Detection of Exo-Planets using Machine Learning Algorithms

Nischal Chandur
PES1201800295
Dept. of ECE
PES University, Ring Road Campus
Bangalore
Email: chandur.nischal@gmail.com

1. Abstract

Since the inception of the Kepler missions, over 2000 planets have been discovered outside our solar system. In the past, possible transit-like events identified by the Kepler satellites were vetted by human inspection to eliminate obvious false positives. The goal of this project is to assess the effectiveness of machine learning as a fast, automated, and reliable means of performing the same functions without human intervention. To this end, I have used training and test data sets made up of "change in stellar luminosity". I use a combination of machine-learning methods such as ensemble learning and cross-validation techniques. The final algorithms through this approach correctly identify planets in the test data ~95 per cent of the time, although each of other possible methods on its own has a significant fraction of false positives. In practice, a combination of different methods as used in this project offers the best approach to identifying the most promising exoplanet transit candidates in the dataset, and by extension similar transit surveys.

2. Introduction

It is estimated that there are more stars in the universe than there are grains of sand in all the beaches in the world combined. The vastness of the universe is beyond the comprehension of the human mind.

Many of these stars are like our very own Sun in terms of size, age and chemical composition. These stars could very well have planets orbiting around much like our Sun does.

The detection of these planets has been the focus of NASA and other space organisations for decades now. A multitude of planets have been found since the inception of the operation.

The transit method has successfully detected many new extra-solar planets (exoplanets), with the Kepler Space Telescope being a major player (Koch et al. 2010). The success of Kepler mission depended in part on the Kepler pipeline (Jenkins et al. 2010) to process the data, identify transit signals, and validate their existence.

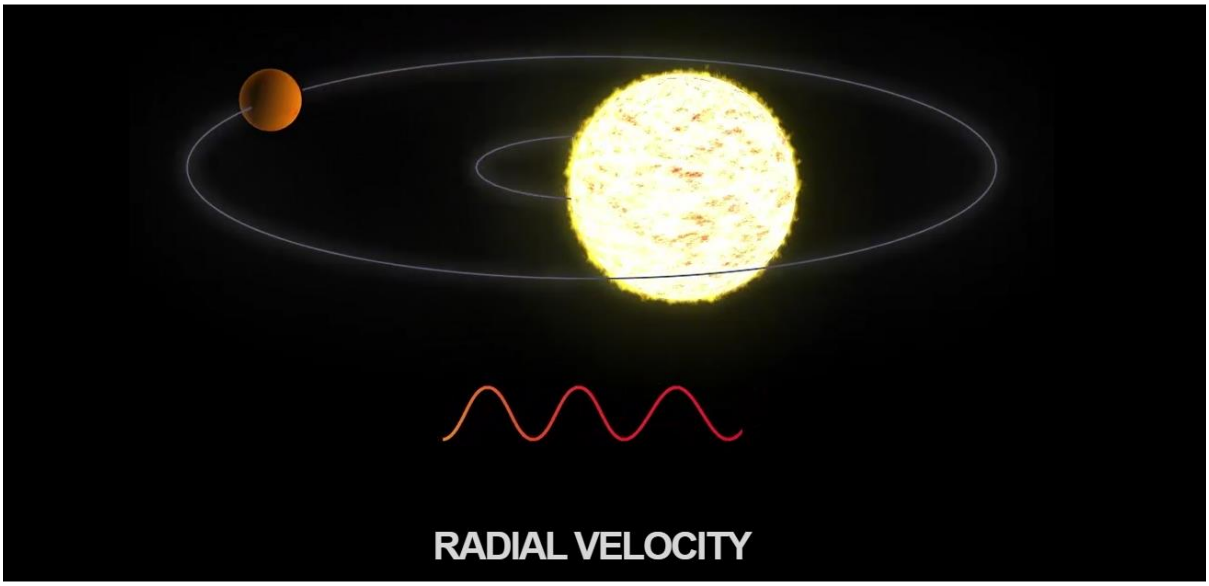
On the other hand, during recent years, machine learning becomes widely used in many fields, including astronomy (Cuevas-Tello et al. 2006). There are many types of machine-learning techniques, and some of them have already been used to analyze transit light curves. For example, k-nearest neighbors (kNN) by Thompson et al. (2015), decision tree by Coughlin et al.

