**Using Machine Learning to Classify Kepler Objects of Interest**

Author: Clara Hu

## Abstract

The Kepler telescope has been deployed for more than a decade, looking for planets and potential life far away from our solar system. Since the Kepler Mission began, Kepler objects of interest and data on those objects have been collected by the Kepler telescope. This data must be analyzed and  studied by scientists to determine if those objects captured by the telescope truly are exoplanets. This project written about in this paper uses the data on Kepler objects of interest as well as other exoplanets to build a machine learning model that will classify the Kepler objects of interest by whether or not they are exoplanets using the features present in the data collected by the Kepler telescope.

## Introduction

Human interest in outer space and far away planets is centuries old. NASA's Kepler Mission continues this interest in space exploration through the discovery of hundreds of exoplanets (Johnson, 2015). The Kepler Mission surveys the Milky Way galaxy in search of potential exoplanets, collecting data on them to see if the exoplanets may be habitable, terrestrial planets similar to our Earth. Not only does the exploring potential habitable exoplanets further scientists' search for extraterrestrial life, studying newly discovered planets and their structure, including distribution, size, orbit, and masses, as well as the stars at the center of the planetary systems is important in helping scientists understand our own planet and sun as well as our own solar system's formation and development.

In order to discover potential exoplanets for the Kepler mission, the Kepler telescope surveys a large sample of stars, looking for transit events, which is when planets pass in front of the stars  (Johnson, 2015). The Kepler telescope uses light curves to detect the transits and makes note of these transits as Kepler objects of interest (KOIs). Through data collected on the transits' movement and changes in brightness, characteristics of the object of interest, such as orbital size, and temperature, can be calculated and from these characteristics, scientists can determine whether or not the object of interest is actually a planet, and if the object of interest turns out to be a planet, they may also be able to determine if the planet is habitable. This data from the Kepler mission will be used in this project, which focuses on Topic 3: Emerging Research and Technologies, Dataset A: Space Exploration.

*Research Question:* Can we build a model using data collected from the Kepler telescope on objects in other planetary or stellar systems to predict whether or not they are planets?

Inspired by the Kepler Mission, the goal of this project is to use the data collected from the mission to determine which characteristics of KOIs can be used as features to create a machine learning model to predict if a candidate KOI should be classified as a confirmed exoplanet or as a false positive (KOI is not actually a planet in transit). The research question is important because it contributes to the Kepler Mission's objective of discovering potentially habitable exoplanets by exploring which characteristics of the planets may be significant in distinguishing between an actual planet and a false detection or another object. The topic of space exploration and exoplanets is also interesting because it allows us to learn more about the hundreds of different planets out in the universe and if it may be possible for there to be another inhabitable Earth-like planet where life as we know it can exist. There has previously been some research pertaining to this topic, and this project will add on and continue to contribute to and complement the existing research.

There has been a plethora of research on exoplanets and extraterrestrial life in general, and some of that research has been done on machine learning models of this topic, which is similar to the goal of this project. The existing research on machine learning models to classify exoplanets focuses solely on Kepler objects of interest, so this project will go beyond the existing research by also considering other exoplanets that have already been discovered outside of the Kepler Mission. This project will also complement existing research by exploring different features that can be used in the models and will explore the applications of models that were not used in previous research. Models that have been used in previous research include Support Vector Machine, K Nearest Neighbors and Random Forest classification (Sturrock, Manry, and Rafiqi, 2019) as well as neural networks (Shallue and Vanderburg, 2018). In addition, because the Kepler Mission is ongoing, the data for the Kepler objects of interest is constantly being updated, and this project will use the newest data available.

## Data

For this project, I will be looking at two of the datasets provided in the DataHub from Dataset A: Space Exploration from Topic 3: Emerging Researches and Technologies as well as an external dataset that is also from the Kepler Mission.

