National College of Ireland

Transiting Planet Search in the Kepler Pipeline Using Automated Machine Learning

MSc Research Project
Data Analytics

Martin Mohan
Student ID: X18191339

School of Computing
National College of Ireland

Supervisor: Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Martin Mohan
Student ID:	X18191339
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Catherine Mulwa
Submission Due Date:	17/08/2020
Project Title:	Transiting Planet Search in the Kepler Pipeline Using Automated Machine Learning
Word Count:	XXX
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	13th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Transiting Planet Search in the Kepler Pipeline Using Automated Machine Learning

Martin Mohan
X18191339

Abstract

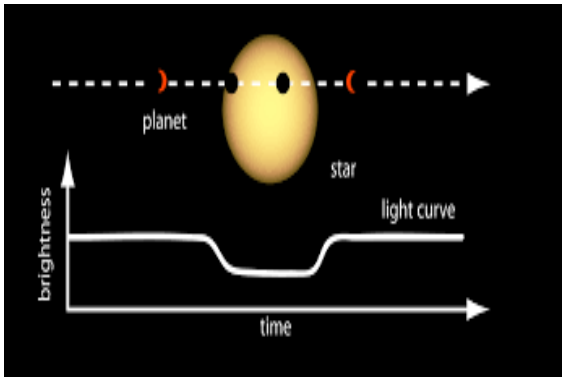
Over 2,300 confirmed planets have been found outside the solar system by the Kepler mission and although data collection finished in May 2013 data cleaning and updating has continued and new exoplanets are still being disovered using this data.
This paper applied the automated machine learning tool TPOT to the most recent kepler data release DR25. Models were trained using the *K*epler *O*bject of *I*nterest (KOI) table which contained 8,198 cases. These models were then used to search the *T*ransit *C*rossing *E*vent table (TCE), containing 34,032 cases for undiscovered exoplanets.
Using TPOT several new *P*lanetary *C*andidates (PCs) were uncovered which merit futher investigation. In addition new and existing PCs were ordered by probability of being a confirmed planets. This allows the prioritization of planets for follow-up observations.

1 Introduction

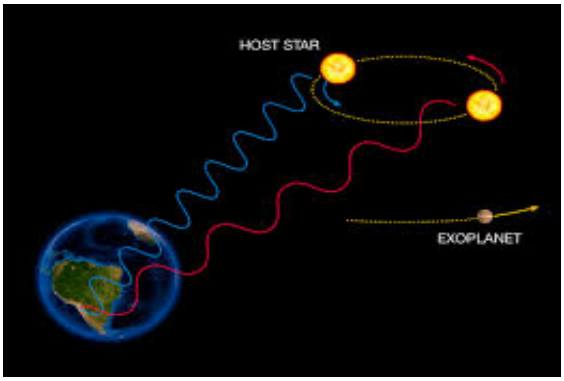
The first exoplanet discovered is considered to be 51 Pegasi b Mayor and Queloz (1995) which was detected using Radial Velocity see Figure 1. For this discovery Michel Mayor and Didier Queloz, shared the 2019 Nobel Prize in Physics. Since this discovery over 4,000 exoplanets have been discovered [1], of which over 2.300 are attributed to Kepler.
The primary Kepler mission was operational from 2 May 2009 until 11 May 2013 Kunimoto et al. (2020). The spacecraft stared at 200,000 stars in a 100 sq. degree patch of sky near Cygnus in order to measure the brightness variations using transit method. In this method exoplanet detection occurs when a planet passes in front of its sun causing a slight dimming Figure 1.

Initially PCs were found manually by astronomers using the series of steps illustrated in the Kepler Pipeline 2.1.1. As the list of planets grew, the focus of the community shifted toward population-level studies and astronomers tried to produce a more uniform exoplanet catalogue to facilitate this Shallue and Vanderburg (2018). The first machine learning was performed by McCauliff et al. (2015) who compared three different methods and found random forest to be the best. Following this several methods were applied which were a mixture of expert systems and machine learning Coughlin (2017) Shallue and Vanderburg (2018) Kunimoto et al. (2020).

[1] `https://exoplanetarchive.ipac.caltech.edu/docs/counts_detail.html` accessed 30 March 2020



(a) *Transit*: Measures dimming effect caused when planet passes in front of it's sun Accounts for *76.2%* off all planets detected



(b) *Radial Velocity: Measures wobble of the planet's sun caused by gravitational pull. Accounts for 19.3% off all planets detected*

Figure 1: Kepler has discovered over half of all known planets using *Transit* method

Humans have been replaced by a combination of machine learning, expert systems and human vetting. This report will examine four different methods which have emerged to automate exoplanet detection the autovetter 2.2.1, the robovetter 2.1.1, Astronet 2.2.2 and Kunimoto 2.1.2.
This report expanded on work by McCauliff et al. (2015) which evaluated machine learning methods to detect exoplanets in the kepler pipeline 2. In this paper the most recent data (DR25) was used and the TPOT automated machine learning package Olson and Moore (2019).

1.1 Motivation and Background

The Kepler mission was the most prolific discoverer of exoplanets. Various methods have been applied to refine the process of exoplanet detection as described in section 2. Traditional machine learning methods have been attempted with older versions of the kepler data 2.2.1 but later methods use expert systems 2.1.1, 2.1.2 or deep learning 2.2.2. This report uses automated machine learning methods to detect exoplanets on the latest Kepler data. The accurate identification of potential candidates means resources are not spent needlessly trying to confirm non-existent planets

1.2 Project Requirement Specifications

This research will contribute to the astronomical community by determining the most effective machine learning method to find planetary candidates using the latest data from the Kepler mission.The research question and objectives are as follows ...

1.2.1 Research Question

RQ: Can automated machine learning be used to improve exoplanet prediction using the latest Kepler data

Sub-RQ: How well would this methodology work on other transit missions such as K2 and TESS

1.3 Research Project Objectives

The research project objectives listed in Table 1 are specified as a solution to aforementioned research question.

Table 1: List of research objectives

	Description
1	Merge TCE and KOI tables from NASA to create a single table containing multiple IV's and a single DV
2	Cleanup table checking for mulitcollinearity and outliers in order to find the best table for predicting exoplanets
3	Use the TPOT automated machine learning tool to find and tune the best models. Some of the models generated by TPOT are ... DecisionTree, RandomForest, Gradient Boosting Classifier and LogisticRegression
4	Evaluate the performance of the models generated and select the best.
5	Use the best models to recover new planets from the TCE table. Order new and previous candidates by probability of being a confirmed planet.
6	Discussion of application to similar missions

1.4 Research Contribution

This paper evaluated machine learning models for identifying exoplanets using the latest Kepler data and is based on the gaps shown in the literary review 2.5. The research is a follow up of the paper McCauliff et al. (2015) 2.2.1 on the classification of Kepler transit candidates and the contribution to research is...

- Four tuned models were generated by TPOT based on Gradient Boost, Random Forest, Linear Regression and Decision Trees.
- Several new PCs were uncovered in the Kepler Data.
- A list of all PCs ordered by probability was created.
- An automated pipeline was produced which could also be applied to other projects.

In order to answer the proposed research question, the project followed the SEMMA methodology. The rest of the technical report is framed as follows:...
Chapter 2 describes related work on machine learning and expert systems for exoplanet detection from 2015 to 2020. The results of the work are compared and gaps identified. Chapter 3 describes the SEMMA methodology used and provides a detailed design for detecting exopanets. Chapter 4 describes the implementation of the design. Chapter 5 evaluates the performance of models retreived by TPOT and show results when these models are run against TCE data. Chapter 6 discusses the results obtained. Finally chapter 7 presents conclusion and talks about future work.

