Hunting for Exoplanets

Habbash Nassim (808292) Villa Giacomo (807462)

Project for the Data and Text Mining course, A.Y. 2019-2020
University of Milano-Bicocca
(e-mail: n.habbash@campus.unimib.it
e-mail: g.villa48@campus.unimib.it).

Abstract: NASA's Kepler Space Observatory is a space telescope launched in 2009 to search for exoplanets in far-away star systems. With the data collected in the mission, an exploratory analysis has been performed to point out any peculiarity in the data, following by a clustering analysis and comparison of different models for the supervised classification task consisting in identifying candidate exoplanets.

Keywords: exoplanets, habitable planets, kepler, analysis, eda, clustering, prediction, machine learning

1. INTRODUCTION

Humanity has looked up and searched the sky since time immemorial. The beginning of space exploration by probes and satellites and advancing technologies has made possible to conduct research on star systems lying hundreds of thousands of light-years away from Earth's proximity.
One of the most recently active fields of research is exoplanetology, a branch of astronomy dedicated to the research of extrasolar planets, motivated by the inquiry about Earth-like planets and the search for extraterrestrial life.
The task of collecting data from far-away systems presents many and varied hardships. The main issue regards data collection itself: while telescopes are able to capture images of stars, galaxies and nebulae, capturing images of exoplanets directly is incredibly hard, as the light reflected by exoplanets is almost always out-shined by their host star brightness.
In the years many different methods of indirect detection have been developed, with varying degrees of success; the Kepler Mission employs the so-called Transit Method: it detects exoplanets by measuring the dimming in brightness that happens when a planet orbits in front of its host star as seen in Fig. 1.
This technique has two main disadvantages: first, the observer (the telescope) has to be aligned to the orbital plane of the planet to be able to detect this event, and second, it takes many observations, spanning possibly many years, to be able to confirm the candidacy of a planet, as this method produces a high rate of false detections.
Disadvantages notwithstanding, this method has produced the highest rate of exoplanets discoveries, as just by scanning for star-systems in large areas of the sky it is possible to produce thousands of candidates, and also, this method allows for some direct inferences on the planets sizes and element composition thanks to techniques such as spectral analysis.

The main intent of the following project is to conduct a satisfiable data mining and analysis on the data collected by the **Kepler Mission** on the so-called Kepler Objects of Interest, candidate exoplanets found by the space telescope during its travel. Following are the main areas inspected:

(1) Data analysis on the concerned dataset
(2) Clustering on stellar and planetary features
(3) Exoplanets supervised classification task

Fig. 1. Exoplanet detection from the dipping in the light-curve during a transit [Google Brain Team, 2018]

2. RELATED WORK

Research work in this field is extensive, discovering a plethora of new confirmed exoplanets and candidate exoplanets with increasing accuracy [Batalha et al., 2013], [Borucki et al., 2011]. There are plenty of different approaches to the research, such as statistically-driven explorations finding kinds of exoplanets according to the chosen metrics of study [Barclay et al., 2013], [Teske et al., 2018].
In the last years there's been a shift in focus in the community from simple analysis to population-studies, thanks to the maturation of the Kepler mission and its accumulation of data: the heterogeneous nature of traditional exoplanets catalogues, having thousands of possible candidates with varying features and degrees of certainty makes them hard to be vetted on a population level. This change in shift saw the rise of different automatic vetting and classifier approaches in the community, such as decision trees-based models applications Coughlin et al. [2016], random forest-based McCauliff et al. [2015], and neural networks Pearson et al. [2017]. While the field is working towards population studies in a more standardized way, there still seems to be many heterogeneities in the way these studies are approached, and as such it's difficult to qualitatively assert

their effectiveness: Kepler's data is indeed extensive, but many approaches still rely on different subsets of the data, bringing thus different outcomes. At last, domain expertise is vital in this field, as technical knowledge on the probes and telescope hardware itself and conceptual knowledge in astronomy and exoplanetology are able to point out effective ways to model tasks and explore data.

3. DATA EXPLORATION

3.1 Dataset Structure

The data of the *Kepler Mission* is released by NASA in the Exoplanet Archive. The dataset used in this project is the Kepler Objects of Interest Table, which contains the cumulative records of all observed Kepler Objects of Interest, containing data on an approximate 10000 possible exoplanets [NASA, 2017]. The dataset is composed by a total of 9564 records and 50 features. The features are subdivided by the Exoplanet Archive in sets, corresponding to different kinds of measurements and assessments on the exoplanet, such as:

- **Transit properties**: consisting of different transit measurements and planetary parameters estimates inferred from the transits.
- **Stellar parameters**: estimates of the host-star parameters.
- **KIC parameters**: Kepler Input Catalog parameters, consisting of the celestial coordinates for the exoplanet
- **Dispositions**: consisting of a set of determined classes and flags.