### *Description of Data*

The data that will be used for this project is from, which includes data from reports of outer space exploration focusing on faraway exoplanets. The two data sets used in this project that are provided in the DataHub for Dataset A: Space Exploration of Topic 3: Emerging Research and Technologies are the Kepler Exoplanet Search Dataset (NASA) and the Kepler Planetary System Composite Dataset (NASA). These two data sets are both from the NASA Exoplanet Archive of the NASA Exoplanet Science Institute and include the data from reports of the Kepler Missions' outer space exploration of faraway exoplanets.

In addition to the two provided data sets, I will also be using an external dataset, Kepler Objects of Interest Composite Dataset (NASA), that is also part of the NASA Exoplanet Archive. The external data set is a necessary supplementary data set for this project because the provided Kepler Exoplanet Search Dataset is missing values for many of the parameters.

The Kepler Exoplanet Search Dataset contains data collected from the Kepler Mission, and focuses on finding candidate exoplanets that may be habitable. Each entry represents a Kepler object of interest (potential exoplanet), and using the characteristics of the KOIs provided in the data set, they can be categorized as either CONFIRMED or FALSE POSITIVE under disposition. A KOI whose disposition has not been finalized yet are categorized under CANDIDATE. The proportion of entries that are CONFIRMED, FALSE POSITIVE, and CANDIDATE exoplanets in the data set is shown in Figure 1. Because the Kepler Exoplanet Search Dataset is missing information for many of the parameters that explain different characteristics of the KOIs, I also included the external dataset, Kepler Objects of Interest Composite Dataset. Similar to the first dataset, each entry represents a Kepler object of interest. The difference between the two data sets is that the Kepler Objects of Interest Composite Dataset includes values for the characteristics that were missing in the first dataset. The data set is continuously updated, so it contains all of the most recent data on the KOIs.

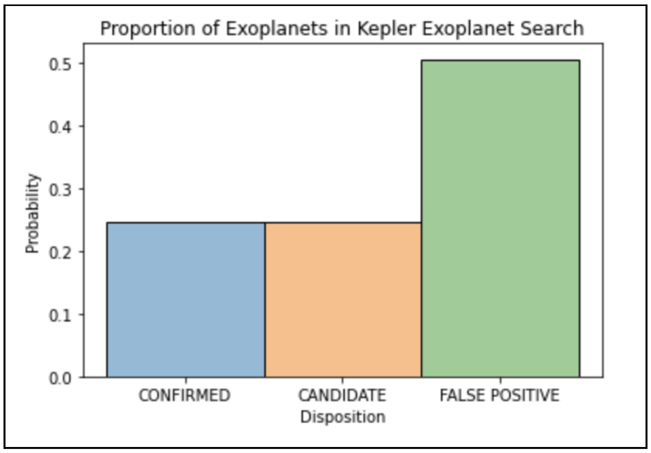

**Figure 1.** Histogram of CONFIRMED, CANDIDATE, & FALSE POSITIVE exoplanets in Kepler Exoplanet Search

The second data set used in the project that was provided in the DataHub is the Kepler Planetary Systems Composite Dataset. Each entry in this dataset contains information on a confirmed exoplanet. The features of the data set include characteristics similar to what is described for the KOIs in the other two data sets. The entries also include data for confirmed exoplanets that were part of the Kepler missions, which means that there may be some overlap between the Kepler Planetary Systems Composite Dataset and the Kepler Exoplanet Search Dataset (Figure 2). Since all of the entries in the Kepler Planetary Systems Composite Dataset are confirmed exoplanets, they could be considered as categorized as CONFIRMED under disposition if the data set had that feature.