2 Related Work on Exoplanet Detection Using the Transit Method (2015-2020)

The primary Kepler mission was retired in May 2013 but over the years the Kepler data has been revised and the planetary catalogues subsequently updated but these exoplanet catalogues are never totally correct as not every planet is found and not every planet is a true planet Bryson et al. (2020). Even following release of all four years of Kepler data more planets have been uncovered by independent authors Shallue and Vanderburg (2018),Kunimoto et al. (2020).
Two types of system are used to derive the list of PCs from Kepler's 200,000 light curves. **Expert systems** 2.1 are rule based systems where the full knowledge of the expert is digitized, and is used in the decision making while **machine learning** 2.2 are based on statistical modelling of data.
After reviewing the main Kepler Pipeline 2.1.1 the review will also investigate other automated solutions which have been used successfully to discover exoplanets using Kepler data 2.2.2, 2.1.2 and will look at solutions produced for other transit missions 2.3. An attempt was made to compare the results 2.4 and finally sections 2.5 and 2.6 explained the gaps which this paper will attempt to address.

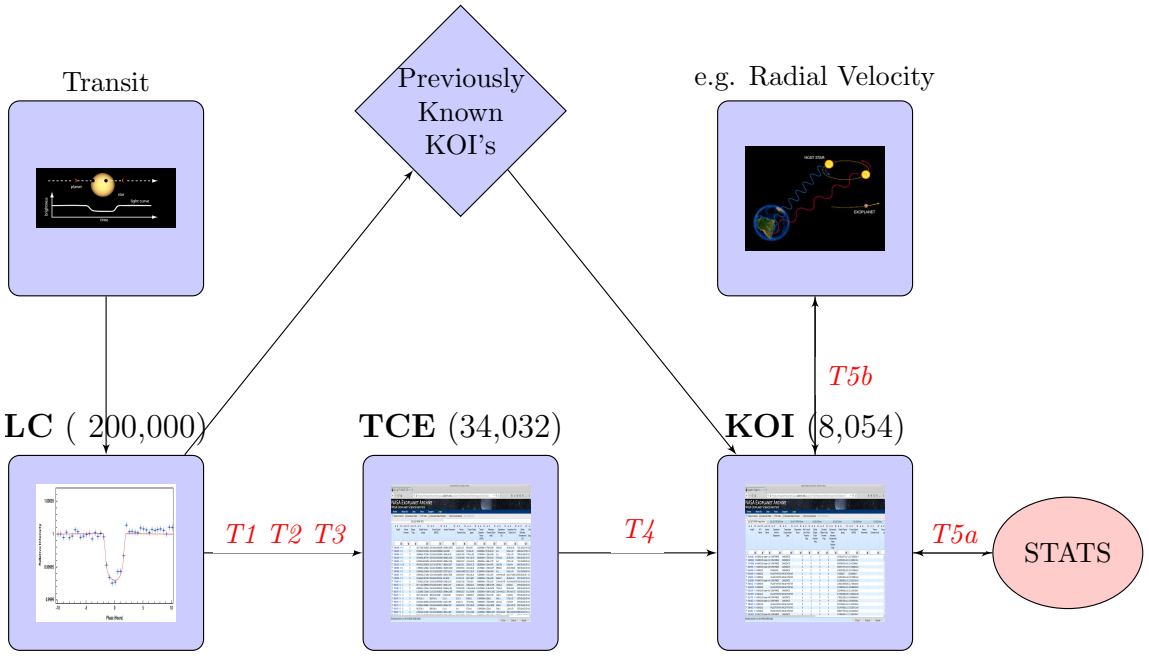
2.1 Expert Systems Used for Transit Detection on Kepler

The Kepler pipeline is an expert system. The pipeline is a series of modular software linked together where each module performs a distinct task as illustrated in Figure 2. A pipeline approach has the advantage of allowing users to replace or upgrade sections independently of the other sections. Most exoplanets were discovered using the Kepler pipeline 2.1.1 although more recently alternative pipelines have also uncovered exoplanets using this data 2.1.2.

2.1.1 Kepler Pipeline (NASA)

At the start of Kepler pipeline features were derived from the light curves using a wavelet-based algorithm (TPS) Jenkins et al. (2010) followed by Data Validation DV Wu et al. (2010). This transformed light curves into a series of tables. Dimensionality reduction was then performed using Locality Preserving Projections (LPP) and K-NN He and Niyogi (2004) Thompson et al. (2015) [2].
After light curves are identified the pipeline used two automated methods to detect exoplanets the autovetter ('machine learning' approach) 2.2.1 and the robovetter ('expert system' approach) 2.1.1. The autovetter's decision rules are 'learned' autonomously from the data, while the robovetter operates with explicitly constructed decision rules Catanzarite (2015). Both systems evolved in parallel and learnt from each other. For example, early robovetter results indicated that the autovetter was initially misclassifying some TCEs with secondary eclipses as planet candidates; by adding new attributes they improved the autovetter's ability to correctly classify secondary eclipses .In the other direction, autovetter results showed that the robovetter was too strongly rejecting candidates, which allowed the robovetter to be tuned to mitigate that problem Catanzarite (2015)

[2]LPP code for Kepler `https://sourceforge.net/projects/lpptransitlikemetric/`



Input	**Output**	*Treatment* [a]
Light Curve (**LC**)	-	*T1*: wavelet-based, adaptive matched filter software Transiting Planet Search (TPS) Jenkins et al. (2010)
-	-	*T2*: Data validation (DV) Wu et al. (2010)
-	Transit Crossing Event (**TCE**)	*T3*:LPP and K-NN Jenkins et al. (2015) He and Niyogi (2004)
Transit Crossing Event (TCE)	Kepler Object of Interest (**KOI**)	*T4*: Veto to exclude Non Transiting phenomena (e.g. instrument noise) Thompson et al. (2018).
Kepler Object of Interest (KOI)	Planet Candidate False Positive	*T5a*: Veto to exclude signals from eclipsing binaries or other more subtle machine noise Jenkins et al. (2015).
Kepler Object of Interest (KOI)	Planet Candidate False Positive	*T5b*: follow-up observation Mullally et al. (2018)

Figure 2: The Kepler pipeline and asociated Treatments

Kepler's penultimate Data Release 24 (DR24) produced two different exoplanet catalogues one based on the autovetter Catanzarite (2015) and one on the robovetter which differed slightly Coughlin (2017). The final data release **DR25** used the robovetter but not the Autovetter.

The Kepler Robovetter uses an expert system technique Coughlin (2017) which operates with explicitly constructed decision rules Catanzarite (2015). The robovetter is used to label each KOI as a Planetary Candidate (PC) or a False Positive (FP) and False Positives are further divided into four categories...

- Not Transit-Like
- Significant Secondary
- Centroid Offset
- Ephemeris Match [3]

It also produces a score ranging from 0.0 to 1.0 that indicates the Robovetter's disposition confidence, where 1.0 indicates strong confidence in PC, and 0.0 indicates strong confidence in FP. Robovetter is available on github [4]

2.1.2 Kepler Pipeline (Kunimoto)

A recent paper announced the discovery of 17 more planets using Kepler data Kunimoto et al. (2020) and had a 98.8% recovery rate of already confirmed planets. They did this by using their own independent pipeline and lowering the snr ratio. The main difference between this and Kepler's is their use of a Box-Least Squares (BLS) algorithm in contrast to the wavelet-based algorithm at the level of light curves. Their approach is "largely inspired by the automated DR25 Robovetter" 2.1.1 but with their own modifications. [5].

2.2 Machine Learning Solutions for Transit Detection

This section will look at the Autovetter 2.2.1 which was the first attempt to use machine learning on Kepler data. It will also examine an alternative machine learning method called Astronet 2.2.2 and other methods applied to ground based transit detection.

2.2.1 Kepler Machine Learning - NASA Autovetter

Autovetter McCauliff et al. (2015) was a machine learning method adopted by the Kepler pipeline which was complementary to the robovetter 2.1.1. The data used for training was taken from tables produced by the Kepler pipeline Figure 2 [6]. Different methods were used to select the attributes including removing several highly correlated attributes. The final training set Catanzarite (2015) labelled the data as follows...

