The Effectiveness of Machine Learning on the Classification of Exoplanets

Anirudh Iyengar

duPont Manual High School

Abstract

The purpose of the experiment was to determine what type of machine learning model would be the best at classifying exoplanets. The Kepler mission was the latest mission on exoplanet exploration, and the data that was collected from the mission was used in this experiment. It was hoped that the experimentation can further develop understanding on exoplanets as well as create more of an automated approach towards exoplanet classification. Four classification models were tested and include Random Forest, Decision Tree, K-Nearest Neighbor, and Logistic Regression. The models were tested based of metric values that include accuracy, percent of error (RMSE), and a Confusion Matrix. Confusion Matrix is a table that considers the true value of the dataset, which shows the number of proper classifications for True Positive, True Negative, False Positive, and False Negative. 3 tests were conducted on each model producing the metric values after the testing and training process. After analyzing the data, the Random Forest model has the best RMSE value and accuracy with a difference of 3% when compared to the 2nd best model. The difference is significant as when the confusion matrix is compared, the Random Forest model can be seen to have the highest number of True classifications and the lowest number of False classifications, thus suggesting that the Random Forest model had the best performance when classifying exoplanet data. Further experimentation includes implementing a visual based classifier for exoplanets, evaluating other classifiers, and conducting more tests on the models.

Keywords: Exoplanet, Confusion Matrix, Random Forest, Decision Tree, KNN, Logistic Regression, RMSE,

Introduction

An exoplanet is any planet beyond the solar system. Most orbit stars, while others are free-floating exoplanets, called rogue planets, which orbit the galactic center but aren't tethered to any star. Currently, 4,551 exoplanets have been discovered, and 7,875 planets are unconfirmed as of October 28, 2021(NASA et al., 2015). Most of the planets are in a small region in the Milky Way galaxy (Gilliand et al., 2011). As exoplanets are discovered, comparisons can be done to see how similar they are to planets in Earth's solar system, such as an exoplanet having a rocky composition like Earth, or a gas heavy composition like Jupiter or Saturn. Exoplanet exploration started around the 1990s and has since progressed rapidly through the larger investment NASA is putting into its exploration. The Kepler mission which ended in 2018, was the main mission in which 2,000 exoplanets were discovered and confirmed and revealed billions of more hidden planets in the Milky Way (NASA et al., 2021). Currently, the data collected from the mission has been used as the primary source of information to confirm exoplanets.

The current process used to classify an exoplanet is through the transit method. When a planet is a star, it is called a transit. Since the planet will be in front of the star, the light emitted from the star will appear to be dimmer. This observation of how bright the light emitted from the star during transit is what helps decide if the planet is an exoplanet or not (Aigrain et al., 2004). Generally, 3 transits are needed to confirm the existence of the planet. The data is compared at a scale of 1% of the star's total brightness. The transit method also provides more information on the planet's attributes and orbit. Since the size of the planet is proportional to the amount of light being blocked, the amount of light being blocked encodes the sizes of the planets. It also shows transit duration, or the time light has been blocked as well as the time between transits or the orbital period. The orbital period is considered the length of a year for that planet since it would be the time the planet comes back from a revolution.

The data collected from the Kepler missions require large amounts of processing and time making it very difficult to make any predictions and classifications of new exoplanets. Machine learning (ML) provides an efficient way for classifying these planets because of its numerous advantages. A finished ML model requires no labor as it can train and improve itself instead of a process of manual fine-tuning.

The question in this project is, "What type of Machine Learning model can most accurately classify an Exoplanet?" The hypothesis is that a Random Forest model will have the greatest accuracy since there are not too many parameters or input dimensions which is good for efficient Random Forest models.

The models being compared are Decision Tree, K-Nearest Neighbor (KNN), Random Forest, and Logistic Regression (control). The dataset is from Caltech's exoplanet archive. It includes about 9000 rows of data and includes the information collected from the Kepler mission (Kepler objects of interest, n.d.). After this, each model will be trained with 60 percent of the data and tested with the other 40 percent of data (this is subject to change). The accuracy and RMSE of the models from this dataset will then be recorded for each of the models.