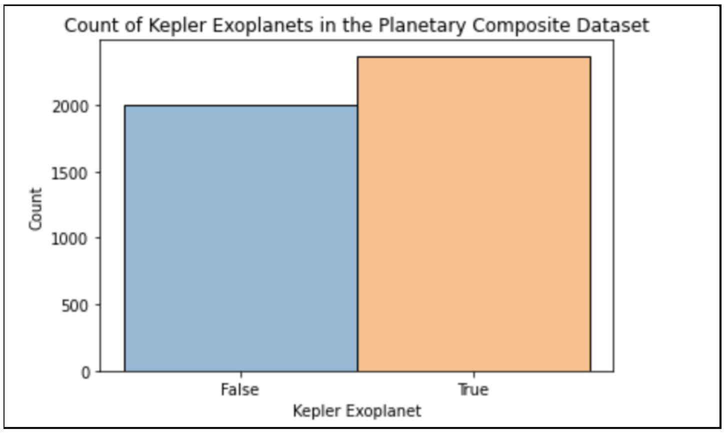

**Figure 2.** Histogram showing the count of the Kepler Exoplanets in the Kepler Planetary System Composite Dataset (CONFIRMED Kepler Objects of Interest in the Kepler Exoplanet Search Dataset)

A bias that may have arisen from the collection of the Kepler Exoplanet Search data was that in the Kepler Mission, the Kepler Space Telescope is trained to look in areas where objects of interest or masses may have already been detected or are expected to be detected, and also there may be observational bias towards the detection of larger planets (Sturrock, Manry, and Rafiqi, 2019).

## Methodology

### *Exploratory Data Analysis*

The first step of the methodology after acquiring the data for the project is to complete the exploratory data analysis (EDA) to explore relationships in the data sets and determine important features that may be useful for the model we will build.

Through looking at Kepler Exoplanet Search Dataset, we can see that the dataset includes many invalid entries, and many of the parameters only include NaN entries. Many of the parameters that were only had NaN values in the Kepler Exoplanet have valid entries in the Kepler Object of Interest Composite Dataset, including planet radius, transit duration, transit depth, planet equilibrium temperature, stellar effective temperature, and insolation flux. Because the entries of both of the data sets describe the same KOIs, I merged the two into a single merged Kepler Object of Interest data frame.

Because we are interested in looking at KOIs as well as the other confirmed exoplanets in the Kepler Planetary System Composite Dataset, we need to find the features that the datasets have in common. This is completed by first exploring the data sets separately to see which features include valid entries in the individual datasets. Then, we find the features that are common between the two datasets. Excluding the disposition and identification parameters of the KOIs and exoplanets, the common list of features are shown in Table 1.

**Common List of Features:**

| Feature Name in Kepler Object of Interest Dataset | Feature Name in Kepler Planetary System Composite Dataset | Description of Feature |
|---|---|---|
| koi_period | pl_orbper | Orbital period (days) |
| koi_prad | pl_rade | Planet Radius (Earth Radius) |
| koi_insol | pl_insol | Insolation Flux (earth flux) |
| koi_teq | pl_eqt | Equilibrium Temperature (K) |
| koi_depth | pl_trandep | Transit Depth (%) |
| koi_duration | pl_trandur | Transit duration (hours) |
| koi_steff | st_teff | Stellar Effective Temperature (K) |
| ra | ra | Right Ascension of the planetary system in decimal degrees (degrees) |
| dec | dec | Declination of the planetary system in decimal degrees (degrees) |

**Table 1.** Features describing characteristics of Kepler objects of interest and exoplanets common to both datasets

Through looking at our dataset, we find that the features that both datasets have in common are planet radius, insolation flux, equilibrium temperature, transit depth, transit duration, stellar effective temperature, right ascension angle, and declination angle. After identifying these features, the next step in the methodology is feature engineering and combining the two datasets together so that the useful features are identified and can be added to the model that we make. Because many of the features' distribution are skewed, in order for them to be used, transformations need to be applied to the features as well.

### *Feature Engineering and Selection*

To prepare the data sets to be able to be used for modeling, the invalid entries are removed, meaning that the entries containing NaN values are removed from the data sets, and only the features which are present in both data sets as previously mentioned are kept along with identification information and and the disposition column in the Kepler object of interest dataset. Because all the exoplanets in the Kepler Planetary Systems Composite Dataset are already confirmed, a column named "disposition" is added to that dataset with "CONFIRMED" as the value for all entries. Then, the columns in the data sets are renamed to match each other, and the Kepler objects of interest data set and the Kepler Planetary Systems Composite Dataset are merged into one single data frame, dropping the duplicate entries from the merged data frame.