[3]The ephereimis is the projected path of a celestial object

[4]`https://github.com/nasa/kepler-robovetter/blob/master/README.md`

[5]`https://github.com/mkunimoto/Transit-Search-and-Vetting`

[6]McCauliff et al. (2015) did not use the publicly available TCE table. His data contained 237 attributes based on the wavelet matched filter used by TPS, transit model fitting, difference image centroids, and some additional tests

- Planetary Candidate
- Non Transiting Phenomena (NTP) - NTP could be instrument noise = False Positive
- Astrophysical False Positives (AFP) - AFP could be a non-transiting binary star = False Positive

Three machine learning methods were tried on this Data Release DR24, Random Forest, Naïve Bayes and K-NN. The results are shown in Table 2.

Critique When choosing the autovetter only three techniques were compared (one of which was inconclusive) and although the Autovetter was used on the DR24 exoplanet catalogue investigations showed that it was not used on the DR25 catalogue which is more complete 2.1.1. This report will address this gap 2.5.

2.2.2 Kepler Machine Learning - Astronet

Astronet Shallue and Vanderburg (2018) used deep learning to detect exoplanets. Convolutional Neural Networks CNN were used to create training data directly using the Kepler light curves. After excluding exoplanets already discovered and signals which may be due to instrument noise or other scenarios they concluded statistically they had discovered 2 new exoplanets.Open source **Astronet** software was written on top of Tensorflow and is available on github. [7].
Comparison of Astronet against autovetter 2.2.1 was not possible directly because it used a different pipeline. Comparison of Astronet against the robovetter was tried but it was necessary to make several assumptions.

Critique Shallue only searched 670 multi-planet systems, which are a rich source of planets.

2.3 Other Transit Missions

The Astronet software 2.2.2 was modified and used in two other missions. After a mechanical failure on Kepler the satellite was re-configured and renamed the **K2** mission Howell et al. (2014). The *Astronet-K2* software reported 98% accuracy Dattilo et al. (2019) and uncovered two previously undiscovered exoplanets. The Astronet software was also applied to the **TESS** satellite mission Osborn et al. (2020) and an accuracy of 97.8% was reported Yu et al. (2019).
The **ground based** Wide Angle Search for Planets (WASP) has found 160 exoplanets using the transit method Schanche et al. (2019). Many machine learning methods were tested including LinearSVC, SVC, Logistic Regression, KNN, Random Forest and CNN. The data was not of good quality but they found best results were achieved using a combination of different methods including Random Forest (RF) and Convolutional Neural Networks (CNNs). Although CNN seemed to perform best alone, it did miss several planets that the RF was able to recover and occasionally let in false signals that were caught by RF or SVC, highlighting the importance of combined methods.

[7] `https://github.com/google-research/exoplanet-ml/tree/master/exoplanet-ml/astronet`

2.4 Kepler Data Processing Comparison

The results reported in literature are shown in Table 2. The measurements are explained in the caption ...

Table 2:
AUC **A**rea **U**nder **C**urve
Accuracy=1-Error rate
Recovery The number of confirmed planets recovered.

AUC	Accuracy	Recovery	Method(s)	Notes
0.9991	0.9415	-	Random Forest (Autovetter)	"We are able to achieve an overall error rate of 5.85% and an error rate for classifying exoplanets candidates of 2.81%" McCauliff et al. (2015) .
0.9894	0.9727	-	Naïve Bayes (Autovetter)	"The resulting error rates for K-NN and naive Bayes are 3.15% and 2.73%, respectively" McCauliff et al. (2015)
No rank	0.9685	-	K-NN (k=1) (Autovetter)	"optimal k (k = 1) does not produce a ranking of predictions" McCauliff et al. (2015)
-	-	98.9%	Kunimoto Pipeline	Kunimoto et al. (2020) 2.1.2
0.974	-	-	Robovetter	The robovetter is 98% reliable, although for signals with low SNR ration this falls to 50.6%. Thompson et al. (2018)
0.988	0.960	-	Astronet	Shallue and Vanderburg (2018) 2.2.2
-	0.87	-	self-organizing maps (SOMs)	This method is extremely fast (minutes) but the accuracy is not high Armstrong et al. (2017)

2.5 Identified Gaps in Exoplanet Detection Techniques

All kepler data originates from the same raw light curves but all the papers describe different processing methods applied on different parts of the Kepler pipeline. A table of results from previous papers was created Table 2 but comparison is difficult due to the aforementioned problems.
There are also disputes about accuracy. The super-earth Kepler-452b was originally confirmed with an accuracy of 99.87% Jenkins et al. (2015) but a subsequent paper disputed the discovery as they calculated the most optimistic probability at only 92% Mullally et al. (2018).
The first paper on machine learning McCauliff et al. (2015) was based on DR24 data and was used to create the NASA Autovetter 2.2.1 but this has since been superseded by DR25. *DR24* produced *18,407 TCE's* and *DR25* produced *34,032 TCE's* and corrected

many mistakes such as misclassification of PCs Seader et al. (2015). The original work done by McCauliff et al. (2015) only compared 3 methods (one of which was inconclusive) on the older DR24 data. The other machine learning system in section 2.2.2 only looked at 670 systems using the CNN method which works directly on the light curves.The other two Kepler systems described in sections 2.1.1 and 2.1.2 are expert systems rather than machine learning systems. A machine learning evaluation of the latest kepler data is missing.

2.6 Conclusion

Several techniques have been used to detect exoplanets although comparison between techniques is difficult due to differences in the way Kepler data was processed. Independent researchers have discovered new planets in the most recent kepler data using deep learning Shallue and Vanderburg (2018) and expert systems Kunimoto et al. (2020) but there has been no investigation using automated machine learning on the most recent TCE data which contains *34,032 TCE's*. A methodology to search for planets using automated machine learning is described in the next section 3.

3 Methodology

3.1 Introduction

A pipeline approach is used in the this project based on the SEMMA approach see Figure 3.

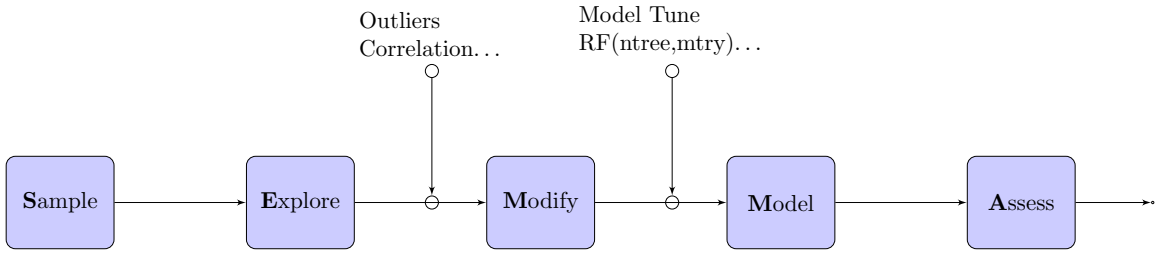
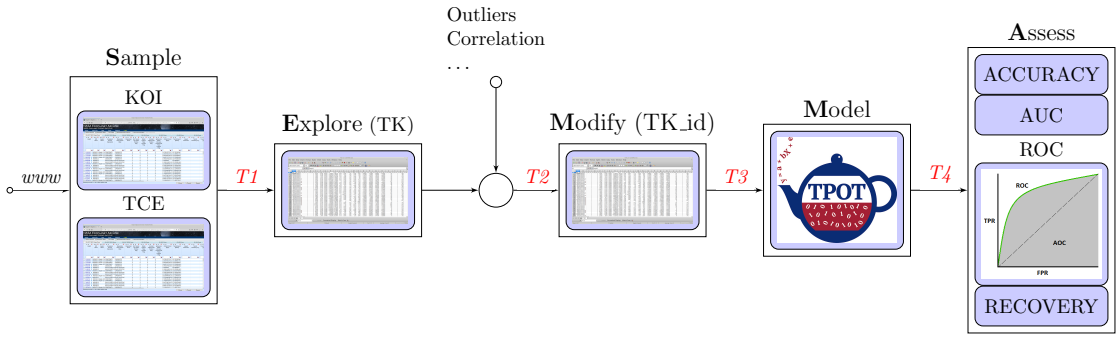


Figure 3: SEMMA (**S**ample,**E**xplore,**M**odify,**M**odel,**A**ssess)

3.2 Technical Design

Detailed design is described in Figure 4. It closely follows the SEMMA methodology described in the previous section 3.1 as does the implementation described in the next section 4.



