astronomers can detect the star's position in the sky. By taking series of images and then comparing them if the star moved or not, they can suspect that this changed position signal of an exoplanet. With this method, 1 exoplanet is discovered.

The Kepler space telescope launched for searching exoplanets by using one of the most successful methods, transit from 2009-2013. Its duty is called the Kepler mission. [2].

We mainly studied with exoplanets that were found using the Transit Method. We have studied the relationship between candidate and false positive exoplanets and realized some differences and similarities.

II. Literature Research

There are a plethora of papers about detecting and exploring exoplanets in literature. Suzanne, A. Favata, F. [3] present a novel algorithm based on a Bayesian approach. The algorithm is based on the Gregory-Loredo method originally developed for the detection of pulsars in X-ray data. In the present paper, the algorithm is presented, and its performance on simulated data sets dominated by photon noise is explored. Batalha, N. M. [4] present the exoplanets found so far, their statistical analysis according to various feature and their important properties. They characterized hundreds of planets over a large range of sizes and compositions for both single- and multiple stars systems. Another paper published by Shallue, C. J. and Vanderburg, A. present a method for classifying a potential planet signals using a deep convolutional neural network to predict whether a given signal is a transiting exoplanet or a false positive caused by astrophysical or instrumental phenomena [5]. Batalha et al. [5] make a test for determining false positive planets that are essential of our data analysis used to determine the disposition type of the KepID's.

III. Dataset Description

A. Cumulative KOI Data

One of the datasets we have used is Cumulative KOI Data from the NASA Exoplanet Archive. NASA Exoplanet Archive is a catalog and a data service that collects and obtains data and information about exoplanets and their host stars. These data sets have stellar and exoplanet parameters and discovery/characterization information about the exoplanets. Data in this archive is obtained from missions like the Kepler Mission and CoRoT. The NASA Exoplanet Archive includes

planets for which the planetary and orbital properties are publicly available, usually through refereed publications [7].

Kepler Mission: The Kepler Mission, which launched on March 6, 2009, was the first space mission that was dedicated to the exploration of exoplanets. Its goal was to find Earth-size and smaller planets and also habitable planets. Kepler was a special-purpose spacecraft that precisely measured the light variations from thousands of distant stars, looking for planetary transits. Exoplanets found from this mission were found using the transit method.

The Cumulative KOI Dataset has many columns but we worked only with the most significant columns and omitted the ones we do not need. We worked with a dataset with 9564 rows and more than 50 columns. This data set also included two error columns for some properties. Some of the important columns and their meanings that are obtained from the Archive are [8]:

- **kepid**: Target identification number, as listed in the Kepler Input Catalog (KIC).
- **kepoi_name**: A number used to identify and track a Kepler Object of Interest (KOI).
- **koi_disposition**: The category of this KOI from the Exoplanet Archive. Current values are CANDIDATE, FALSE POSITIVE, NOT DISPOSITIONED or CONFIRMED.
- **koi_pdisposition**: The pipeline flag that designates the most probable physical explanation of the KOI. Typical values are FALSE POSITIVE, NOT DISPOSITIONED, and CANDIDATE.
- **koi_score**: A value between 0 and 1 that indicates the confidence in the KOI disposition.
- **koi_period**: Orbital period, given in days.
- **koi_impact**: The sky-projected distance between the center of the stellar disc and the center of the planet disc at conjunction, normalized by the stellar radius.
- **koi_prad**: Planetary Radius, given in Earth radii.
- **koi_slogg**: Stellar Surface Gravity, given in log10(cm s-2.
- **koi_srad**: Stellar Radius, given in solar radii.
- **koi_smass**: Stellar Mass, given in solar mass.
- **koi_depth**: The fraction of stellar flux lost at the minimum of the planetary transit.
- **koi_steff**: The photospheric temperature of the star.
- **koi_teq**: Approximation for the temperature of the planet.
- **koi_duration**: Transit Duration (hours).