The **Disposition** set of features presents two target classes for each exoplanet:

- **empirical_state**: represents the classification outcome of the exoplanet by the automated Kepler pipeline, and can assume the values of [Candidate, False Positive]
- **state**: represents the classification outcome of the exoplanet just as before, but incorporates peer-reviewed information by researchers, and can assume the values of [Confirmed, Candidate, False Positive]

The additional flags contained in the Disposition set represent the occurrence of events during the data collection for a certain exoplanet that weight positively towards said exoplanet being a misclassified exoplanet (a False Positive). These flags are spuriously collected and sometimes vetted manually by researchers in the archive. These are boolean attributes and are: *light_curve_consistent* (consistency of the light curve as shown in Fig. 1), *secondary_events* (eclipses or other unexpected events), *star_signal* (obfuscation by a nearby star signal) and *flux_contamination*.

3.2 Feature selection

A preliminary analysis of the dataset has deemed necessary a step of feature selection, as the dataset presented many attributes regarding **measures errors** of different other features, either relating to hardware or best-fitted parameters confidence bounds. These error values might be of use in the analysis of the **effectiveness** of a measurement method in the presence of a time-series for the type of measurement, or to smooth and regulate the weighting of certain measurements with respect to the others, but for the purpose of the project they have been deemed unnecessary, and thus have been deleted. As such it's assumed that the referenced attributes each have around the same bounds of confidence in correctness.

Another type of discarded feature are nominal attributes such as KOI Id, KOI Name and Kepler Name, which do not possess any kind of intrinsic information to be analyzed, and also presented varying degrees of missing values. The remaining features selected for the analysis, are:

- **Transit and planetary parameters**: *planet_radius*, *planet_temperature*, *orbital_period*, *transit_epoch*, *impact_parameter*, *transit_duration*, *transit_depth*
- **Stellar parameters**: *star_radius*, *star_temperature*, *star_gravity*
- **Coordinates**: *ra* (Right Ascension), *dec* (Declination)
- **Dispositions**: *state*, *empirical_state*, *secondary_events*, *star_signal*, *flux_contamination*, *light_curve_consistent*

3.3 Missing Value Replacement

Of the 9564 records only 363 (3.79%) instances present missing values on one or more columns. The values are always missing in blocks of the same features, possibly attributing the issue with a lack of confirmed measures for some exoplanets. The missing data has been thus interpreted as a **Missing At Random** case. For replacement, two different approaches have been used: the first applies to target attributes having high correlations to others. For those, a **linear regression** has been modeled, choosing those attributes that maximize the correlation as independent variables. As such, the number of parameters in the regression may for each missing attribute. As the inference of the regression sometimes produced erroneous values for the attribute's domain (such as a negative values for *planet_radius*), these outlying values are subsequently replaced by a **trimmed mean** (calculated without taking into account the $\frac{1}{4}$ and $\frac{3}{4}$ quantiles) of the attribute itself, further distorted by a random value between 1 and 15%. The second approach applies to those attributes that do not present any appreciable correlation to other attributes. These have been replaced by a **trimmed mean** of the attribute itself, as explained above. Following is a list of the processed attributes:

- **Linear Regression**: *planet_radius*, *planet_tempeature*, *insolation_flux*, *star_temperature*, *star_gravity*, *star_radius*
- **Trimmed mean**: *impact_parameter*, *transit_depth*

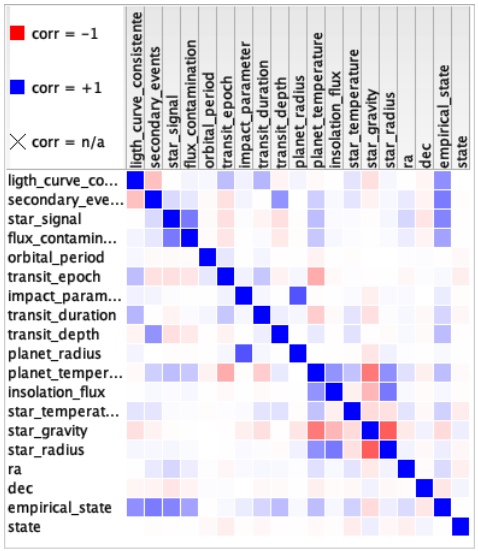


Fig. 2. Correlation matrix

4. DATA ANALYSIS

The distribution of the *state* attribute, as reported in Fig. 3 shows a fairly balanced dataset, where half of the exoplanets have been identified as False Positives and the other half as a mixed-bag of Candidates and Confirmed exoplanets.