Before further evaluating the features and building the model for the project, the merged data set is split into a training set and a test set. The training set includes all the entries where the disposition has already been determined as CONFIRMED or FALSE POSITIVE, whereas the test set consists of all the CANDIDATE entries. This is because the goal of the project is to build a model which will be able to predict whether the CANDIDATE entries should be classified as CONFIRMED or FALSE POSITIVE given characteristics of each object of interest. To make the classification possible when modeling, in the training set, a new column is added which encodes CONFIRMED exoplanets as 1 and FALSE POSITIVE exoplanets as 0. The proportion of CONFIRMED and FALSE POSITIVE exoplanets in the training set is shown in Figure 3.

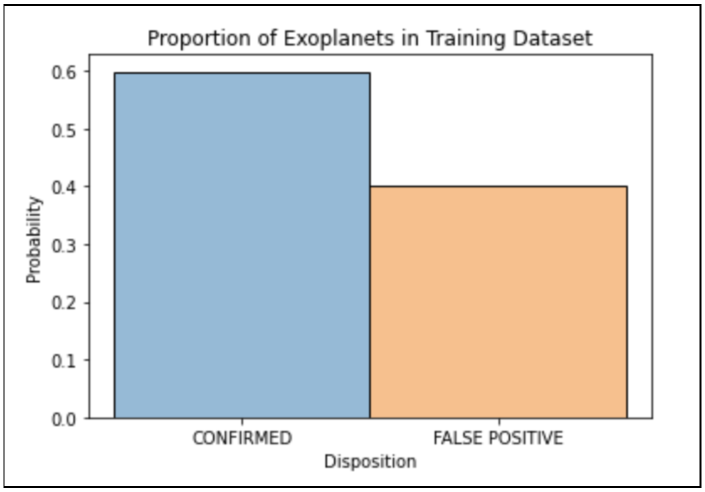

**Figure 3.** Histogram of CONFIRMED and FALSE POSITIVE exoplanets in the training dataset

In order to determine which features may be useful in the model, a correlation heatmap was made to explore any primary relationships between the features (Figure 4). The correlation heatmap shows that there seems to be a high level of correlation between Right Ascension angle and Declination angle. In addition, there is also a moderate level of correlation between Equilibrium Temperature and Stellar Effective Temperature, Equilibrium Temperature and Insolation Flux, as well as Transit Duration and Orbital Period, so only one feature from each of those pairs likely needs to remain and be considered for the model.

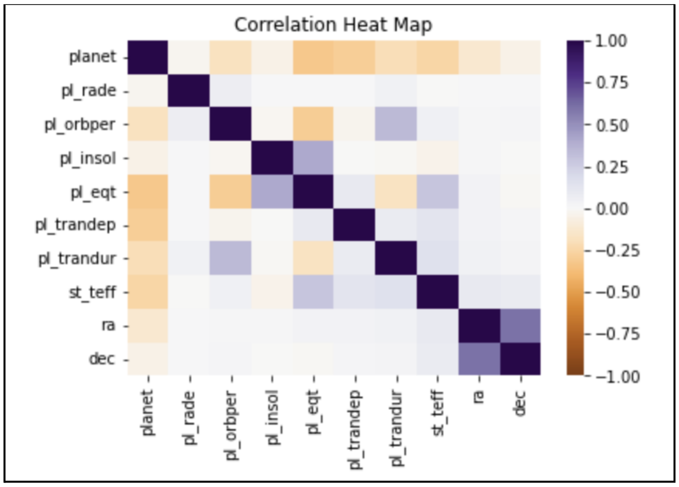

**Figure 4.** Correlation Heat Map of possible features to use in the model

Many features in the data set are highly skewed, so they had to be transformed in order to be useful. These features include orbital period, transit depth, and planet radius. Log transformations were applied. After transforming these features, joint plots were made to explore the differences in distribution of the features between CONFIRMED and FALSE POSITIVE objects (Figure 5). While some areas of the plots overlap, there are clear differences in the distributions of log transformed transit depth and log transformed orbital period as well as log transformed radius and equilibrium temperature between CONFIRMED and FALSE POSITIVE exoplanets, so these features are likely to be useful in classifying the KOI.