B. Flux Changing Dataset

The data describe the change in flux (light intensity) of several thousand stars. The planets themselves cannot emit light, but the stars can. When these stars are examined in certain periods, there may be a regular extinction of the flux. This is proof that the star could be an object orbiting it. Therefore, such a star can be considered as a candidate star. So we wanted to examine data containing the changing in luminous flux over a period in confirmed exoplanet-stars and non-exoplanet-stars. Therefore, it is important to either detecting and imaging exoplanets. S.B Wright, T. J., Gaudi, S. B. [9] point out the various methods to detect exoplanets and explain the ways to image the planet's planet/star flux ratio and the angular separation between the planet and star in their paper.

The datasets were originally designed in 2016. This dataset is a cleaned version of the original form which can be available NASA Archive. This project consists of various mission and 99% of this dataset is taken from campaign 13. [10]

The data consist of 5087 rows and 3198 columns. Columns represent the different flux values recorded at a different time while rows represent exoplanets which are both confirmed and non-confirmed. 37 of these stars are confirmed and 5050 are unconfirmed stars. Thus, we have an unbalanced dataset.

IV. Analysis and Results

A. Analysis of KOI Data

We have analyzed the KOI Data set. In this data set, while 5068 of the rows are FALSE POSITIVE, 4496 of them are CANDIDATE.

To know whether a planet is an exoplanet or not, scientists look at four methods for classification and they are four columns in our data set called: $fpflag\_nt$, $fpflag\_ss$, $fpflag\_co$ and $fpflag\_ec$.

If at least one of these columns has 1 as a value, the planet is most likely FALSE POSITIVE. However, Hot Jupiters (a type of exoplanets that are physically similar to Jupiter) also have 1 as a value in $fpflag\_ss$ column, as seen in Figure 1. Even all of these columns have 0 as value, some can still be FALSE POSITIVE, as seen in Figure 2.

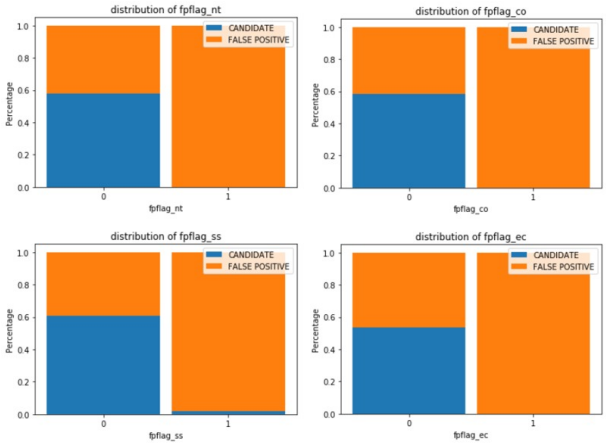


Fig. 1: Flag values in percentage

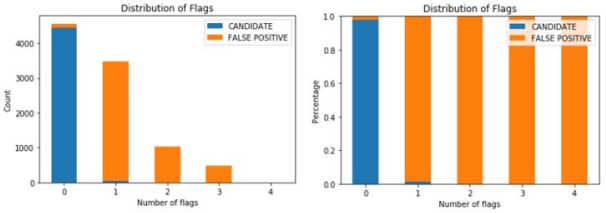


Fig. 2: Count of flags and the percentage

We looked at the density graphs of all columns between FALSE POSITIVE and CANDIDATE exoplanets and realized how some of the properties are significant and some are insignificant.

Columns koi_depth, koi_impact, koi_prad, and koi_teq showed very different patterns. Therefore, we think they are significant properties to decide whether a suspected planet is an exoplanet or not.

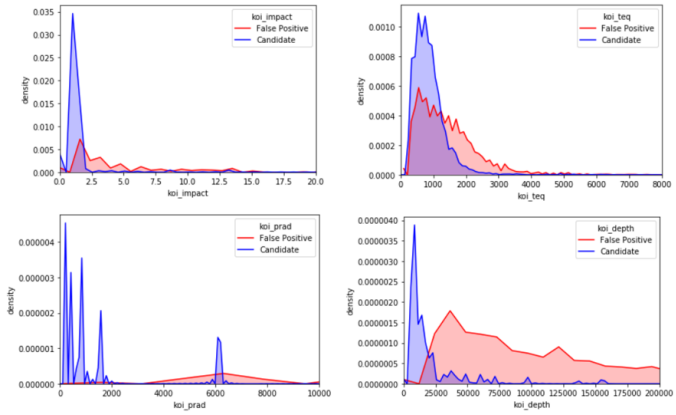


Fig. 3: Density graphs of given columns

Columns $koi\_duration$, koi_steff, and koi_slogg showed very similar patterns, which can mean that these properties are not so significant. Since koi_steff and koi_slogg are the properties of the star and not related to the planet in question, this can be the reason why they are not significant.