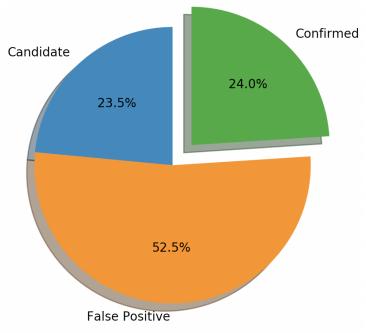


Fig. 3. Distribution of the classes in the *state* attribute

Before delving into a more in-depth analysis of the data, it has been deemed necessary to understand how do the flags present in each exoplanet interact with the classification result. Fig. 4 shows that the number of flag highly correlates with the identification of False Positives exoplanets. As the number of active flags increases, the possibility of observing celestial bodies that could actually be exoplanets decreases.

Plotting the observed exoplanets on a cartesian plane allows to observe their spatial distribution, as shown in Fig. 5. The plot does present a checkerboard pattern as a result of the Kepler probe's own sensors. It is notable how some areas appear to be fuller than others (e.g. the bottom area of plot), which is due to those observations taking space on the galactic plane, and as such containing a higher celestial bodies density [inasmuch NASA, 2009b, Section: What are the Squares on the Kepler Mission Star Field?].

The **Star Parameters** of each exoplanets, *star_temperature*, *star_gravity* and *star_radius*, can allow some additional

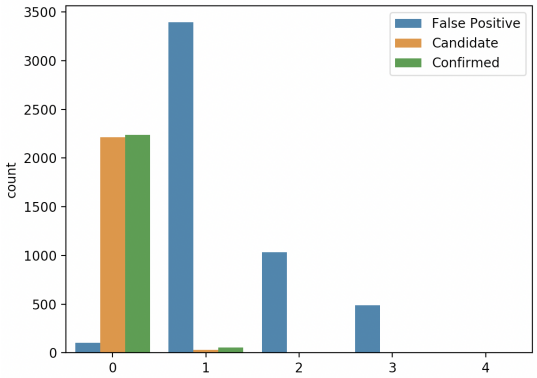


Fig. 4. Distribution of number of flags for each class in state. On the x axis the number of flags is presented with three different bars identifying the count of corresponding classes

deductions and analysis to be conducted. The pair plot in Fig. 7 plots the pairwise relationship of the attributes compared to their target class, showing a how just the star parameters exhibit a good discriminating ability for each pair of features in regards to the exoplanet target class. Confirmed exoplanets and False Positive exoplanets are somewhat separated in different regions, while Candidate exoplanets do appear in various parts of both regions.

From the pair plot it's possible to deduce that exoplanets tend to orbit around stars of smaller size but with a higher gravity and not relatively hot. The graph shows some outliers candidate KOI, which does seem to be somewhat far from the confirmed ones in terms of size, possibly making them future False Positives.

Based on the *star_temperature* it's possible to analyze the **spectral class** of the actual star, as defined by the **Harvard spectral classification system**. The system classifies stars seven classes, in descending order, O ($\geq$ 30000K), B (10000–30000K), A (7500–10000K), F (6000–7500K), G (5200–6000K), K (3700–5200K), M (2400–3700K) based on their temperature. Fig. 6 shows the distribution of the target classes and spectral classification.

The analysis does show a tendency for detected exoplanets of any class to orbit around stars of class F, G and K, which have temperatures ranging from 3700K to 7500K (note, the Sun's temperature is 6000K, putting it between classes