4 different types of models are used in this experiment. Each model represents an independent variable level. The data split will remain constant throughout all models. The hardware used will remain the same for the training process. The dependent variable of this project is the accuracy of the different models. All the model's accuracy and results will be tested by hand, as graphs of example transits will be created to see if the model's predictions are accurate. Statistical methods like r squared and cross-validation may be used for data comparison. All conduction of this research will be done in a residential area on a computer. A "winner" will be found by comparing the RMSE (root mean squared error) values of each of the

models since it shows the percentage of error of each model meaning that the model with the lowest percentage of error is the best.

Methodology

The primary focus of this project is to figure out what Machine Learning model is most efficient to classify an Exoplanet. This means that multiple (4 distinct model types were created) Machine Learning models must be written and trained on a specific dataset that is related to exoplanets. First a dataset was found on recent Exoplanet information. NASA's most recent mission for Exoplanets called the Kepler mission has collected data on planets that have been spotted. This dataset is used as the primary dataset in the experiment as it provides all the necessary information needed to know about a planet, and that it is also from a trusted source of the Jet Propulsion Laboratory of Nasa. The dataset can be located at this link, "https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi."

After the dataset was selected, the process of creating the models was next. The models were tested and built using a Windows 10 machine. Python was selected as the programming language of use in this project because of its computational capabilities and the Machine Learning and mathematical libraries that come with it needed to write, train, and test the machine learning models.

A Windows 10 machine was used as it was the machine at access. Python 3.9 is the specific version of Python used as it was the latest version at this period. Vim was used as a text editor since it was the preferences editor of choice and Google Collab was also used as the interface for training as it provided access to a GPU (Graphics Processing Unit) and RAM (Random Access Memory) which are essential components for proper model training.

For Python to be used for Machine Learning, the packages NumPy, pandas, and SciPy were needed to be downloaded using PIP (python package manager). NumPy allows for mathematical capabilities and arrays, pandas allow for dataset management, and SciPy allows for scientific computations and technical computations.

First using pandas, a data frame was created to modify the dataset to relevant information and values such as Transit Epochs, Transit Duration, Number of Transit, etc. Then using NumPy and SciPy, Machine Learning models were coded. Models were created with different weights and biases to see how they would affect the results and predictions of the model. Each model had about 3 weight and bias configurations that were tested to see its efficiency.

To go more in depth on the testing procedure, each of the 4 models were trained by 60 percent of the data and tested by the other 40 percent of the data. This is because the dataset is limited to around 9000 entries, so having more training or less training would throw off the balance between the two. After the accuracy reading was outputted for each model, the weight and biases of the models were adjusted and then tested again for accuracy. These steps were repeated 2 more times for a total of 3 tests per model meaning 12 tests in total since there are 4 distinct model types. The model's accuracy is tested by their RMSE (root mean square error) values as it shows the percent of error each model in their predictions. The model's efficiency is tested on accuracy from a separate part of the dataset which includes False Positive and Positive exoplanets that is partitioned from the Caltech dataset.

A model was considered good when the accuracy was above 80 percent. This was chosen as it shows that the model is making correct predictions consistently, but not every prediction was correct. The higher the accuracy the better for that certain model as it showed that the

predictions were getting better and more precise. The lower the RMSE value the better as a lower percent of error is better for the model.

Data and Results

Table 1: Collected Raw Data by Test

	Logistic Regression	Random Forest	Decision Tree	K-Nearest Neighbor(KNN)
RMSE(Test 1)	0.4204874996332874	0.19929407769481777	0.2518350334021991	0.4456351052387371
RMSE(Test 2)	0.4135749054697794	0.19523470984497382	0.2587359010571127	0.4234151052387371
RMSE(Test 3)	0.4061506617323485	0.2008948590547275	0.25373570212777097	0.4345351052387371
Confusion Matrix	TN, FP, FN, TP [1186 308 207 1421]	TN, FP, FN, TP [1460 34 90 1538]	TN, FP, FN, TP [1396 98 100 1528]	TN, FP, FN, TP [1132 362 258 1370]
Variance	0.3431428416368176	0.848239394926142	0.7458499024764085	0.20861825681102752
Accuracy	0.8350416399743754	0.9618834080717489	0.9365791159513133	0.8014093529788597
Average Root Mean Square Error	0.41340435561	0.1984745489	0.25476887886	0.43452843857207