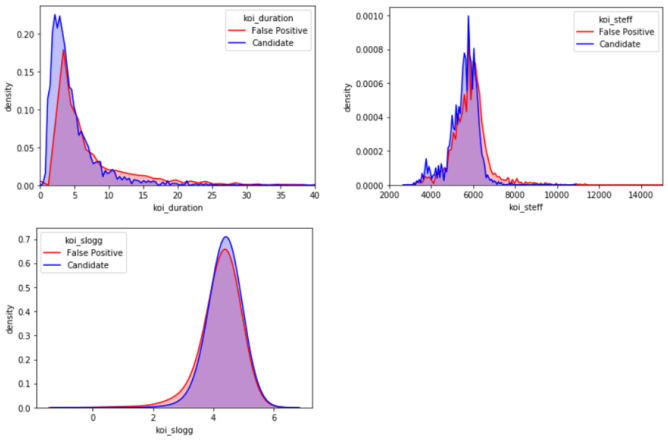


Fig. 4: Density graphs of given columns

We have plotted the correlation matrix between CANDIDATE and FALSE POSITIVE exoplanets, shown in Figure 5. koi_slogg and koi_srad have a high negative correlation, as one would expect. As explained in the Section III-A, koi_slogg is the stellar gravity and koi_srad is the stellar radius. From Newton's law of universal gravitation, we already know that they are in inverse ratio. Also, there is a positive correlation between koi_prad and koi_impact. This is also expected from their definition because koi_impact is simply the distance between the planet and the star normalized by the stellar radius and koi_prad is the planet's radius. Thus, they are correlated.