(2016), random forest by McCauliff et al. (2015), and self-organizing map (SOM) by Armstrong et al. (2017).

In Section 3, various methods employed for the detection of exoplanets are introduced. In Section 4, the basic concept of machine learning is introduced and seven machine learning models are constructed. The dataset used in the project is described in Section 5. The observations and results are presented in Section 6. Conclusions are made in Section 7.

3. Methods of Detection of Exoplanets

3.1 Radial Velocity Method



The radial-velocity method for detecting exoplanets relies on the fact that a star does not remain completely stationary when it is orbited by a planet. The star moves, ever so slightly, in a small circle or ellipse, responding to the gravitational tug of its smaller companion. When viewed from a distance, these slight movements affect the star's normal light spectrum or colour signature. The spectrum of a star that is moving towards the observer appears slightly shifted toward bluer (shorter) wavelengths. If the star is moving away, then its spectrum will be shifted toward redder (longer) wavelengths.

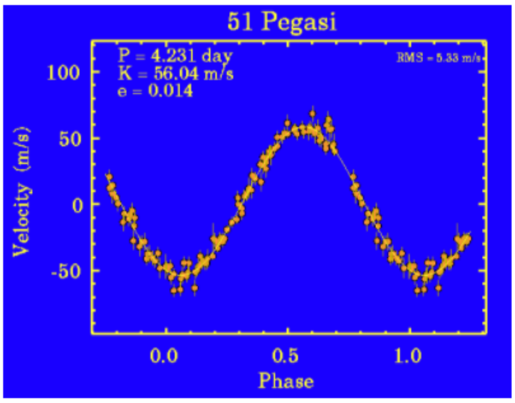
Using highly sensitive spectrographs attached to ground-based telescopes, planet hunters can track a star's spectrum, searching for periodic shifts spectral wobbles. The spectrum appears first slightly blue-shifted, and then slightly red-shifted. If the shifts are regular, repeating themselves at fixed intervals of days, months, or even years, it is almost certainly caused by a body orbiting the star, tugging it back and forth over the course of its orbit. If the body has a mass lower than about 10 times that of Jupiter (about 3,000 times the mass of Earth), then it is probably a planet. (Larger-mass objects are probably stars.)

Advantages:

- Finding big exoplanets in close orbits around the parent star.
- Measuring the mass of the exoplanet.
- Can be employed using ground-based telescopes.

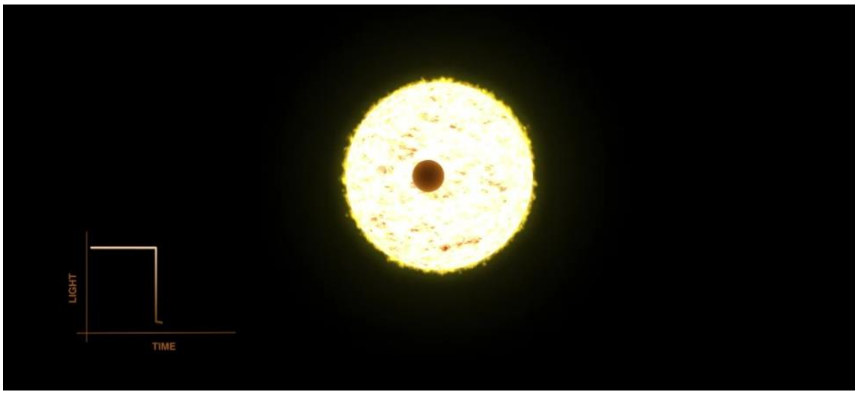
Disadvantages:

- Difficult to detect exoplanets in distant orbits.
- Measuring the radius of the exoplanets accurately is difficult.
- Finding small exoplanets is hard.



THE RADIAL VELOCITY GRAPH OF 51 PEGASI 51 Pegasi was the first exoplanet detected and confirmed. The points on the graph indicate actual measurements taken. The sinusoid is the characteristic shape of the radial velocity graph of a star rocking to the tug of an orbiting planet. *exoplanets.org*

3.2 Transit Photometry Method



When a planet passes in front of a star as viewed from Earth, the event is called a “transit”. On Earth, we can observe an occasional Venus or Mercury transit. These events are seen as a small black dot creeping across the Sun—Venus or Mercury blocks sunlight as the planet moves between the Sun and us. Kepler finds planets by looking for tiny dips in the brightness of a star when a planet crosses in front of it—we say the planet transits the star.