Table 1 shows RMSE (Root Mean Squared Error) being used as the main metric to compare the different independent variable levels which are the 4 different machine learning models. It represents the spread of the potential error that comes from the predictions of the models and is represented by a quantitative value. All these different tests were run to help test the model's accuracy. In the first test KNN had the highest RMSE, of 0.4456351052387371. This implies that the model had a low accuracy during the testing and training process against all the models. This pattern seems to have kept up for the other 2 tests as the RMSE value of the KNN model was the highest in the remaining 2 tests. When compared to other algorithms with the 3rd best accuracy (Logistic Regression (Control)) there exists a minor difference between all the tests. This implies that there isn’t much variability between data when compared with each other.

When looking at the variance row for all models, all values are tenths greater than 0, meaning that the spread of data is very minimal for all RMSE data collected.

Another metric represented in this table is the Confusion matrix. TP stands for Total Positive information, FP stands for False Positive information, FN stands for False Negative information, and TN stands for True Negative information. The Random Forest model had the highest number of current TP and TN classifications and the lowest amount of FP and FN classifications. This model was able to accurately classify these pieces of information from the dataset. The other 3 models appear to have similar Confusion Matrix results, implying that they may have gone through some of the same circumstances during the training process.

Two algorithms that stood out were the Random Forest and Decision Tree models. Their respective RMSE and accuracy suggest that either of them can be selected as statistically better and support the alternate hypothesis.