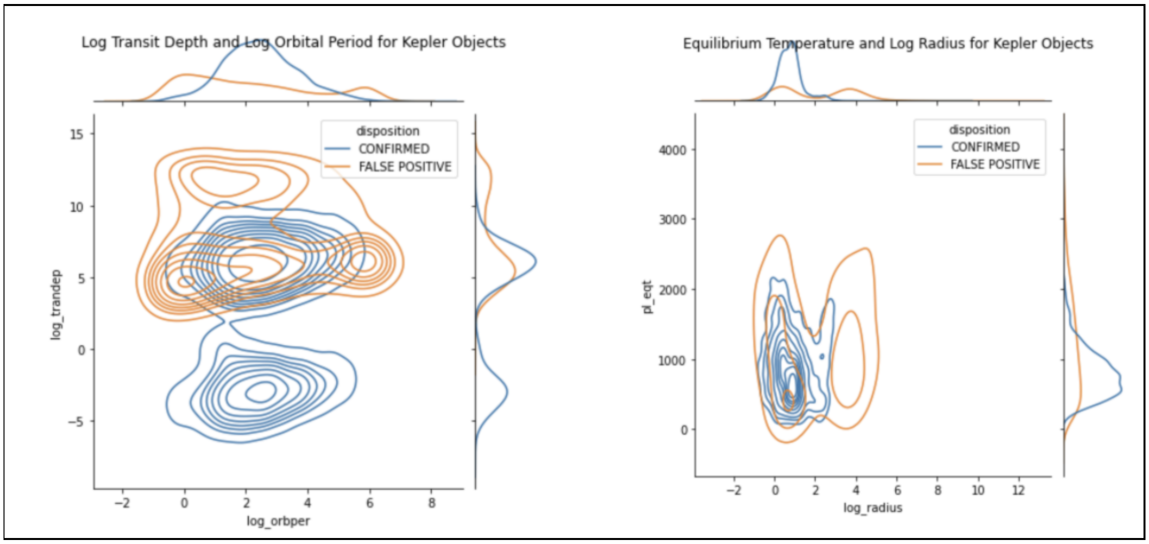

**Figure 5.** Joint KDE plots of Log Transit Depth and Log Orbital Period as well as Equilibrium Temperature and Log Radius for Kepler objects.

To explore other features in relation to their categories of CONFIRMED and FALSE POSITIVE exoplanets, KDE plots were made as well (Figure 6). For the features transit duration (as well as log transit duration), declination angle, and right ascension angle, the distributions for CONFIRMED and FALSE POSITIVE exoplanets almost completely overlap, so those features are likely not to be of use in the model.