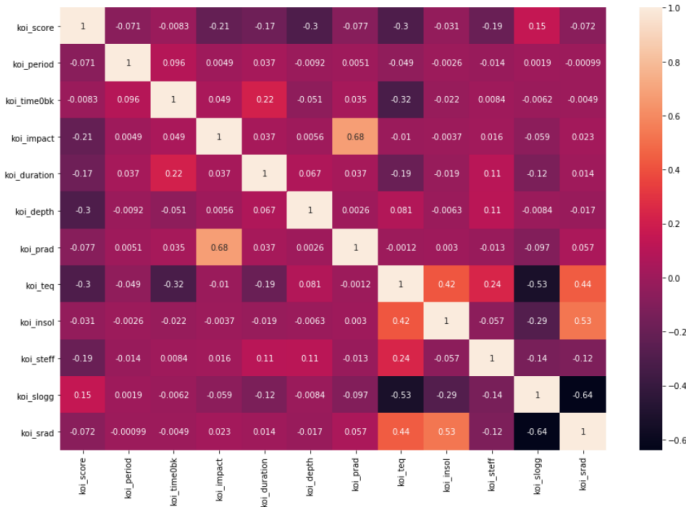


Fig. 5: Correlation matrix between CANDIDATE and FALSE POSITIVE planets

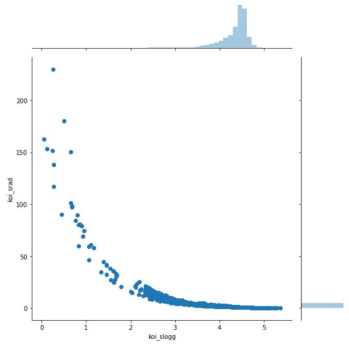


Fig. 6: Correlation between koi_slogg and koi_srad

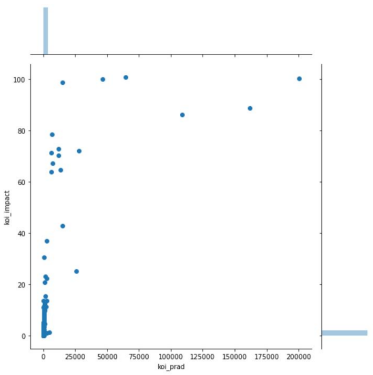


Fig. 7: Correlation between koi_prad and koi_impact

B. Newton's law of universal gravitation

As known from the law,

$$F = \frac{GMm}{r^2}$$

holds universally. By this equation, one can see that gravity has a linear relation with mass and an inverse relation with radius. An inverse relation between gravity and radius can be seen in the data too, as shown in Figure 6. After seeing this, we have decided to take a look at this equation ourselves.

Our data set has columns koi_slogg, koi_smass, and koi_srad and their error columns. We calculated F using these columns and the constant G and compared it with data from our dataset. Because the planet's mass is relatively much smaller than its star's mass, that mass was not used. Comparison of calculated and observed data is shown in Figure 8.

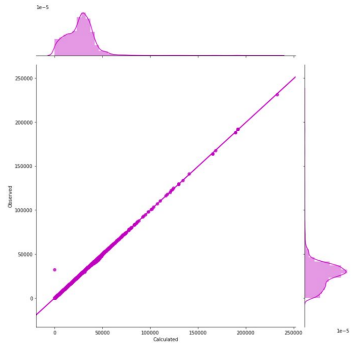


Fig. 8: Calculated and observed data graph

C. Analysis of flux values

We looked at the histogram plot of the flux changes of stars of both types with respect to time. The first row illustrates histogram plots for the 3 randomly selected confirmed stars, while the second line for the unconfirmed stars(Figure 9). We found that the range of current values for the approved ones changes too much, while the range for the approved ones is narrower.