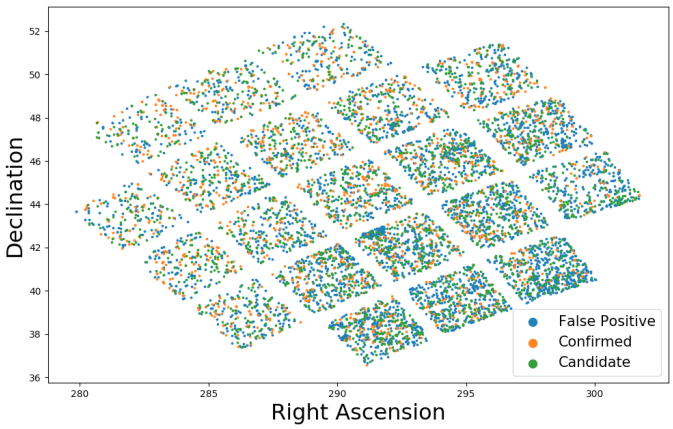


Fig. 5. Spatial distribution of the exoplanets

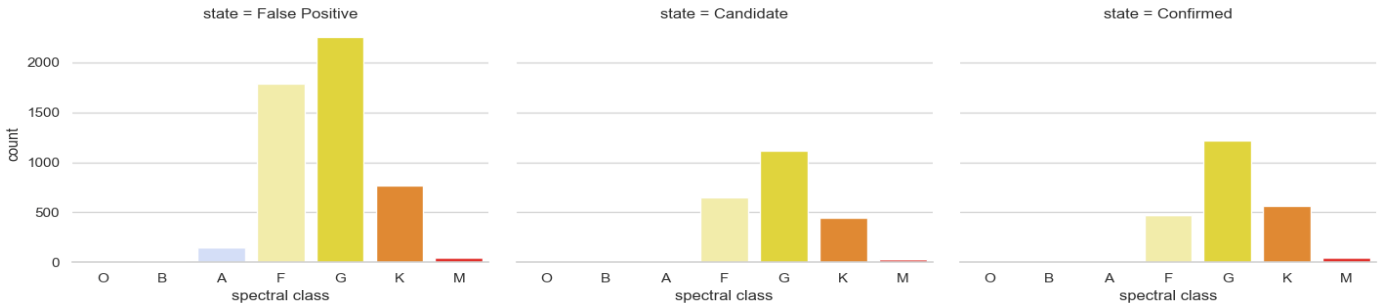


Fig. 6. Spectral class distribution based on *state*

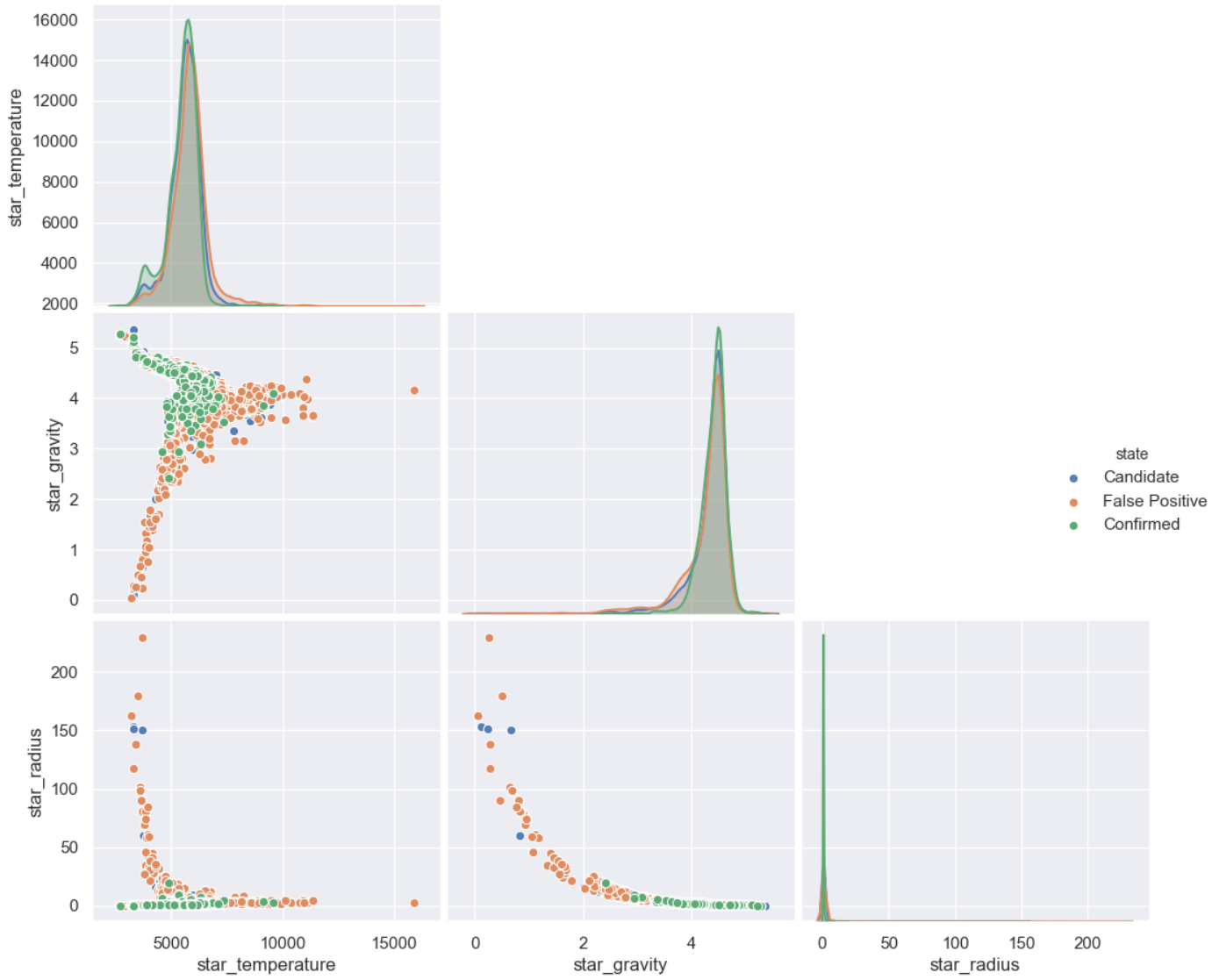


Fig. 7. Pair plot of the features *star_temperature*, *star_gravity* and *star_radius*. On the diagonal plots, the distribution of the classes for each pair of features. On the other plots, a scatter plot of the jointed features.

F and G). Class O to B do not seem to have exoplanets (except outliers) orbiting around them, possibly due to the extremely high temperature and gravity.

Fig. 8 shows how the spectral class relates to *stellar_gravity*, *stellar_radius*: spectral classes are clearly delimited by gravity - it is also notable how highly correlated *star_temperature* is to star_gravity during the preliminary analysis. The investigations shown in Fig. 7, Fig. 6 and Fig. 8 do give interesting empirical confirmations of astronomical concepts such as **circumstellar habitable zones**, as