Once detected, the planet's orbital size can be calculated from the period (how long it takes the planet to orbit once around the star) and the mass of the star using Kepler's Third Law of planetary motion.

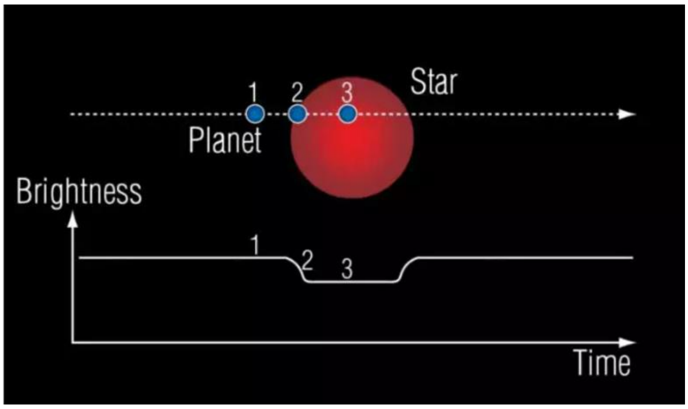
The size of the planet is found from the depth of the transit (how much the brightness of the star drops) and the size of the star. From the orbital size and the temperature of the star, the planet's characteristic temperature can be calculated. From this the question of whether or not the planet is habitable (not necessarily inhabited) can be answered.

Advantages:

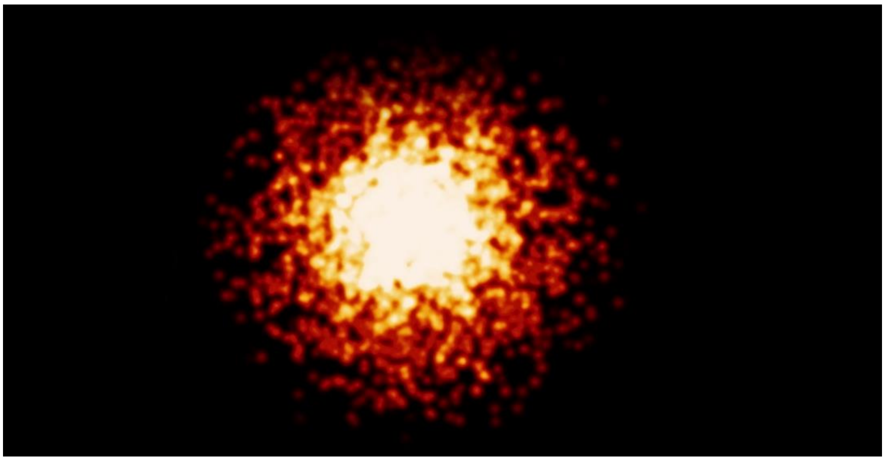
- Best for finding exoplanets in close orbits.
- Appropriate for measuring planet diameter.

Disadvantages:

- Not great for finding exoplanets that do not orbit in the plane of view from Earth. This is the case for the majority of the exoplanets.



3.3 Direct Imaging



Direct imaging of exoplanets is extremely difficult and, in most cases, impossible. Being small and dim, planets are easily lost in the brilliant glare of the stars they orbit. Nevertheless, even with existing telescope technology, there are special circumstances in which a planet can be directly observed.

In visible wavelengths, the reflected light from the planets would be swallowed up by the brilliance of the star, but at longer, infrared wavelengths, the planets' intrinsic heat glows comparatively brightly. With repeat observations, astronomers can observe the planets move in their orbits.