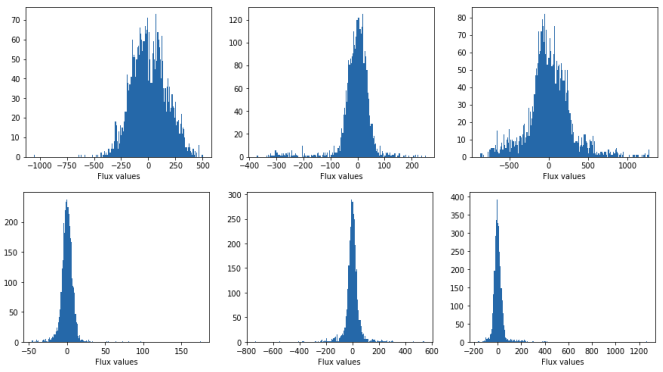


Fig. 9: Histogram plots

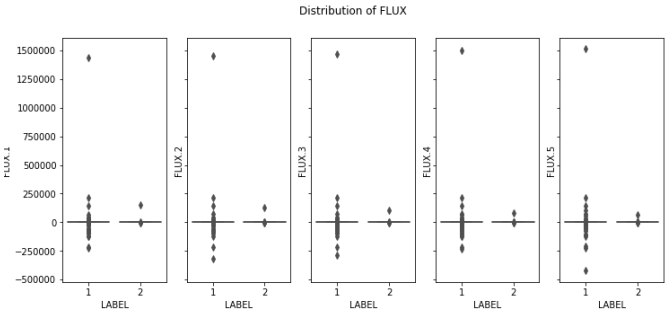


Fig. 10: Box plot for 5 different flux features

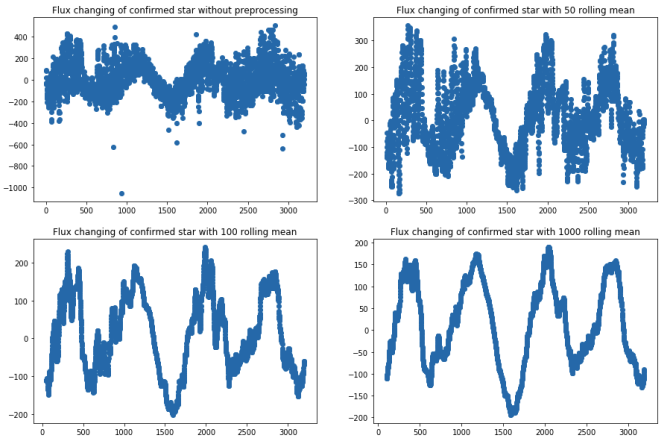


Fig. 11: Flux values for confirmed stars

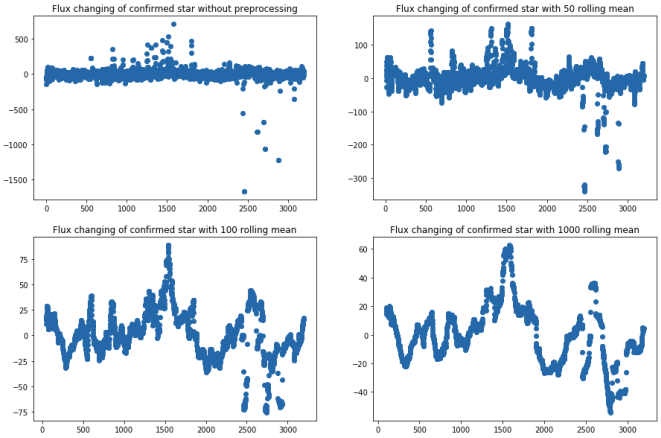


Fig. 12: Flux values for non-confirmed stars

When we take random samples from confirmed and non-confirmed stars and visualize the change of luminous flux over time, we see that the approved stars follow a sinusoidal pattern. At the same time, when we looked at both Figure 9 and Figure 10, we noticed that there is some kind of anomaly flux values. Therefore, we have obtained a more general view by visualizing the rolling mean values in 3 different scrolling windows in order to better detect the pattern. We saw a more

regular sinusoidal movement for the confirmed ones, while we were unable to detect a specific pattern for the disapproved ones.

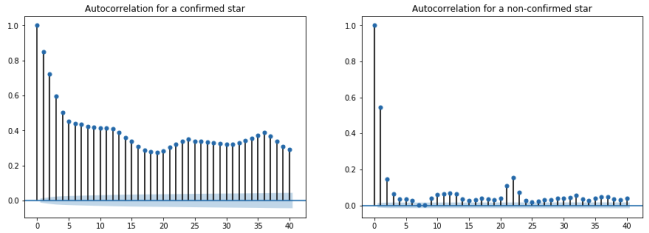


Fig. 13: Flux values for non-confirmed stars

Autocorrelation is a way of measuring the linear relationship between an observation at time t and observations at previous times. We looked at a lag 40 autocorrelations of two types of stars at 95% confidence interval. Autocorrelation at k lag values means the correlation between values that are k time periods apart. For confirmed stars, the flux values seem to depend on previous flux values, while for unconfirmed stars this correlation does not exist. We have also verified the interdependence of the flux values that we did only get from the graphs and show a sinusoidal pattern by examining their autocorrelations in Figure 13.

D. Dimension Reduction

We visualized it in 3-dimensional space by reducing the size of the properties that reflect the flux value we have. The purpose of this is whether there are similar properties between approved and unapproved stars, and whether the outlier flux values that we are aware of from other visualizations make these stars different from each other. We reduced the data size to 3 dimensions, explaining 81% of the variance of the data. The most important properties acting on these 3 basic components are FLUX.2447, FLUX.1249, FLUX72.

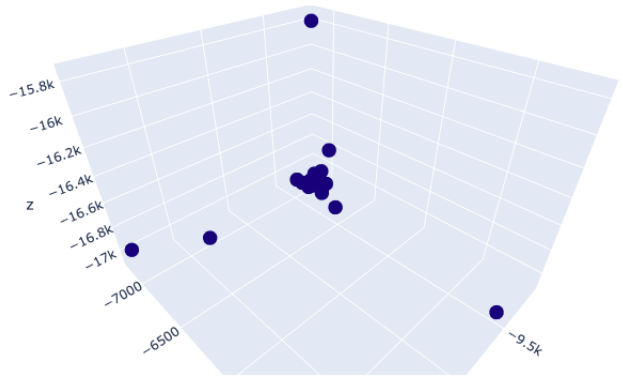


Fig. 14: Flux values for non-confirmed stars

We found outliers in 3-D visualization in Figure 14, especially among non-confirmed stars. For those stars, the flux values could differ greatly from each other. This result led to the need for further examination of outliers and to the theses