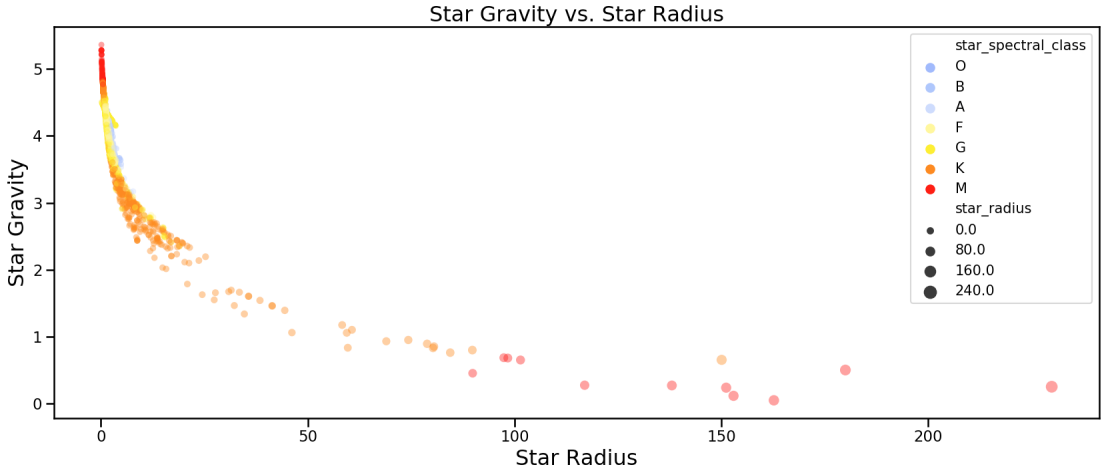


Fig. 8. Scatter plot of each exoplanet host-star's *stellar_gravity*, *stellar_radius* and *spectral class* for Confirmed exoplanets

the data seems to confirm that Earth-like planets happen to form around mid-sized stars like the Sun.

5. CLUSTER ANALYSIS

Clustering has been applied on two parallel paths: **stellar features clustering** and **planetary features clustering**. For both cases the K-Means algorithm has been used. Given a training set of points in the feature space, K-Means works by randomly initializing k cluster centroids, and then iteratively recomputing each cluster centroid until convergence. To identify the optimal number of clusters k, the two different types of indices have been tried and compared: **elbow method** and the **silhoutte coefficient**.