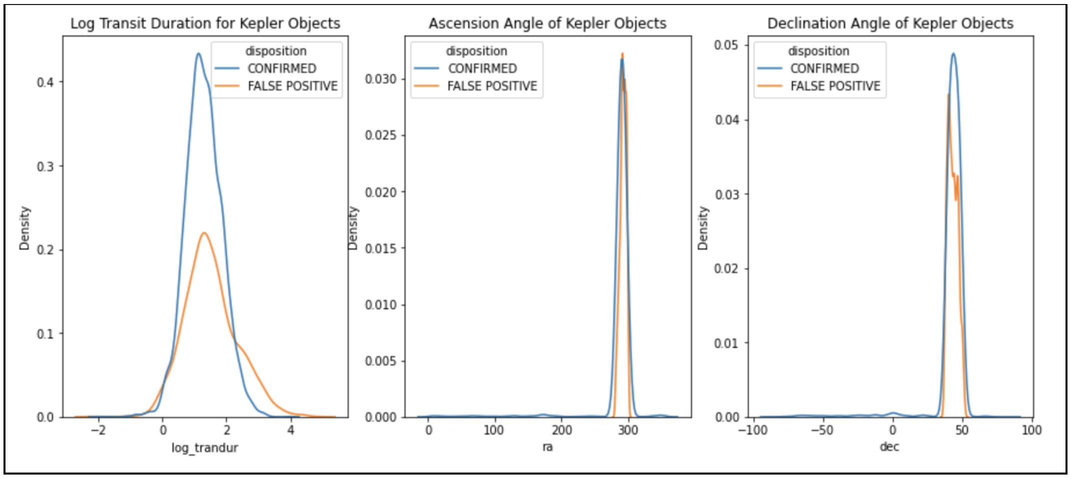

**Figure 6.** KDE plots of Log Transit Duration, Right Ascension Angle, and Declination Angle for Kepler objects.

After exploring the data through EDA and visualizations, the features selected to be included in the machine learning model for predicting if a KOIt is a CONFIRMED exoplanet or a FALSE positive are log radius, log orbital period, equilibrium temperature, and log transit depth.

### *Modeling*

In order to answer this project's research question, a model that can classify potential exoplanets using the characteristics of those objects needed to be built. The characteristics, or features, to be used in the model were listed above. For this project, two machine learning models were built to answer the research question. The first baseline model used was a logistic regression model because logistic regressions are useful in solving binary (0, 1) classification problems, which is what this project is. We are trying to find a way to classify candidate exoplanets as either CONFIRMED (1) exoplanets or FALSE POSITIVE (0) exoplanets. A second machine learning model was also used in this project to improve upon the classifications made by the logistic regression model. For the second model, a decision tree model was used. The decision tree is an effective supervised machine learning model used for classification. A decision tree model works well for binary classification as it is used in this project, but it can also be used when there are multiple categories.

In both of the models, L2 regularization was applied to reduce variance. From a bias-variance tradeoff standpoint, the reduction in variance increases the accuracy of the model, but it also increases bias and error for the model. In addition, for the decision tree model, the overfitting problem was addressed through increasing the parameter min_samples_leaf to 5 when building the model. In the model, the min_samples_leaf parameter controls the number of splits considered when informing the decision made by the model, ensuring that multiple samples are used to inform every decision made by the tree. This allows the decision tree to avoid low variance, overfit nodes.

The same set of data was used for both of the models so that they could be evaluated and compared with each other. The original training set of data was further split into a 90% block as training data for the models and a 10% used as the validation set. This was to ensure that cross-validation could be performed on the models and so that the models could additionally be checked for overfitting.

## Results

Through creating the logistic regression model and decision tree model, we were able to classify Kepler objects of interest as CONFIRMED or FALSE POSITIVE exoplanets using the features log radius, log orbital period, equilibrium temperature, and log transit depth. While both models were able to make classifications, their performance was very different. Overall, the performance of the decision tree model was much better than the performance of the logistic regression model. The decision tree model's predictions had higher accuracy, recall, and precision in comparison to the logistic regression model. The accuracy of the model is the proportion of classifications that the model correctly predicts while the precision of the model is the proportion of true positives out of all the predicted positives, and the recall the proportion of observations that were actually positives that were correctly predicted. In the case of this project, the precision is the proportion of predictions that were actually CONFIRMED out of all the observations that were predicted to be CONFIRMED, and the recall was the proportion of CONFIRMED observations that the model correctly predicted. The comparison of these performance metrics can be seen for the training set in Table 2.

**Training Set Accuracy, Precision, and Recall**

| | **Logistic Regression Model** | **Decision Tree Model** |
|---|---|---|
| **Accuracy** | 0.785 | 0.943 |
| **Precision** | 0.888 | 0.957 |
| **Recall** | 0.783 | 0.949 |

**Table 2.** Accuracy, precision, and recall of training set for Logistic Regression and Decision Tree Models

When looking at binary error and cross-validation (cv) error, we can also see that the decision tree model performed better (Table 3). The binary error is the proportion of the validation set that the model classifies incorrectly, while the cv error determines the performance of the model through using the loss function. For this project, the loss function used is cross-entropy loss, or log loss. Both the binary error and cv error for the logistic regression model was much larger than for the decision tree model.