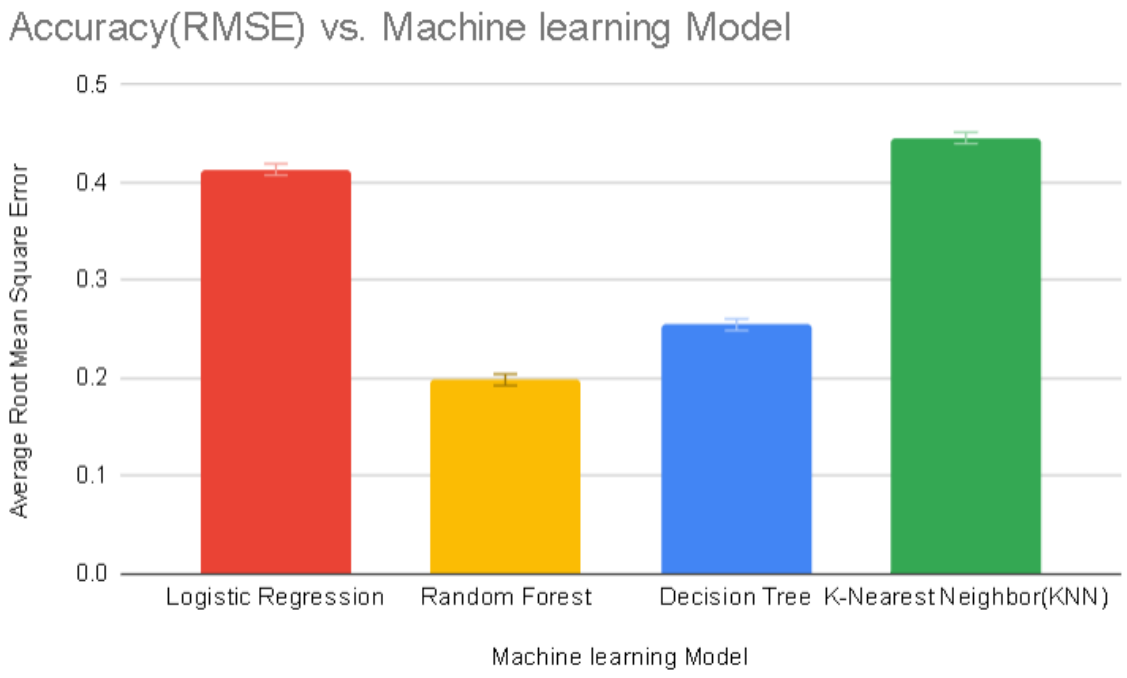

Fig 1

Figure 1 is comparing the different RMSE accuracy values between all the independent variable levels which are the different Machine Learning models. This is a comparison of the average RMSE values for each model, and here it shows that these models had a much higher percent of error compared to the other two. Since this is a comparison of averages, no variability can be seen here, but if the graph is compared to Table 1, there are no instances of major variability between the data showing that certain comparisons can be made. When looking at the other two models, Decision Tree, and Random Forest, it shows that the Random Forest has a lower percent of error by a value of around .5 which indicates a 5 percent difference. Though this may seem like a minor difference, in a real-world usage of these models, a 5 percent difference wavers the true accuracy of the model by around 7 percent as shown from Table 1, as the accuracy values of these models had a difference of 0.7. Generally, a conclusion can be made that Random Forest and Decision Tree are the most accurate and consistent due to their low RMSE values. Though there is an evident difference between RMSE values of the models, no true definite predictive model can be chosen to prove what model is most suited for this situation. Taking these models to the field right now would not be a well-suited decision as minimal variability would still need to be decreased to the thousandths or the hundred thousandths, as in a real-world situation, thousands of tests would be running on the application meaning such differences wouldn't be acceptable. More trials would have to be conducted, but that means more data would have to be introduced as the size of data is limited.

Table 2: Descriptive Statistics of Accuracy(RMSE)				
Machine Learning Model	Logistic Regression	Random Forest	Decision Tree	K-Nearest Neighbor(KNN)
Mean	0.4134043556	0.1984745489	0.2547688789	0.4345284385
Standard Deviation	0.0058542318	0.0019136043	0.0029104506	0.0090712782
1 SD(68% Band)	0.4075501238 - 0.4192585874	0.1965609446 - 0.2003881532	0.2518584283 - 0.2576793295	0.4254571603 - 0.4435997167
2 SD(95% Band)	0.401695892 - 0.4251128192	0.2023017575 - 0.1946473403	0.2489479777 - 0.2605897801	0.4163858821 - 0.4526709949
3 SD(99% Band)	0.395841660 - 0.430967051	0.192733736 - 0.204215362	0.2460375271 - 0.2635002307	0.4073146039- 0.4617422731
T test(a = 0.05)		T value = -56.0002, p < .00001	T value = -75.0282, p<.00001	T value = -441.9984, p<.00001
		Significant	Significant	Significant

Table 2 represents a collection of Descriptive statistics that were performed based on the main accuracy metric RMSE. Logistic Regression was used as the control model throughout the experiment. Random Forest had the lowest mean for RMSE, and KNN had the largest mean in terms of RMSE. The greatest standard deviation is seen in the KNN model meaning that it has the highest variance since variance is $(\sigma)^2$, meaning that the data collected from KNN had the highest spread. The research hypothesis is that the Random Forest model would have the highest performance in terms of accuracy. The null hypothesis of the experiment was that all the models would have the same levels of performance in terms of accuracy, and one tailed T-tests were conducted at 0.05 significance to test it. The comparison between Logistic Regression and Random Forest yielded statistically significant results, rejecting the null hypothesis, and not rejecting the alternative hypothesis. Similarly in the comparisons between Logistic Regression and Decision Tree, and Logistic Regression and KNN, the T-tests yielded statistically significant results, rejecting the null hypothesis, and not rejecting the alternative hypothesis. Overall, the data supported the research that the Random Forest model would have the highest performance in terms of accuracy compared to other popular classification models, and that the data that was a

result from testing is significant. To prove or disprove the research hypothesis, a comparison can be done with another project on the models built in the industry. A similar project(https://arxiv.org/abs/2011.14135), used the same types of models that had an accuracy of around 94.1 percent with precision up to 80 percent.

Conclusion

The purpose of the experiment was to examine which type of machine learning classification model was the best at classifying exoplanets. There were four different models assessed which include Logistic Regression, K-Nearest Neighbor, Decision Trees, and Random Forest model. It was hypothesized that a Random Forest model will have the greatest accuracy and lowest RMSE value since there are few parameters or input dimensions. Accuracy and RMSE are metric values generated by the models that show how well the model classifies the information alongside its' percent of error.