The elbow methods consists the comparison of intra-cluster distance relative to inner-cluster distance per number of clusters. Finding the optimal number of clusters consists in finding the bending spot in the graph (hence elbow), that maximizes cluster density.

$$W_k = \sum_{r=1}^{k} \frac{1}{n_r} \cdot D_r$$

Where:

- k is the number of the clusters.
- n_r is the number in cluster r.
- D_r is the sum of distances between all points in a cluster.

$$D_r = \sum_{i=1}^{n_r - 1} \sum_{j=1}^{n_r} ||d_i - d_j||_2$$

The silouhette coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each element. The closer the coefficient is to 1 the better the more cohesive the cluster structure is.

$$Silhoutte(C_m) = \frac{1}{|C_m|} \cdot \sum_{i \in C_m} \frac{b(i) - a(i)}{max(b(i), a(i))}$$

Where:

- $b(i)$ is the mean distance between the *i-th* element and the elements of the nearest cluster that the sample is not a part of

$$b(i) = min_{k \neq m} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

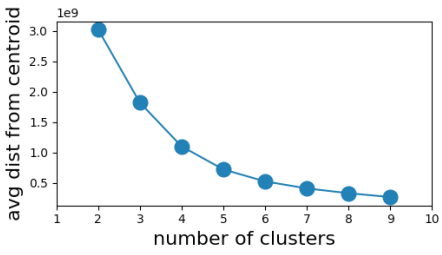
- $a(i)$ is the mean distance between the *i-th* element and every other element in the same cluster

$$\frac{1}{|C_m - 1|} \cdot \sum_{j \in C_m | i \neq j} d(i,j)$$

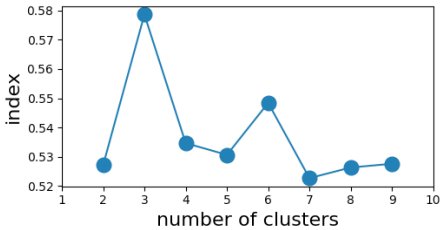
The elbow method is often ambiguous, making hard to identify the correct k, and not very reliable; as such the measure has been used in conjunction with the silhouette coefficient to justify the selection of k, while giving a higher weight in the selection of k to the silhouette.

5.1 Stellar Clustering

The feature used to analyze stellar clustering are *star_temperature*, *star_gravity* and *star_radius*. The silhoutte coefficient has been computed for k ranging from 2 to 8, finding the optimal number of cluster to be 3.



(a) Elbow plot



(b) Silhouette coefficient

Fig. 9. Silhouette coefficient and elbow method plot at varying k for stellar clustering, k=3 was selected from the silhouette peak, confirmed in the range of the bend for the elbow plot

Table 1 shows how the biggest difference between the centroids concerns the temperature of the stars; in particular

Centroids				
	temperature	**gravity**	**radius**	**count**
Cluster 1	*5935.37*	*4.33*	*1.23*	*6703*
Cluster 2	*7871.81*	*3.97*	*2.39*	*346*
Cluster 3	*4803.10*	*4.29*	*2.91*	*2514*

(a) Table 1. Stellar clustering centroids

the values of the centroids belong to difference Harvard classes:

- Cluster 1: class G (5200–6000K)
- Cluster 2: class A (7500–10000K)
- Cluster 3: class K (3700–5200K)

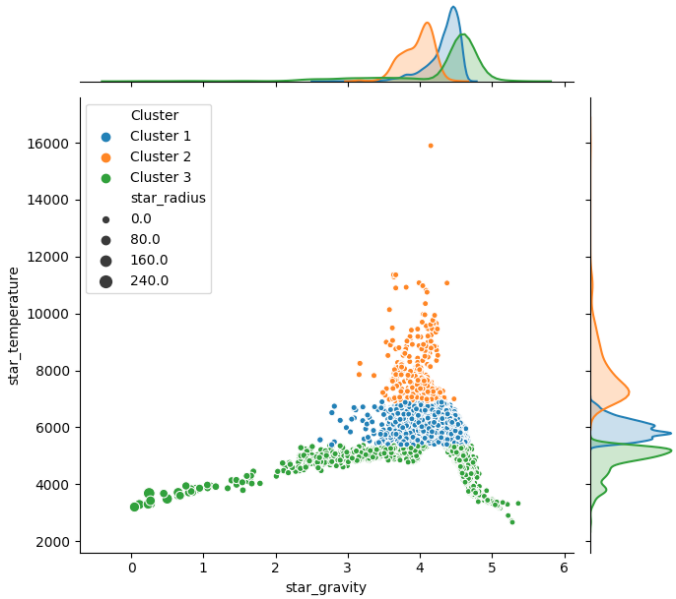
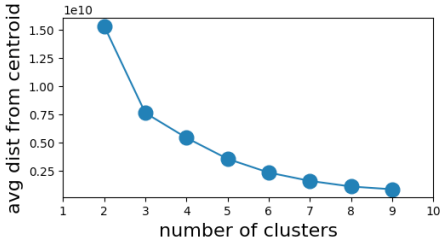
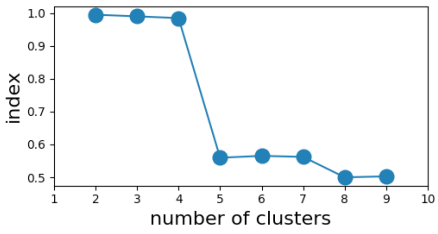