**Binary Error and Cross-Validation Error**

| | Logistic Regression Model | Decision Tree Model |
|---|---|---|
| **Binary Error** | 0.221 | 0.127 |
| **Cross-Validation Error** | 7.620 | 4.394 |

**Table 3.** Comparison of binary error and cv error between the Logistic Regression and Decision Tree Models

Through our results, we can answer our research question by saying that potential exoplanets can be classified with a decision tree model using the features  log radius, log orbital period, equilibrium temperature, and log transit depth. However, there is still room for improvement in the model, which can be made by further exploring different features and testing other machine learning models.

## Discussion

### *Evaluation of Models*

The two different models used in this project ended up with one model performing significantly better than the other. This is likely because when the logistic model classifies observations based on their data points, it only divides the space into two, whereas the decision tree model will continue to bisect the space into smaller and smaller regions, which mean that it is more specific and better at breaking down complex data into manageable parts to classify. To further evaluate and compare the two models' performances we can look at their false positive and false negative rates. The counts of these metrics (as well as true positives and negatives) can be seen through confusion matrices (Figure 7). The false negative and false positive rates are much lower in the decision tree model compared to the logistic regression model.

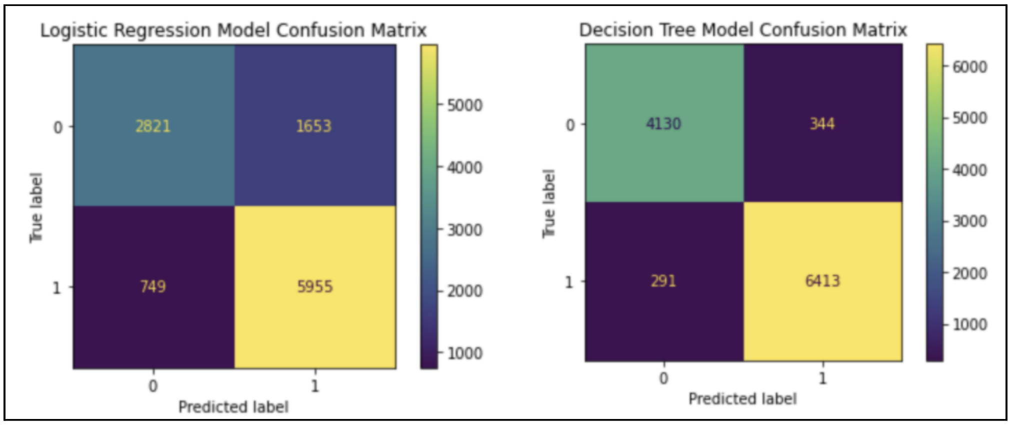

**Figure 7.** Confusion Matrix for Logistic Regression Model and Decision Tree Model

The reason for the high false positive rate for the logistic regression model can be seen through a receiver operating characteristic curve (ROC curve). The ROC curve shows the tradeoffs for false positives and false negatives at various cutoffs for a binary classifier system. The ROC curve for the logistic regression model shows that quite a high proportion of false positives is needed to capture all the true positives (Figure 8).

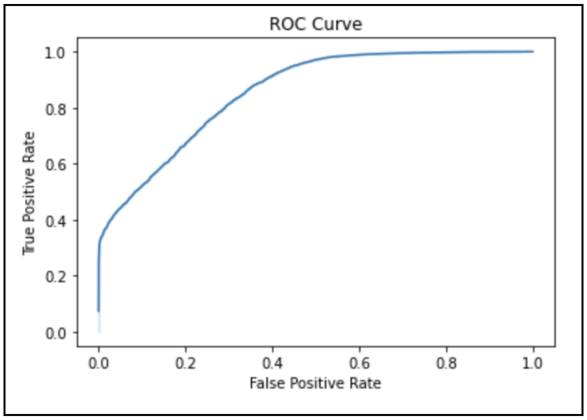

**Figure 8.** ROC Curve for Logistic Regression Model

As mentioned in the results, although the decision tree model already improved upon the baseline logistic regression model, there is still room for improvement. The model may be able to be further improved if more features related to the characteristics of KOIs or existing exoplanets were used. This would mean that searching through other external datasets for features or future collection of more data is needed.