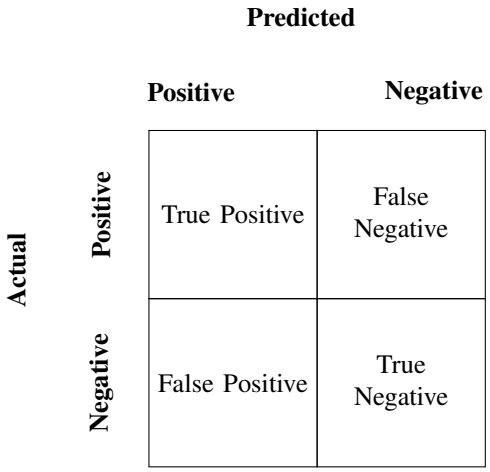


Fig. 15: Confusion Matrix

of how much the flux value recorded in which period of time affected on the result.

E. Classification Model

We used 2 different machine learning models to classify unconfirmed stars. As mentioned in Chapter 1, our data is an unbalanced dataset. In order for our model not to be affected by the imbalancing problem and not overfitting the properties of non-confirmed stars, we have equalized the number of two classes by increasing the number of confirmed stars by the SMOTE oversampling method [11] to the number of non-confirmed stars.

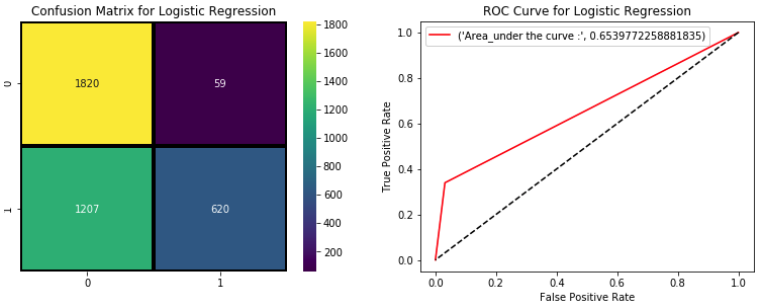
It consists of 4 different terms formed by different combinations of actual and predicted values, namely true positive, true negative, false positive, false negative. The meaning of the terms are as follows;

- True Positive: Predicting the positive actual value as positive
- True Negative: Predicting the negative actual value as negative
- False Positive (also known as Type 1 Error): Predicting the negative actual value as positive
- False Negative (also known as Type 2 Error): Predicting the positive actual value as negative

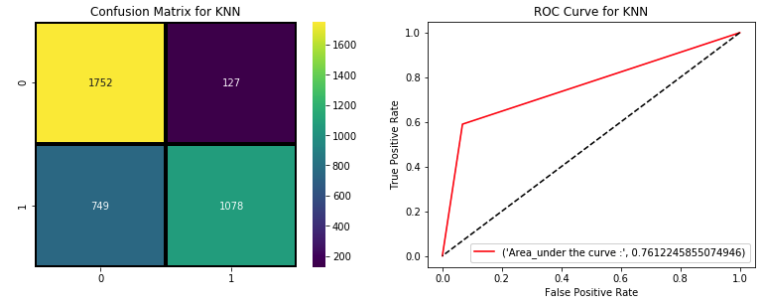
The confusion matrix is a technique to measure the performance of a classification model. The rows of the matrix represent predicted values while its columns represent actual values(and vice versa). Since we reconstruct our input data after compressing, it is suitable for comparing the input and the output data. There are many derivatives, including values such as precision and recall value that we use to measure the success of our model.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1a}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{1b}$$



(a) Results for Logistic Regression Model



(b) Results for KNN Model

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision + Recall}} \tag{1c}$$

We measured our results using the confusion matrix and the AUC score. Confusion matrix consists of different measures that explanations of those like below;

- True Positive: Predicting the confirmed planet as confirmed.
- True Negative: Predicting the non-confirmed planet as non-confirmed.
- False Positive: Predicting the non-confirmed planet as confirmed.
- False Negative: Predicting the confirmed planet as non-confirmed.