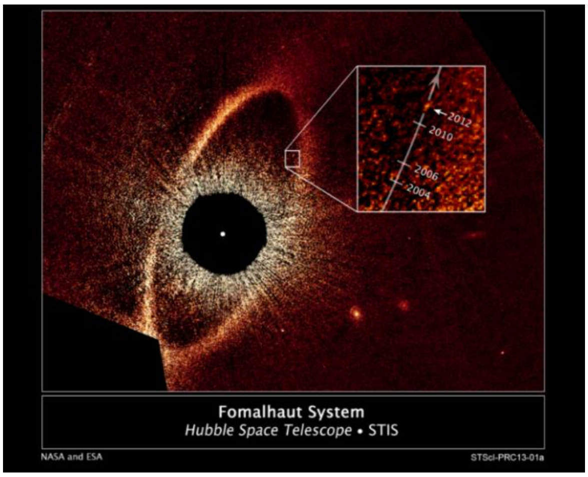
There are plans for future missions and projects that would make direct imaging easier. Ground-based telescopes with adaptive optics systems obtain sharper images, helping astronomers separate planet and star light. Ground-based or spaceborne telescopes equipped with coronagraphs can block the light from the star just like you might use your hand to shade your eyes from strong sunlight, making it easier to spot planets. And missions have been proposed to fly a star-shade in formation with the telescope, blocking starlight before it ever gets to the imaging instrument.

Advantages:

- Actually, seeing the planets.
- Find planets that do not transit the parent star.

Disadvantages:

- Difficult to find exoplanets in distant orbits.
- Hard to find exoplanets around bright stars.



3.4 Gravitational Microlensing Detection

Microlensing is an astronomical effect predicted by Einstein's General Theory of Relativity. According to Einstein, when the light emanating from a star passes very close to another star on its way to an observer on Earth, the gravity of the intermediary star will slightly bend the light rays from the source star, causing the two stars to appear farther apart than they normally would. This effect was used by Sir Arthur Eddington in 1919 to provide the first empirical evidence for General Relativity.

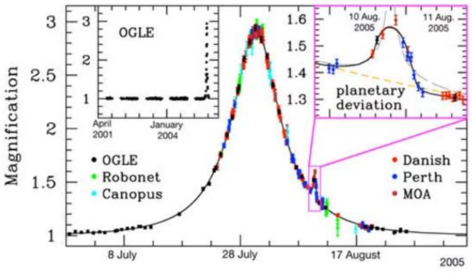
If a planet is positioned close enough to the lensing star so that it crosses one of the two light streams emanating from the source star, the planet's own gravity bends the light stream and temporarily produces a third image of the source star. When measured from Earth, this effect appears as a temporary spike of brightness, lasting several hours to several days, superimposed upon the regular pattern of the microlensing event. For planet hunters, such spikes are the telltale signs of the presence of a planet. Furthermore, the precise characteristics of the microlensing light-curve, its intensity and length, tell scientists a great deal about the planet itself. Its total mass, its orbit, and its period can all be deduced with a high degree of accuracy and probability from the microlensing event.

Advantages:

- Finding exoplanets far away from Earth
- Finding exoplanets with distant orbits around their parent stars.
- Finding free floating exo-planets.

Disadvantages:

- Discovering many exoplanets at once is difficult.
- Detecting the exoplanet more than once.



Light Curve of OGLE-2005-BLG-390

ESO PR Photo 03b/06 (January 25, 2006) © ESO

4. Machine Learning Algorithms

4.1 K-means Clustering

A data set of items has certain features and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the k-means algorithm; an unsupervised learning algorithm.

The algorithm works as follows:

- First, we initialize k points, called means, randomly.
- We categorize each item to its closest mean and we update the mean’s coordinates, which are the averages of the items categorized in that mean so far.
- We repeat the process for a given number of iterations and at the end, we have our clusters.

4.2 Decision Trees

A decision tree is a flowchart-like structure in which each internal node represents a test on a feature, each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules.

Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks.

4.3 Naïve Bayes Classification

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Algorithm:

- Convert the data set into a frequency table.
- Create a table of probabilities.
- Use the Naïve Bayesian equation to calculate the probability of each class.

4.4 Discriminant Analysis

Discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature. The term categorical variable means that the dependent variable is divided into a number of categories.

The objective of discriminant analysis is to develop discriminant functions that are nothing but the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It enables the researcher to examine whether significant differences exist among the groups, in terms of the predictor variables. It also evaluates the accuracy of the classification.

Discriminant analysis is described by the number of categories that is possessed by the dependent variable.

4.5 Support Vector Machines

Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

4.6 Cross-Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it is simple to understand and because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split.

4.7 Ensemble Learning

Ensemble methods combine several weak learning base algorithms to construct better predictive performance than a single base algorithm. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model. When we try to predict the target variable using any machine learning technique, the main causes of difference in actual and predicted values are noise, variance, and bias. Ensemble helps to reduce these factors (except noise, which is an irreducible error).