Fig. 10. Joint plot of a scatter and histogram of the stellar parameters subdivided in clusters



(a) Elbow plot



(b) Silhouette coefficient

Fig. 11. Silhouette coefficient and elbow method plot at varying k for planetary clustering, k=2 was selected from the silhouette peak, not in the range of the elbow, but empirically gives the most cohesive clusters

Another point to pay attention to is the distribution of the classes according to the clusters, which is shown in Fig. 12.

It's curious to note how in cluster 2 (whose centroid seems to be referring to class A stars) there are a very low probabilities of finding confirmed exoplanets orbiting around the clustered stars.

Fig. 10 visually represents the clustering subdivision. The histogram observable on the axes shows a different trend according to the cluster, while the scatter plot itself shows how exoplanets have been distributed according to their stellar features.

5.2 Planetary Clustering

The steps for planetary clustering are analogous to those already shown for the stellar clustering. The attributes used, in this case are those concerning the planet itself, *orbital_period*, *transit_epoch*, *transit_duration*, *impact_parameter*, *planet_radius*, *planet_temperature*.

A preliminary preprocessing applied in this step only was the removal of outliers: K-Means is particularily sensitive to them, and a first attempt only generated inconclusive clusters. The optimal number of clusters found with the silhouette method this time was 2.

As Table 2a shows, Cluster 1 seems to be defined by larger, hot celestial bodies with shorter orbital periods. This could be referring to the so-called **Hot Jupiters** planets: these are planets similar to Jupiter, but with really short orbital periods. Their close proximity to their host-star does rise their surface temperature, and they're one of the easiest kind of planets to detect. Cluster 2 is composed by mostly smaller planets, with lower temperatures and longer orbital periods.

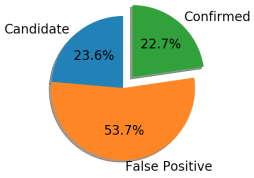
It's notable how celestial bodies belonging to Cluster 1 are, for the most part, false positives, while almost the opposite can be said about Cluster 2.

6. CLASSIFICATION

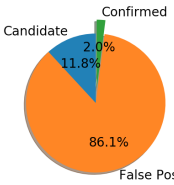
The task consists in the prediction of the class state, representing whether a exoplanets is confirmed or not. For this, different considerations are to be made: the flag attributes of the disposition set are occasional and highly discriminating observations, as shown in 4. As such these observation haven't been considered in the context of classification, as a model being able to predict exoplanets from the main observations seems much more serviceable. The *empirical_state* attribute, being parallel to *state*, also hasn't been considered.

Following are the models used:

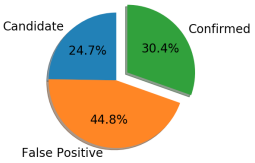
- **J48**, an open source Java implementation of C4.5 (Salzerg). C4.5 is a decision-tree classifier based on ID3, working with the concept of information entropy. The algorithm generates a decision tree where each nodes splits the classes based on the highest normalized information gain. The hyperparameters set up are:
  - *Confidence factor*: 0.25, used for pruning
  - *Minimum number of objects*: 2, minimum number of instances per leaf
  - *Number of folds*: 3, determining the amount of data used to reduce error pruning



(a) State distribution for Cluster 1 (6703 elements)



(b) State distribution for Cluster 2 (346 elements)

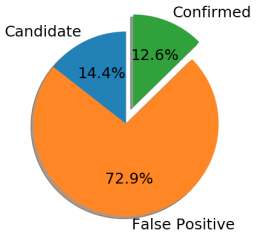


(c) State distribution for Cluster 3 (2514 elements)

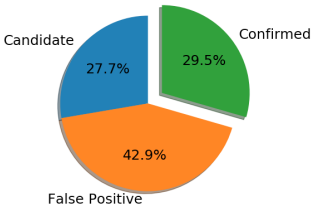
Fig. 12. Class distribution for each (stellar) cluster