Model	Label	Precision	Recall	F1-Score
Logistic Regression	Non-Confirmed	0.60	0.97	0.74
	Confirmed	0.91	0.34	0.49
KNN	Non-Confirmed	0.70	0.93	0.80
	Confirmed	0.89	0.59	0.71

TABLE I: Precision, Recall and F1 Scores

While the precision value (1a) expresses how much of what we predicted as confirmed is actually confirmed, our recall value (1b) shows how many of the values that are actually confirmed are correctly predicted. F1 score represents the harmonic average of precision and recall values. (1c)

According to the results, we achieved a prediction success of about 80% for both classes in the KNN algorithm(Table I). We can see the effect of luminous flux being a potential planet in modeling.

V. Conclusion

This paper analyzed different types of exoplanets by using statistical methods and built a classification model. In conclusion, there are many features of exoplanets and their stars but some features show different patterns for CANDIDATE and FALSE POSITIVE exoplanets. These are the significant features. Also, some features show almost the same pattern between CANDIDATE and FALSE POSITIVE exoplanets and they are not so significant. We have studied a Physics law ourselves with the dataset we have and compared the observed data with calculated data. Also, each exoplanet has its own unique luminous flux changing pattern. Especially stars that appear to be outliers have separable properties from other stars in the same category. Confirmed stars follow a sinusoidal pattern, while unconfirmed ones have different patterns. We found that these changes in the luminous flux provide significant success in detecting stars. Creating a hybrid dataset using other data sources will affect the success of new classification models.

VI. Future Prospects

Humanity's curiosity towards the other galaxies and habitable planets is growing day by day. People want to know whether they are alone in this universe or not. Humanity could see their planet's beginning and future, if new planets are found. There is a science field called Exoplanetology, also called exoplanetary science, which is specifically on searching and studying exoplanets.

NASA is launching new missions and tools for searching of new exoplanets. Missions like Kepler and K2, and tools like Kepler telescope and TESS(Transiting Exoplanet Survey Satellite) are aimed to find exoplanets and discover the outer space.

The field of exoplanets will only grow with developing technologies. Therefore, studying the data these researches created and understanding the properties of exoplanets is necessary. Our inferences can be used to create more advanced models and these models can be used to detect if a suspected exoplanet is an actual exoplanet or not.

References

[1] Brennan, P. (2020). Retrieved from https://exoplanets.nasa.gov/what-is-an-exoplanet/in-depth/.

[2] Brennan, P. (2020). Retrieved from https://exoplanets.nasa.gov/alien-worlds/ways-to-find-a-planet//2.

[3] Aigrain, Suzanne, and F. Favata. "Bayesian detection of planetary transits-A modified version of the Gregory-Loredo method for Bayesian periodic signal detection." Astronomy Astrophysics 395.2 (2002): 625-636.

[4] Batalha, Natalie M. "Exploring exoplanet populations with NASA's Kepler Mission." Proceedings of the National Academy of Sciences 111.35 (2014): 12647-12654.

[5] Shallue, Christopher J., and Andrew Vanderburg. "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 (2018): 94.

[6] Batalha, N. M., Rowe, J. F., Bryson, S. T., Barclay, T., Burke, C. J., Caldwell, D. A., ... Dupree, A. K. (2013). Planetary candidates observed by Kepler. III. Analysis of the first 16 months of data. The Astrophysical Journal Supplement Series, 204(2), 24.

[7] Anonymous. (2020). Retrieved from https://exoplanetarchive.ipac.caltech.edu/docs/faq.html.

[8] Anonymous. (2020). Retrieved from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html.

[9] Wright, J. T., Gaudi, B. S. (2012). Exoplanet detection methods. arXiv preprint arXiv:1210.2471.

[10] Anonymous. (2017). Exoplanet Hunting in Deep Space. Retrieved from https://www.kaggle.com/keplersmachines/kepler-labelled-time-series-data.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.