Shown by the data, the models were able to show great accuracy and RMSE values, but if these models were taken into production, they would not be suitable for any professional environment. This is because only three tests were conducted for each model meaning that there is not much of a testing environment for the models to show true consistency in each prediction. Based on the performance of the three trials for each of the models, the Decision Tree and Random Forest model were shown to have the highest accuracy and lowest RMSE values. Both models were able to classify exoplanets at 93% and 96% accuracy respectively, while the other two models had accuracies of 83% for regression and 80% for KNN. The number of testing enumerations as well as how the models handle the data played a role in the results shown from the models.

Based on the data collected, the hypothesis was supported that the Random Forest model would have the lowest RMSE value and greatest accuracy value. When the accuracy values are compared, the difference between Random Forest and Decision Tree is. A true comparison would have to be made with the Confusion Matrix values, as we see that the Random Forest model was much better at classifying False Positive and False Negative data. The difference between the False Positive data is 64 False Positive Classifications when Random Forest is compared to Decision Tree. A t-test was conducted in comparison of the control Linear Regression to each of the models RMSE values. Every single model had a t-test that was conducted that showed significance in terms of the importance of the statistical values showing that the comparison and assumption can be that the hypothesis is supported. Every p value was less than .00001 at a significance level of 0.05 meaning that the data is significant. The significance of the t-tests allow show how the models having different structures of handling the data have a direct effect on the accuracy and RMSE of the classifications.

The reasoning is because while the Decision Tree goes through a process of sequential steps that choose the path taken to the final prediction, the Random Forest model uses a collection of Decision Trees to provide sufficient results. A Random Forest does not rely on the feature importance of one Decision Tree when going through sequential processing but randomly chooses different features during the training process. When compared to KNN or Linear Regression, their way of handling data is through a linear process of a best of fit line that groups common data together. This process is inefficient for the dataset that was used, since the dataset was very abstract which is no suited for linear handling. This allows for the Random Forest model to generalize over the data in a better way, thus resulting in higher accuracy.

For future research involving this project, building a visual-based classifier would be the next step. Currently, the exoplanets are being classified from the numerical data that represents the planets that have been discovered from the Kepler mission. The numerical data and the classification data gathered can be used in the creation of the visual classifier, as they are all necessary pieces of data to create an accurate visual classification system. Building a visual classifier that is practical and functioning can be beneficial to future space exploration, as exoplanets are still a topic that scientists are still trying to broaden their understanding of.

Another extension of this product is evaluating more classification models. There are more classification models out there, and only four have been chosen for this project. All the models have specific ways they train and test the data to make predictions and running the same procedure on several types of models will allow for a deeper analysis into these structures. Using more algorithms can help make the data less complex which assists in creating opportunities for better predictive models.

Appendix

Full data set: https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi

References

Aigrain, S., Favata, F., & Gilmore, G. (2004, January 27). *Characterising stellar micro-variability for planetary transit searches*. Astronomy & Astrophysics. Retrieved September 23, 2021, from https://www.aanda.org/articles/aa/abs/2004/06/aa0039/aa0039.html.

Gilliland1, R. L., Chaplin2, W. J., Dunham3, E. W., Argabright4, V. S., Borucki5, W. J., Basri6, G., Bryson5, S. T., Buzasi7, D. L., Caldwell8, D. A., Elsworth2, Y. P., Jenkins8, J. M., Koch5, D. G., Kolodziejczak9, J., Miglio2, A., Cleve8, J. van, Walkowicz5, L. M., & Welsh10, W. F. (2011, October 10). *IOPscience*. The Astrophysical Journal Supplement Series. Retrieved September 23, 2021, from https://iopscience.iop.org/article/10.1088/0067-0049/197/1/6/meta.

Malik, A., Moster, B. P., & Obermeier, C. (2021, March 5). *Exoplanet Detection Using Machine Learning*. arXiv.org. Retrieved September 23, 2021, from https://arxiv.org/abs/2011.14135.

Kepler objects of interest. (n.d.). Retrieved September 23, 2021, from https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi.

NASA. (2021, October 27). *Exoplanet program: Missions*. NASA. Retrieved October 29, 2021, from https://exoplanets.nasa.gov/exep/about/missions-instruments/?page=0&per_page=40&order=position%2Basc&search=&category=160.