Input	Output	Treat [a]
TCE.csv[8] KOI.csv[9]	TK.csv	*T1*: The TCE and KOI tables were downloaded from the NASA website. They were then merged and pre-processed to create a single table TK.csv containing mulitple IV's and one DV.
TK.csv	TK_**id**.csv	*T2*: After exploration TK.csv was modified by removing correlation or outliers. These modifications were identified using a unique **id**.
TK_id**.csv**	TK_id**.py**	*T3*: The file TX_id was input to the TPOT software Randal S. Olson and Jason H. Moore (2018) which performs feature selection and tuning in order to provide a **python** program with the best prediction model.
TK_id.py	Performance results	*T4* The models created by TPOT were evaluated against holdout data for Precision, Accuracy and AUC. They were also tested for overfitting.
TK_id.py	Recovery results	*T5* The best models were run against the original **TCE** file to check if any new exoplanets could be found

[a] *Treatment* refers to programs which modify or evaluate the data. *T1*, *T2*, *T3*, *T4*, *T5*

Figure 4: Treatment pipeline: Each *T*reatment has a defined input and output which is usually a csv file

4 Implementation

The pipeline design in Figure 4 was implememented and CSV files were used if possible to store data at each stage of the pipeline allowing the use of standard packages such as IBM-SPSS,Python and R.

4.1 Sample

Two csv files were downloaded from the NASA exoplanet archive: **TCE.csv** [8] and **KOI.csv** [9]. After downloading and exploration 4.2 some initial treatment was performed using a python program *Treat1.py* see Figure 4 to produce a single csv file **TK.csv** containing multiple IVs and one DV.
The initial treatment consisted of removing obsolete columns, replacing error values with signal to noise (SNR) the small amount of missing entries were imputed using 'most frequent' method. In addition nine cases marked by the tce_rogue_flag were also removed.[10].

4.2 Explore

The table generated in section 4.1 called **TK.csv** conisted of 126 IV's and 8020 cases which were mainly continuous and categorical values. The value koi_dispostion which indicated the status of the exoplanet was used as the DV see Table 3

Table 3: Distribution of koi_disposition (The Dependent Variable (DV))

koi_disposition	**Number of Cases**
FALSE POSITIVE	3963
CONFIRMED	2286
CANDIDATE	1771

4.3 Modify

4.3.1 Missing values

A small amount of missing values were found. These were imputed using "most frequent" method using the program *Treat1* see Figure 4.

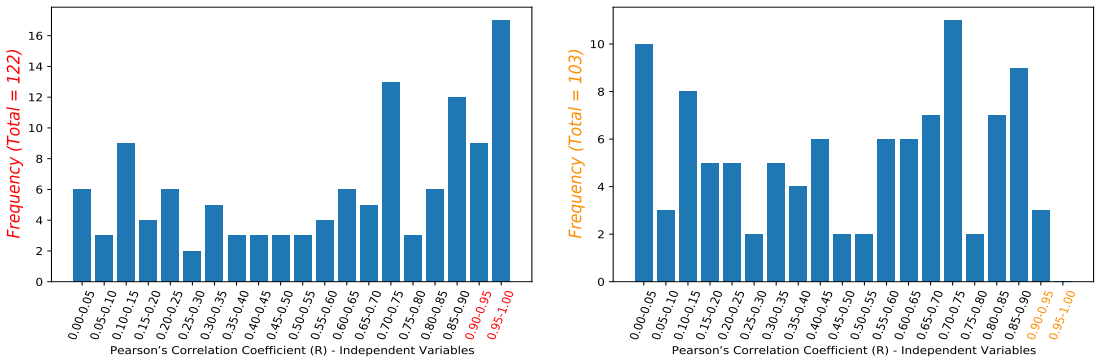
4.3.2 Multicollinearity

Multicorrelation between independent variables (IVs) is not a good idea because they are redundant and increase the size of error terms Barbara G. Tabachnick and Linda S. Fidell (2014). Previous reports found a large number of IV's with correlations were near unity McCauliff et al. (2015). High mulitcollinearity among IV's was reduced by iteratively removing IV's with Variance Inflation Factor (VIF) greater than 10 which reduced the number of IV's from 122 to 99 see Figure 5.

[8] Threshold crossing event (TCE) table: `https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=tce` last accessed 24/07/2020

[9] Kepler Object of interest (KOI) Data: `https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi`ast accessed 09/05/2020

[10]This flag indicates cases due to a NASA software bug which should not have been included.



(a) Before (b) After

Figure 5: Mulitcorrelation before and after removing IVs with VIF>10

4.3.3 Outliers

The outliers were capped using 3*(Inter Quartile Range) and data treated in this way was identified by adding the id "_cap2". The results are shown in Table 4 and the effects on accuracy were inconclusive.

4.4 Model

Automated machine learning packages automatically choose the best algorithm by training and tuning the best models and then measuring the performance. The models also perform feature selection as shown in Figure 6. These packages can match or improve upon expert human performance Waring et al. (2020). Benchmarking had identified auto-sklearn and TPOT [11] as the most promising automated machine learning packages Balaji and Allen (2018) and TPOT was chosen for this pipeline.

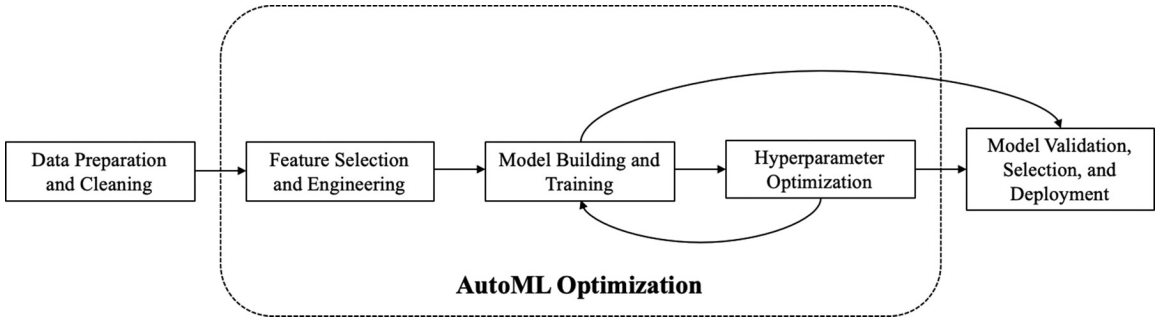


Figure 6: Automated Machine Learning

Tpot uses 10-fold cross-validation when validating the data but before Tpot was applied the data was split into 90% train and 10% test data, which was mainly used to test for overfitting.
Models are chosen by TPOT and Gradient Boost was selected as the best model using the default configuration. Different configuration options, such as TPOT light were tried to obtain more models and three other models were obtained and ordered by precision which is defind as the proportion of positive values that are truly positive Lantz (2019) see Table 4.

4.5 Conclusion

This implementation section looked at **S**ample, **E**xplore, **M**odify and **M**odel from the **SEMMA** model. The next section 5 looks at **A**ssessment of the data.