4.8 Regression Models

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

5. About the dataset

The data used here is cleaned and is derived from observations made by the NASA Kepler space telescope. The mission came to a conclusion on October 30th, 2018. The dataset was downloaded from Kaggle, a website dedicated to hosting a multitude of datasets for various machine learning projects. The datasets were prepared in the summer of 2016 by Winter Delta. The GitHub profile of this individual can be found in the reference section.

NASA open-sources the original Kepler Mission data and it is hosted at the Mikulski Archive. After being beamed down to Earth, NASA applies de-noising algorithms to remove artefacts generated by the telescope. The data - in the ".fits format" - is stored online. The dataset used in this project, contains 750 samples of stars and the change in their apparent brightness, also known as, the flux of the stars.

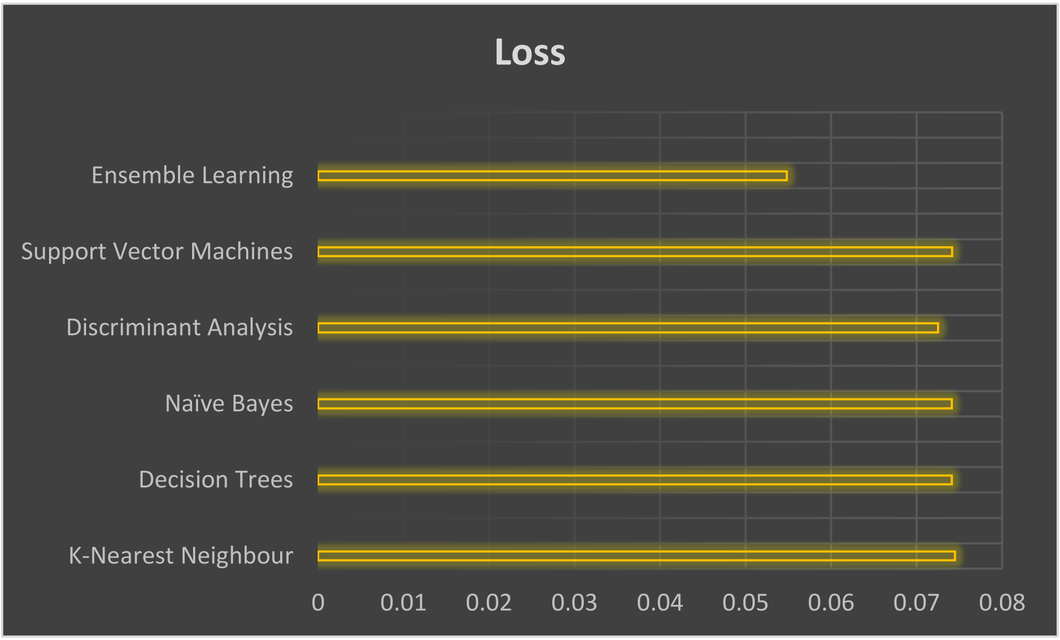
The change in flux of the stars have been noted for a period of 3196 days (8 years, 9 months and 1 day). Therefore, the dataset contains 3196 columns. The dataset is further divided into 500 stars that are meant to train the model and 250 stars that are meant to test the accuracy of the model. The response variable of the data is called LABEL and indicates whether the star in question has an exo-planet orbiting around it or not. If the value is 1, it means that there is no exo-planet and if the value of LABEL is 2, it indicates the existence of an exo-planet.

6. Observations and Results

In section 4, several machine learning algorithms used in this project have been mentioned. Employing these algorithms on the dataset obtained would give us a model that can be used to predict the existence of exo-planets around distant stars. In this section, the ML model which provides the least loss is selected. In order to make the model more robust to noise, methods such as cross-validation and ensemble learning are employed here. This ensures a reduced error while also preventing the system from overfitting the curve.