*Applying Model to the Test (CANDIDATE) Data*

While the machine learning models were evaluated using the training and validation sets, an interesting application of the models for this project was to predict the classification of the test (CANDIDATE) data, even though we are currently unable to determine whether or not those predictions are correct (Table 4).

**CANDIDATE Exoplanet Predictions**

| | **Logistic Regression Model** | **Decision Tree Model** |
|---|---|---|
| **Predicted as CONFIRMED (count; proportion)** | 2532 (0.79) | 1841 (0.58) |
| **Predicted FALSE POSITIVE (count; proportion)** | 650 (0.21) | 1341 (0.42) |

**Table 4.** Count and proportion of CONFIRMED and FALSE POSITIVE predictions for each model

Using the decision tree model, less KOIs were predicted to be CONFIRMED compared to when using the logistic regression model. This result makes sense because of the much higher false positive rate that the logistic regression model had, meaning that it was more likely to classify KOIs as CONFIRMED.

*Potential Societal Impacts*

The potential societal impacts of this project include contributing to research that searches for habitable exoplanets. Being able to build a machine learning model using measurable and calculable characteristics of objects in outer space and determine whether or not the objects are exoplanets can contribute immensely to the discovery of new planets. Discovering planets outside of our solar system can lead to potential insights on life outside our solar system as well. In conjunction with the Kepler Mission and other research on exoplanets, this project can help provide the public with greater scientific knowledge and understanding of outer space as well as better understand our own solar system.

*Ethics*

When conducting research and working on projects like this one, it is important to keep ethics in mind. Ethics encompass the moral principles that should be followed in research. This project relates machine learning and data science to astronomy and earth and planetary sciences. In these fields, ethics dictates that it is important to produce unbiased research that contributes to the field in a positive manner so that knowledge in the field can continue to grow, and those who are learning can be educated with accurate research. Unethical behavior such as falsifying data that influences future work could cause the field to regress. In addition, it is important to document research so that it is transparent, verifiable, and replicable. The documentation of analysis and this report allows for this project to satisfy those requirements.Finally, ethical practices are also important for the development of scientific fields and building trust between scientists and the public.

## Conclusion

Through this project, the question of could a machine learning model that can classify Kepler objects of interest be built using features made up of characteristics of those objects of interest was answered. The project's results found that a decision tree model could classify the KOIs as exoplanets using the features log radius, log orbital period, equilibrium temperature, and log transit depth. This result is only the introduction to a plethora of potential future work that may be done on this topic. The machine learning model may be further improved if more features related to the objects of interest are used, meaning that future research can be done with data from other external data sets or with future collection of data. In addition, future work may include verifying the predictions made by the model as well as looking into what objects the objects which were not classified as exoplanets truly are. The classification of exoplanets can lead to more research on far away planetary systems and extraterrestrial life, paving the way for more discoveries to be made.

## References

*Data columns in Kepler objects of interest table*. Data columns in Kepler Objects of Interest Table. (n.d.). Retrieved December 7, 2021, from https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html.

Johnson, M. (2015, April 14). *Mission overview*. NASA. Retrieved December 6, 2021, from https://www.nasa.gov/mission_pages/kepler/overview/index.html.

*NASA exoplanet archive*. NASA Exoplanet Archive. (n.d.). Retrieved December 7, 2021, from https://exoplanetarchive.ipac.caltech.edu/.

*Planetary systems and planetary systems composite parameters table definitions*. Planetary Systems and Planetary Systems Composite Parameters Data Column Definitions. (n.d.). Retrieved December 7, 2021, from https://exoplanetarchive.ipac.caltech.edu/docs/API_PS_columns.html.

Sturrock, George Clayton; Manry, Brychan; and Rafiqi, Sohail (2019) "Machine Learning Pipeline for Exoplanet Classification," *SMU Data Science Review*: Vol. 2 : No. 1 , Article 9.

Shallue, C. J., & 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), 94.