(create https://github.com/mmohan/transit-ml not x18191339@student.ncirl.ie To Do pass: WaitAndSee99)

5 Assess

A list of models found by TPOT are shown in Table 4 At each stage of data cleaning TPOT was run to find the best model and a naming convention was adopted to distinguish which models had been treated for multicorrelation (_vif) or outliers (_cap2).
Stacked models based on Gradient Boosting (GB_vif),Random Forest (RF_vif), Linear Regression (LR) and Decision Tree (DT) were chosen for further investigation. The performance results are presented in section 5.1, the recovery results in section 5.2 and these are discussed in section 6.

[11] https://epistasislab.github.io/tpot/

Table 4: List of models selected by TPOT in order of precision. All models except those generated by TPOT light (LR and DT) were overfitted.
Name (_vif=Treated for multi-correlation 4.3.2), (_cap2=outliers capped 4.3.3)

Name	**Precision CONFIRMED**	**Accuracy**	**Python sklearn model**
GB_vif_cap2	0.835	0.859	sklearn.ensemble.GradientBoostingClassifier
GB_vif	0.830	0.859	sklearn.ensemble.GradientBoostingClassifier
RF_vif	0.817	0.855	sklearn.ensemble.RandomForestClassifier
GB	0.816	0.855	sklearn.ensemble.GradientBoostingClassifier
RF_vif_cap2	0.811	0.859	sklearn.ensemble.RandomForestClassifier
GBtest	0.810	0.863	sklearn.ensemble.GradientBoostingClassifier
RF	0.806	0.855	sklearn.ensemble.RandomForestClassifier
LR	0.804	0.85	sklearn.linear_model.LogisticRegression
DT	0.793	0.844	sklearn.tree.DecisionTreeClassifier
LR_vif_cap2	0.791	0.834	sklearn.linear_model.LogisticRegression
DT_vif	0.769	0.809	sklearn.tree.DecisionTreeClassifier
BernoulliNB	0.441	0.555	sklearn.naive_bayes.BernoulliNB
KNeighbors Classifier	0.416	0.522	sklearn.neighbors.KNeighborsClassifier
GaussianNB	0.348	0.409	sklearn.naive_bayes.GaussianNB

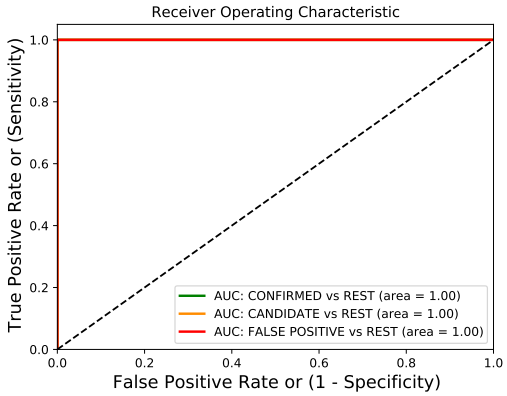
5.1 Performance Evaluation

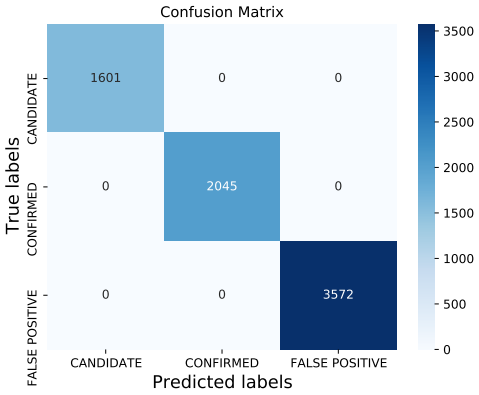
Model performance was assessed using sklearn's OneVsRestClassifier [12] Gradient Boost and Random Forest were selected by TPOT using default setting and showed similar performance. Running the model against the training data showed that **both were overfitted** although the accuracy was still good.
Logistic Regression and Decision Tree classifiers were the models chosen using TPOT light which forces the selection of simple and fast-running pipelines and **neither were overfitted**.
Detailed results for Gradient Boost and Logistic Regression are presented in sections 5.1.1 and 5.1.2 and result for Random Forest and Decision Tree are in the accompanying Technical Manual.

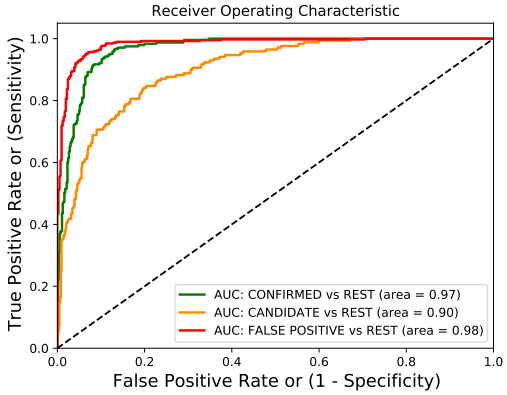
[12]OneVsRestClassifier: `https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html`

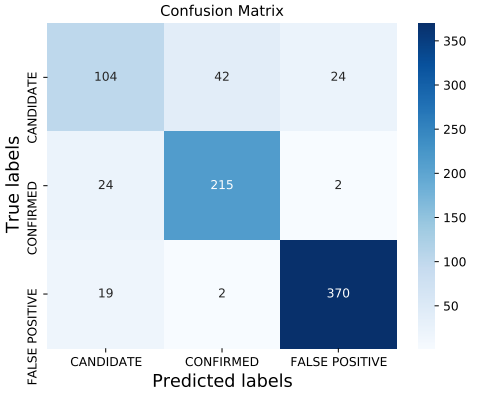
5.1.1 Gradient Boost(GB) and Random Forest(RF)

GB and RF had very similar results. GB results are shown in Figure 7 and RF results are in the Technical Manual. For both models 100% recovery was obtained with training data indicating overfitting.







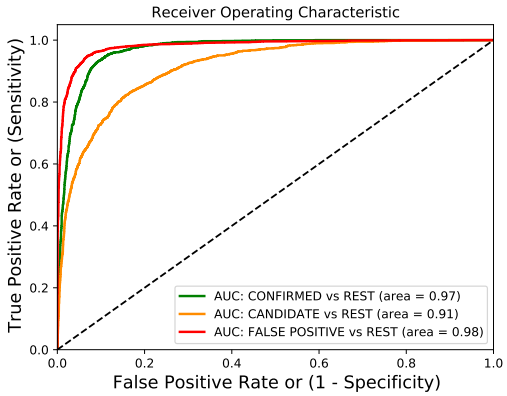


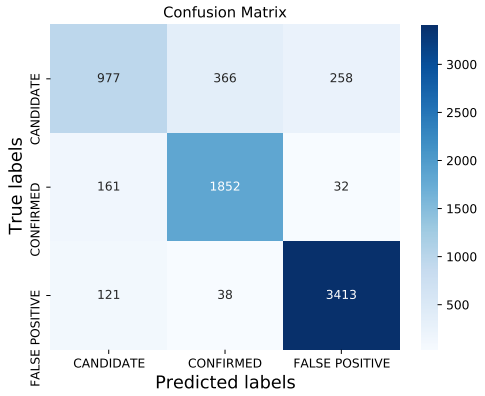
	f1-score	**precision**	**recall**	**support**
CANDIDATE	0.656	0.707	0.612	170.0
CONFIRMED	0.86	0.83	0.892	241.0
FALSE POSITIVE	0.94	0.934	0.946	391.0
accuracy	0.859	0.859	0.859	0.859
macro avg	0.819	0.824	0.817	802.0
weighted avg	0.856	0.855	0.859	802.0

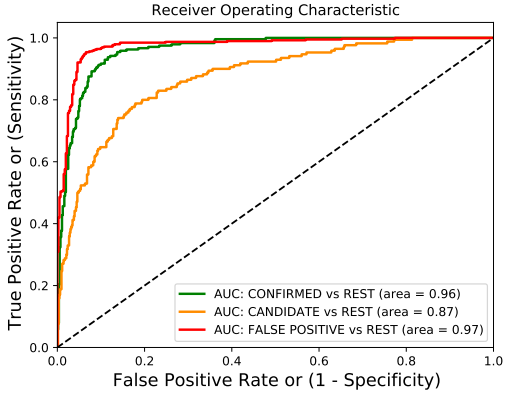
Figure 7: GBvif: Top Row: overfit test. Middle and bottom row test data