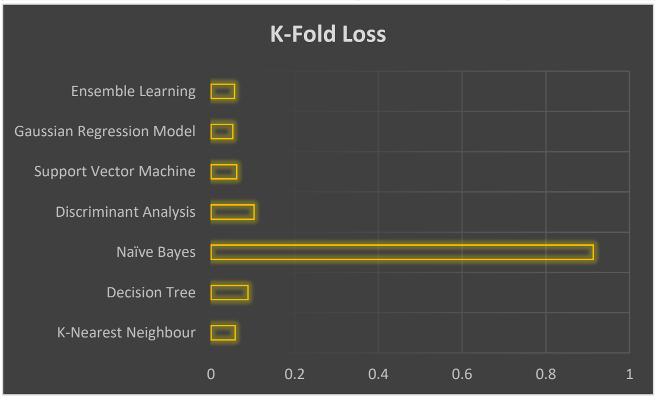
6.1 Weak Learners

METHOD OF CLASSIFICATION	PARAMETERS				LOSS
K- Nearest Neighbour Classficiation	NumNeighbours 3	Distance spearman	DistanceWeight squaredinverse	Standardize FALSE	0.0745
Decision Trees	MinLeafSize 40	MaxNumSplits 469	SplitCriterion gdi	NumVariablestoSamples 3196	0.0741
Naïve Bayes	DistributionNames kernel	Width 2.23E-06	Kernel normal		0.0741
Discriminant Analysis	Delta 995.1	Gamma 0.984	DiscrimType diagLinear		0.0725
Support Vector Machines	Coding onevsone	BoxConstraint 29.86	KernelFunction polynomial	Standardize FALSE	0.07415
Ensemble Learning	Method Bag	NumLearningCycles 187	MinLeafSize 9		0.0548



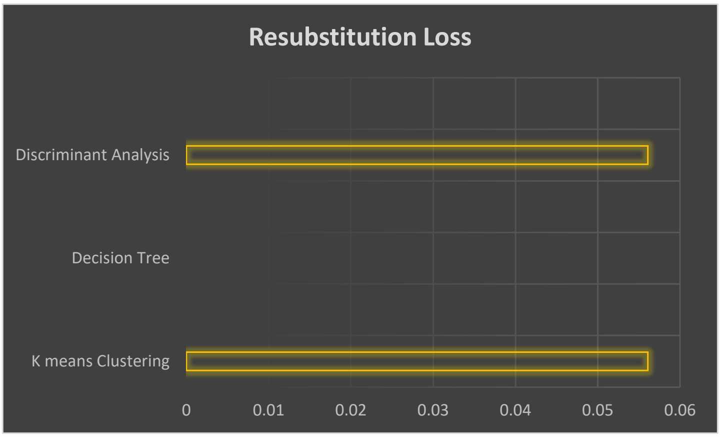
6.2 K-fold Cross Validation

Method of Classification	K-Fold	K-Fold Loss
K means Custering	32	0.0588
Decision Tree	40	0.0882
Naïve Bayes	5	0.913
Discriminant Analysis	15	0.1029
Support Vector Machine	25	0.0615
Gaussian Regression Model	24	0.0519
Ensemble Learning	10	0.602



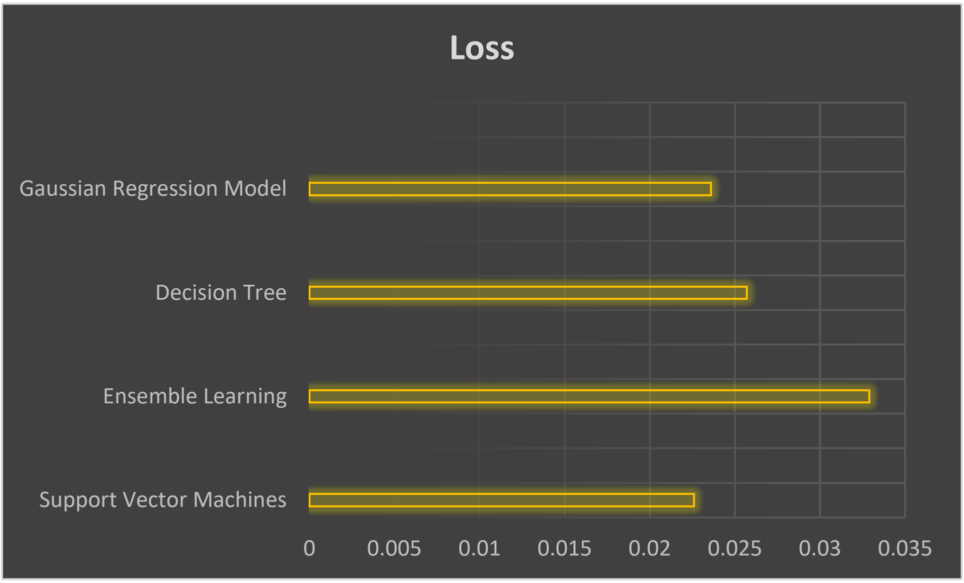
6.3 Ensemble Learning

Method of Classification	Parameters	Resubstitution Loss
K means Clustering	NumNeighbors 3	0.0561
Decision Tree	MinLeafSize 40	0
Discriminant Analysis	DiscrimType diaglinear	0.0561