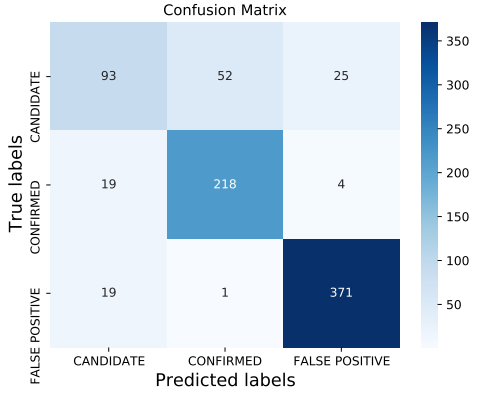
5.1.2 Logistic Regression(LR) and Decision Tree (DT)

LR and DT had very similar results. LR results are shown in Figure 8 and DT results are in the Technical Manual









	f1-score	precision	recall	support
CANDIDATE	0.618	0.71	0.547	170.0
CONFIRMED	0.852	0.804	0.905	241.0
FALSE POSITIVE	0.938	0.927	0.949	391.0
accuracy	0.85	0.85	0.85	0.85
macro avg	0.803	0.814	0.8	802.0
weighted avg	0.844	0.844	0.85	802.0

Figure 8: LR: Top Row: overfit test. Middle and bottom row test data

5.2 Recovery

Four models highlighted in Table 4 were run on the TCE to find confirmed planets. Planets designated as confirmed are displayed in sections 5.2.1 and 5.2.2 along with the original designation. The four models combined predicted 984 planets as confirmed which had not been confirmed previously.
Due to overfitting the models GB and RF Figure 9 acted more like expert systems than predictive models and very few CANDIDATES or FALSE POSITIVES were returned. The models LR and DT see Figure 10 were not overfitted and nearly one quarter of PCs were predicted to be confirmed planets.
Eclipsing binaries may be mistaken for planets see section 2.1.1 so for simplicity 213 cases in systems with eclipsing binaries [13] were exlcuded leaving 771 confirmed planets.
An extract of the results is shown Table 5 and further detailed results can be found in the Technical Manual.

Table 5: Probability of being a confirmed planet sorted by Logistic Regression >0.916

kepid	**plnt num**	**kname**	**dispos**	**LR**	**DT**	**GB_vif**	**RF_vif**
8456679	2	K00102.02	FP	0.972	0.951	-	-
8480285	1	K00691.01	CAND	0.971	0.948	-	0.503
8804455	2	K02159.02	FP	0.963	0.986	-	-
10788461	1	K03925.01	CAND	0.959	0.571	-	-
8644365	1	K03384.01	CAND	0.948	0.943	-	-
8804845	1	K02039.01	CAND	0.945	0.919	-	-
3831053	1	K00388.01	CAND	0.944	0.966	-	0.642
12505076	1	K02154.01	CAND	0.944	0.971	-	-
4149450	1	K01864.01	CAND	0.943	0.992	-	-
2581316	2	K03681.02	CAND	0.942	0.932	1.0	0.952
5709725	2	K00555.02	CAND	0.937	0.919	-	0.516
5374854	2	K00645.02	CAND	0.93	0.94	-	-
3247268	2	K01089.02	CAND	0.927	0.886	-	-
11098013	1	K02712.01	CAND	0.924	0.939	-	-
9411166	2	-	-	0.921	0.684	0.974	0.979
3641726	1	K00804.01	CAND	0.917	0.847	-	-
7938496	1	K00900.01	CAND	0.916	0.853	-	-

[13] http://archive.stsci.edu/kepler/eclipsing_binaries.html

5.2.1 Gradient Boost (GB) and Random Forest(RF)

When confirmed planets were plotted against probability using GB and RF they produced similar plots and the GB plot is shown in Figure 9.

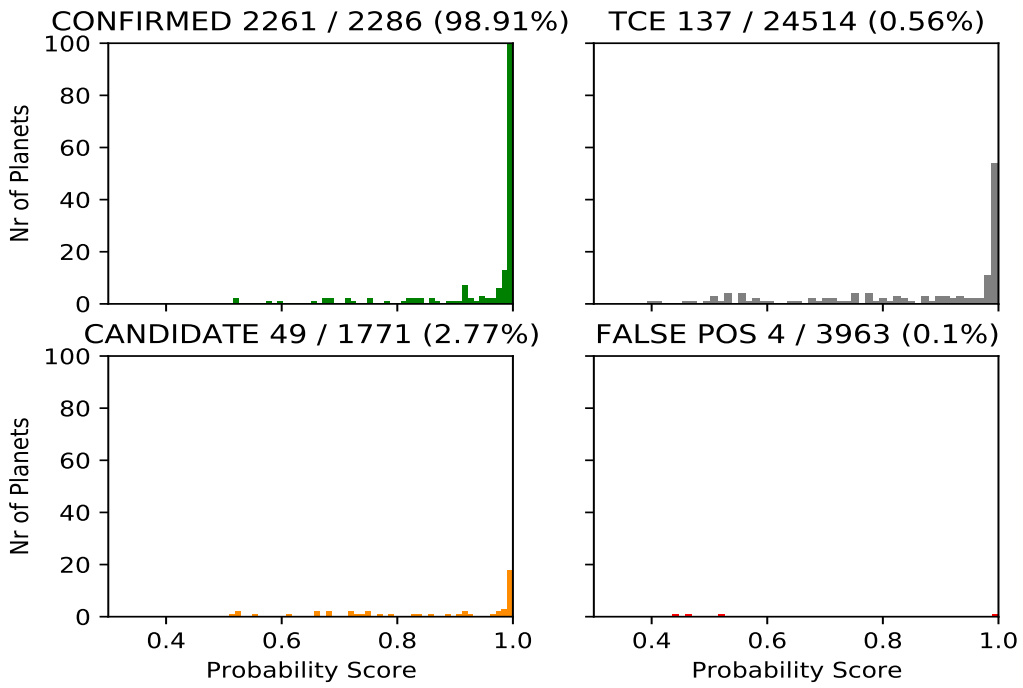


Figure 9: GB_vif: Planets predicted as CONFIRMED

5.2.2 Logistic Regression (LR) and Decision Tree (DT)

When confirmed planets were plotted against probability using LR and DT they produced similar plots and LR plot is shown inFigure 10

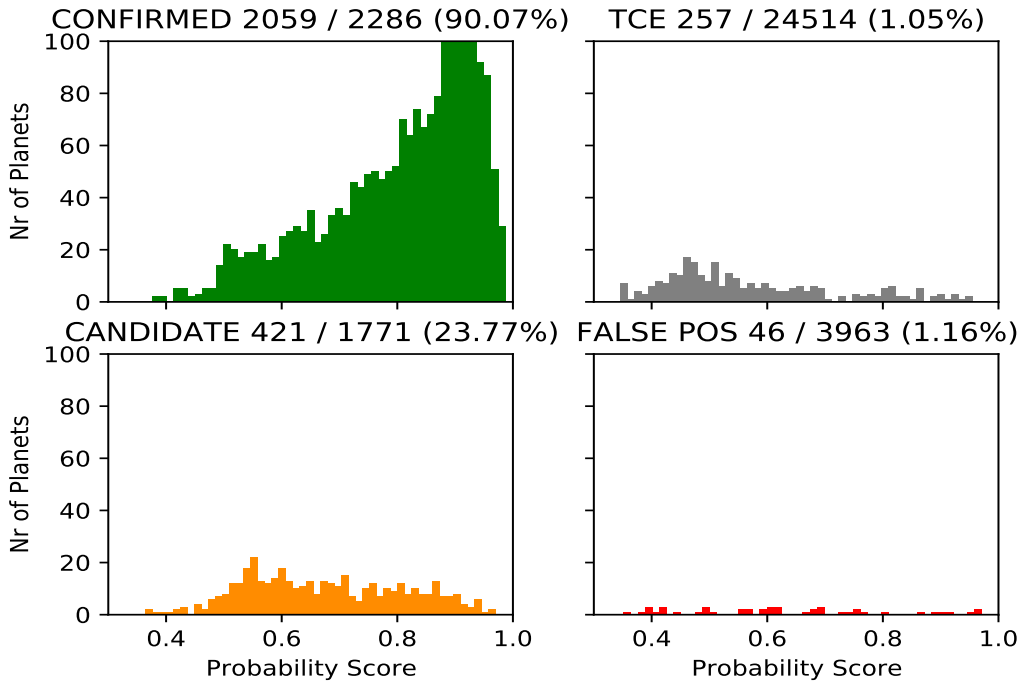


Figure 10: LR: Planets predicted as CONFIRMED

6 Discussion

The LR model predicted nearly one quarter of PCs to be confirmed planets and seventeen of these planets had a probability of 0.916 or above see Table 5.
In comparison Shallue and Vanderburg (2018) predicted four planets above 0.916 of which two were confirmed planets. They used a single deep learning model to obtain the results whereas two models (LR and DT) where used. Confirmation of planets was achieved by deeper investigation of each individual case.
Kunimoto et al. (2020) used expert systems and reported the recovery rate of planets as 98.8% which is nearly 9% better than that of the models in this report (although overfitted models had 100% recovery) see Figure 9 and Figure 10.
McCauliff et al. (2015) reported much higher AUC 0.9991 and accuracy 0.945 Table 2 than any of the models in this report see section 5.1 but the probability of PCs were not reported.
In this report PCs were sorted in order of probabilty allowing prioritization of the best PCs for further investigation. Sorting was done by LR although alternatives are available see Technical manual. The software pipeline is generic enough to be used in other projects apart from Kepler. With some modifications it could be adapted for projects such as K2 and Tess Glidden (2019) as long as enough data is available to allow population studies.

7 Conclusion and Future Work

The TPOT automated machine learning tool was applied to predict exoplanets using the latest Kepler data. Data from the NASA exoplanet archive was downloaded and four models were generated which predicted exoplanets with high probability.
Pipelines using Gradient Boost, Random Forest, Logistic Regression and Decision Trees were generated. Performance tests revealed that models using Gradient Boost and Random Forest were overfitted although they still had good accuracy. Logistic Regression and Decision Tree were not overfitted and were able to predict a quarter of PCs as confirmed planets. These PCs were ordered by quality and seventeen were predicted with high probability of being exoplanets by more than one of the models.

Future Work to confirm planets would require investigation of each PC on an individual basis using statistics or follow-up observations. Planetary confirmation by NASA require that a finding is confirmed by one or more papers in a peer reviewed journal.
The entries in Table 5 with a designated kname such as **2581316** come from the KOI table and have already been vetted. The entries which do not have a kname such as **2581316** come from the TCE table and have not been vetted. Planetary confirmation require that each case be investigated individually more deeply.
Many entries in Table 5 from the KOI table contain only LR and DT probabilities because only 10% of data from the other models was not overfitted. In future this could be fixed by trying alternatives to TPOT (auto-sklearn) or omitting CANDIDATES from the training phase. The improved software could also be applied to other similar projects.

References

Armstrong, D. J., Pollacco, D. and Santerne, A. (2017). Transit shapes and self-organizing maps as a tool for ranking planetary candidates: application to Kepler and K2, *Monthly Notices of the Royal Astronomical Society* **465**(3): 2634–2642. Publisher: Oxford Academic.
URL: *https://academic.oup.com/mnras/article/465/3/2634/2447827*

Balaji, A. and Allen, A. (2018). Benchmarking Automatic Machine Learning Frameworks, *arXiv:1808.06492 [cs, stat]* . arXiv: 1808.06492.
URL: *http://arxiv.org/abs/1808.06492*

Barbara G. Tabachnick and Linda S. Fidell (2014). *Using Multivariate Statistics: Pearson New International Edition*, number 6th Edition.

Bryson, S., Coughlin, J., Batalha, N. M., Berger, T., Huber, D., Burke, C., Dotson, J. and Mullally, S. E. (2020). A Probabilistic Approach to Kepler Completeness and Reliability for Exoplanet Occurrence Rates, *arXiv:1906.03575 [astro-ph]* . arXiv: 1906.03575.
URL: *http://arxiv.org/abs/1906.03575*

Catanzarite, J. H. (2015). Autovetter Planet Candidate Catalog for Q1-Q17 Data Release 24, p. 36.

Coughlin, J. L. (2017). Robovetter Completeness and Effectiveness for Data Release 25, p. 22.

Dattilo, A., Vanderburg, A., Shallue, C. J., Mayo, A. W., Berlind, P., Bieryla, A., Calkins, M. L., Esquerdo, G. A., Everett, M. E., Howell, S. B., Latham, D. W., Scott, N. J. and Yu, L. (2019). Identifying Exoplanets with Deep Learning. II. Two New Super-Earths Uncovered by a Neural Network in *K2* Data, *The Astronomical Journal* **157**(5): 169.
URL: *http://stacks.iop.org/1538-3881/157/i=5/a=169?key=crossref.dba1f1e9293965e63d4418f0ce5*

Glidden, A. (2019). The TESS Objects of Interest Process, **233**: 140.04. Conference Name: American Astronomical Society Meeting Abstracts #233.
URL: *http://adsabs.harvard.edu/abs/2019AAS...23314004G*

He, X. and Niyogi, P. (2004). Locality Preserving Projections, p. 8.

Howell, S. B., Sobeck, C., Haas, M., Still, M., Barclay, T., Mullally, F., Troeltzsch, J., Aigrain, S., Bryson, S. T., Caldwell, D., Chaplin, W. J., Cochran, W. D., Huber, D., Marcy, G. W., Miglio, A., Najita, J. R., Smith, M., Twicken, J. D. and Fortney, J. J. (2014). The K2 Mission: Characterization and Early Results, *Publications of the Astronomical Society of the Pacific* **126**(938): 398–408.
URL: *http://iopscience.iop.org/article/10.1086/676406*

Jenkins, J. M., Caldwell, D. A., Chandrasekaran, H., Twicken, J. D., Bryson, S. T., Quintana, E. V., Clarke, B. D., Li, J., Allen, C., Tenenbaum, P., Wu, H., Klaus, T. C., Middour, C. K., Cote, M. T., McCauliff, S., Girouard, F. R., Gunter, J. P., Wohler, B., Sommers, J., Hall, J. R., Uddin, A. K., Wu, M. S., Bhavsar, P. A., Van Cleve, J., Pletcher, D. L., Dotson, J. A., Haas, M. R., Gilliland, R. L., Koch, D. G. and Borucki, W. J. (2010). OVERVIEW OF THE KEPLER SCIENCE PROCESSING PIPELINE, *The Astrophysical Journal* **713**(2): L87–L91.
URL: *https://iopscience.iop.org/article/10.1088/2041-8205/713/2/L87*

Jenkins, J. M., Twicken, J. D., Batalha, N. M., Caldwell, D. A., Cochran, W. D., Endl, M., Latham, D. W., Esquerdo, G. A., Seader, S., Bieryla, A., Petigura, E., Ciardi, D. R., Marcy, G. W., Isaacson, H., Huber, D., Rowe, J. F., Torres, G., Bryson, S. T., Buchhave, L., Ramirez, I., Wolfgang, A., Li, J., Campbell, J. R., Tenenbaum, P., Sanderfer, D., Henze, C. E., Catanzarite, J. H., Gilliland, R. L. and Borucki, W. J. (2015). DISCOVERY AND VALIDATION OF Kepler-452b: A 1.6R SUPER EARTH EXOPLANET IN THE HABITABLE ZONE OF A G2 STAR, *The Astronomical Journal* **150**(2): 56. Publisher: IOP Publishing.
URL: *https://doi.org/10.1088%2F0004-6256%2F150%2F2%2F56*

Kunimoto, M., Matthews, J. M. and Ngo, H. (2020). Searching the Entirety of Kepler Data. I. 17 New Planet Candidates Including 1 Habitable Zone World, *The Astronomical Journal* **159**(3): 124. arXiv: 2003.04397.
URL: *http://arxiv.org/abs/2003.04397*

Lantz, B. (2019). *Machine Learning with R;Expert techniques for predictive modeling*, 3rd edn, Sage publications.