6.4 Regression Models

Method of Classification	Paramaters			Loss
Support Vector Machines	Box Costraint 3	KernelScale 1.25	Epsilon 6.5	0.0226
Ensemble Learning	Method LSBoost	NumLearningCycles 16	LearnRate 0.32	0.0329
Decision Tree	MinLeafSize 63	MaxNumSplits 1	NumVariablesToSample 3196	0.0257
Gaussian Regression Model	Sigma 0.0245	BasisFunction purequadratic	KernelFunction exponential	0.0236



Results:

After carefully studying the various models and the errors in their prediction, we can choose a system with around 5% loss. Anything less, could result in the system overfitting the data and anything more could render the system inaccurate.

Therefore, to predict the existence of planets around distant stars we can use the following models:

- Ensemble Learning.
- Cross-Validation on KNN clustering, SVMs, Gaussian Regression models and ensemble learning.
- Ensemble learning with learners as KNN clustering and discriminant analysis.

*Refer the table for the loss of the aforementioned methods.

7. Conclusion

While the Kepler data alone is not of sufficient quality to definitively identify planets from the data, it has proven to be very effective in producing new candidates for future follow-up and eventual planet status.

The machine-learning framework provides a tool for the observer wanting to re-examine the full set of data holdings in any Kepler field, enabling fast re-classification of all targets showing transit-like behaviour and identification of new targets of interest.

An additional advantage of this approach is that the algorithms can be quickly re-trained as new information is added to the data set.

Using multiple machine-learning models is an effective framework that can be modified and applied to a variety of different large-scale surveys in order to reduce the total time spent in the target identification and ranking stage of exoplanet discovery. Combining the results from additional machine-learning methods could further improve the prediction.

With the launch future missions, we can be certain that the humans are going to be surveying a much larger volume of space and the number of stars observed will be in the billions. The number of light curves in these data sets is clearly beyond manual classification, so machine-learning techniques will be essential to their success.

The application and performance analysis of machine learning on current sky surveys such as Kepler and WASP are integral to the successful understanding and implementation on future large surveys.

8. References

Websites referenced include:

- https://www.nasa.gov/mission_pages/kepler/overview/index.html
- http://www.exoplanetes.umontreal.ca/transit-method/?lang=en#:~:text=The%20transit%20method%20is%20a,by%20the%20radial%20velocity%20method/
- https://exoplanets.nasa.gov/keplerscience/
- https://github.com/winterdelta/KeplerAI/
- https://www.kaggle.com/keplersmachines/kepler-labelled-time-series-data
- https://exoplanetarchive.ipac.caltech.edu/
- https://matlabacademy.mathworks.com/

Scientific Papers referenced include:

- N. Schanche, A. C. Cameron, G. Hébrard, L. Nielsen, A. H. M. J. Triaud, J. M. Almenara, K. A. Alsubai, D. R. Anderson, D. J. Armstrong, S. C. C. Barros, F. Bouchy, P. Boumis, D. J. A. Brown, F. Faedi, K. Hay, L. Hebb, F. Kiefer, L. Mancini, P. F. L. Maxted, E. Palle, D. L. Pollacco, D. Queloz, B. Smalley, S. Udry, R. West, and P. J. Wheatley, "Machine-learning approaches to exoplanet transit detection and candidate validation in wide-field ground-based surveys," *OUP Academic*, 22-Nov-2018. [Online]. Available: https://doi.org/10.1093/mnras/sty3146. [Accessed: 07-Dec-2020].

- K. A. Pearson, L. Palafox, and C. A. Griffith, "Searching for exoplanets using artificial intelligence," *OUP Academic*, 25-Oct-2017. [Online]. Available: https://doi.org/10.1093/mnras/stx2761. [Accessed: 07-Dec-2020].