Mayor, M. and Queloz, D. (1995). A Jupiter-mass companion to a solar-type star, *Nature* **378**(6555): 355–359. Number: 6555 Publisher: Nature Publishing Group.
URL: *https://www.nature.com/articles/378355a0*

McCauliff, S. D., Jenkins, J. M., Catanzarite, J., Burke, C. J., Coughlin, J. L., Twicken, J. D., Tenenbaum, P., Seader, S., Li, J. and Cote, M. (2015). AUTOMATIC CLASSIFICATION OF KEPLER PLANETARY TRANSIT CANDIDATES, *The Astrophysical Journal* **806**(1): 6. Publisher: IOP Publishing.

Mullally, F., Thompson, S. E., Coughlin, J. L., Burke, C. J. and Rowe, J. F. (2018). Kepler's Earth-like Planets Should Not Be Confirmed without Independent Detection: The Case of Kepler-452b, *The Astronomical Journal* **155**(5): 210. Publisher: American Astronomical Society.
URL: *https://doi.org/10.3847%2F1538-3881%2Faabae3*

Olson, R. S. and Moore, J. H. (2019). TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning, *in* F. Hutter, L. Kotthoff and J. Vanschoren (eds), *Automated Machine Learning: Methods, Systems, Challenges*, The Springer Series on Challenges in Machine Learning, Springer International Publishing, Cham, pp. 151–160.
URL: *https://doi.org/10.1007/978-3-030-05318-5_8*

Osborn, H. P., Ansdell, M., Ioannou, Y., Sasdelli, M., Angerhausen, D., Caldwell, D., Jenkins, J. M., Räissi, C. and Smith, J. C. (2020). Rapid classification of TESS planet candidates with convolutional neural networks, *Astronomy & Astrophysics* **633**: A53. Publisher: EDP Sciences.
URL: *https://www.aanda.org/articles/aa/abs/2020/01/aa35345-19/aa35345-19.html*

Randal S. Olson and Jason H. Moore (2018). TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning, *Automated Machine Learning : Methods, Systems, Challenges* p. 151. Place: Cham Publisher: Springer International Publishing.

Schanche, N., Cameron, A. C., Hébrard, G., Nielsen, L., Triaud, A. H. M. J., Almenara, J. M., Alsubai, K. A., Anderson, D. R., Armstrong, D. J., Barros, S. C. C., Bouchy, F., Boumis, P., Brown, D. J. A., Faedi, F., Hay, K., Hebb, L., Kiefer, F., Mancini, L., Maxted, P. F. L., Palle, E., Pollacco, D. L., Queloz, D., Smalley, B., Udry, S., West, R. and Wheatley, P. J. (2019). Machine-learning Approaches to Exoplanet Transit Detection and Candidate Validation in Wide-field Ground-based Surveys, *Monthly Notices of the Royal Astronomical Society* **483**(4): 5534–5547. arXiv: 1811.07754.
URL: *http://arxiv.org/abs/1811.07754*

Seader, S., Jenkins, J. M., Tenenbaum, P., Twicken, J. D., Smith, J. C., Morris, R., Catanzarite, J., Clarke, B. D., Li, J., Cote, M. T., Burke, C. J., McCauliff, S., Girouard, F. R., Campbell, J. R., Uddin, A. K., Zamudio, K. A., Sabale, A., Henze, C. E., Thompson, S. E. and Klaus, T. C. (2015). DETECTION OF POTENTIAL TRANSIT SIGNALS IN 17 QUARTERS OF KEPLER MISSION DATA, *The Astrophysical Journal Supplement Series* **217**(1): 18. Publisher: IOP Publishing.
URL: *https://doi.org/10.1088%2F0067-0049%2F217%2F1%2F18*

Shallue, C. J. and Vanderburg, A. (2018). Identifying Exoplanets with Deep Learning: A Five-planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90, *The Astronomical Journal* **155**(2): 94. Publisher: American Astronomical Society.
URL: *https://doi.org/10.3847%2F1538-3881%2Faa9e09*

Thompson, S. E., Coughlin, J. L., Hoffman, K., Mullally, F., Christiansen, J. L., Burke, C. J., Bryson, S., Batalha, N., Haas, M. R., Catanzarite, J., Rowe, J. F., Barentsen, G., Caldwell, D. A., Clarke, B. D., Jenkins, J. M., Li, J., Latham, D. W., Lissauer, J. J., Mathur, S., Morris, R. L., Seader, S. E., Smith, J. C., Klaus, T. C., Twicken, J. D., Cleve, J. E. V., Wohler, B., Akeson, R., Ciardi, D. R., Cochran, W. D., Henze, C. E., Howell, S. B., Huber, D., Prša, A., Ramírez, S. V., Morton, T. D., Barclay, T., Campbell, J. R., Chaplin, W. J., Charbonneau, D., Christensen-Dalsgaard, J., Dotson, J. L., Doyle, L., Dunham, E. W., Dupree, A. K., Ford, E. B., Geary, J. C., Girouard, F. R., Isaacson, H., Kjeldsen, H., Quintana, E. V., Ragozzine, D., Shabram, M., Shporer, A., Aguirre, V. S., Steffen, J. H., Still, M., Tenenbaum, P., Welsh, W. F., Wolfgang, A., Zamudio, K. A., Koch, D. G. and Borucki, W. J. (2018). Planetary Candidates Observed by Kepler . VIII. A Fully Automated Catalog with Measured Completeness and Reliability Based on Data Release 25, *The Astrophysical Journal Supplement Series* **235**(2): 38. Publisher: American Astronomical Society.
URL: *https://doi.org/10.3847%2F1538-4365%2Faab4f9*

Thompson, S. E., Mullally, F., Coughlin, J., Christiansen, J. L., Henze, C. E., Haas, M. R. and Burke, C. J. (2015). A MACHINE LEARNING TECHNIQUE TO IDENTIFY TRANSIT SHAPED SIGNALS, *The Astrophysical Journal* **812**(1): 46. Publisher: IOP Publishing.

Waring, J., Lindvall, C. and Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare, *Artificial Intelligence in Medicine* **104**: 101822.
URL: *http://www.sciencedirect.com/science/article/pii/S0933365719310437*

Wu, H., Twicken, J. D., Tenenbaum, P., Clarke, B. D., Li, J., Quintana, E. V., Allen, C., Chandrasekaran, H., Jenkins, J. M., Caldwell, D. A., Wohler, B., Girouard, F., McCauliff, S., Cote, M. T. and Klaus, T. C. (2010). Data validation in the Kepler Science Operations Center pipeline, San Diego, California, USA, p. 774019.
URL: *http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.856630*

Yu, L., Vanderburg, A., Huang, C., Shallue, C. J., Crossfield, I. J. M., Gaudi, B. S., Daylan, T., Dattilo, A., Armstrong, D. J., Ricker, G. R., Vanderspek, R. K., Latham, D. W., Seager, S., Dittmann, J., Doty, J. P., Glidden, A. and Quinn, S. N. (2019). Identifying Exoplanets with Deep Learning III: Automated Triage and Vetting of TESS Candidates, *The Astronomical Journal* **158**(1): 25. arXiv: 1904.02726.
URL: *http://arxiv.org/abs/1904.02726*