Centroids							
	Orbital period	**Transit epoch**	**Transit duration**	**Impact parameter**	**Radius**	**Temperature**	**Dims**
Cluster 1	*5.287*	*138.90*	*4.06*	*1.18*	*328.84*	*1708.22*	2372
Cluster 2	*85.97*	*177.65*	*6.28*	*0.56*	*14.60*	*690.75*	6750

(a) Table 2. Planetary clustering centroids



(a) State distribution for Cluster 1 (2372 elements)



(b) State distribution for Cluster 2 (6750 elements)

Fig. 13. Class distribution for each (planetary) cluster

- **SVM** trained with John Platt's sequential minimal optimization algorithm (Platt). SVM is a supervised machine learning algorithm that uses what's known as kernel trick to perform non-linear classification of the data in high-dimensional feature space. SMO uses heuristics to partition the training problem into smaller problems that can be solved analytically. Typically, it speeds up training by quite a bit. Multi-class problems, as the concerned one, are solved using pairwise classification. The hyperparameters set up are:
  - C: 1, misclassification cost
  - *Kernel*: Polynomial kernel

- **Feed-forward Neural Networks** based on RProp algorithm (Martin Riedmiller).
  - *Maximum number of epochs*: 100
  - *Hidden layers*: 1
  - *Neurons per layers*: 1

The models were validated through the use of 10-fold cross-validation, after which performance measures and error confidence intervals have been measured.

The confusion matrices in Fig. 14 show an irregular response from SVM, as it seems to not be able to classify Candidate exoplanets. This however should not weight to harshly towards the usability of the model: candidate exoplanets are exoplanets waiting to be confirmed as such, and thus are a "buffer class" of currently undecidable exoplanets. Measuring confidence intervals on the errors of each models jointly, as shown in Fig. 15, shows the which model performs with the lowest error, and consequently works better. Given two models, M_1 and M_2, if the upper bound of the interval is less than zero, then M_1 is better than M_2 given the confidence value *alpha*. Otherwise, if the upper bound is greater than zero, and zero is contained in the interval, the difference is not statistically significant given the confidence value *alpha*. If both upper and zero is not contained in the interval, M_2 is better than M_1.

More formally:

$$CI = (\overline{d} - t^{K-1}_{1-\frac{\alpha}{2}} \cdot \overline{\sigma}_d; -\overline{d} + t^{K-1}_{1-\frac{\alpha}{2}} \cdot \overline{\sigma}_d)$$

Where:

- $\overline{d}$ is the mean difference between the errors made on each test fold.
- σ_d is the standard deviation between error values and mean errors value.
- $t^{K-1}_{1-\frac{\alpha}{2}}$ is a value from the T-student distribution equal to *1.645*, so with α equals to *0.05* and $K-1$ equals to ∞ (30-50 degrees of freedom is sufficient for handling the t-distribution as the normal distribution, dataset has more than 9000 measurements).

Being the upper limit lower than zero, it's possible to affirm that the first model (J48) is better than the second (VSM). The comparison between the average errors of the various models after 10-fold cross-validation is therefore also provided.

Following are the accuracy values obtained by each model:

- J48: 0.691
- Neural Network: 0.69
- SVM: 0.56

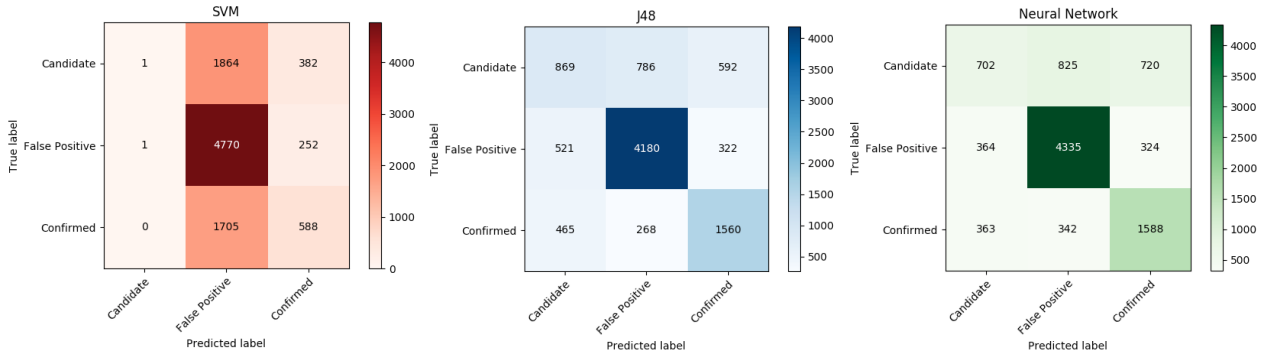
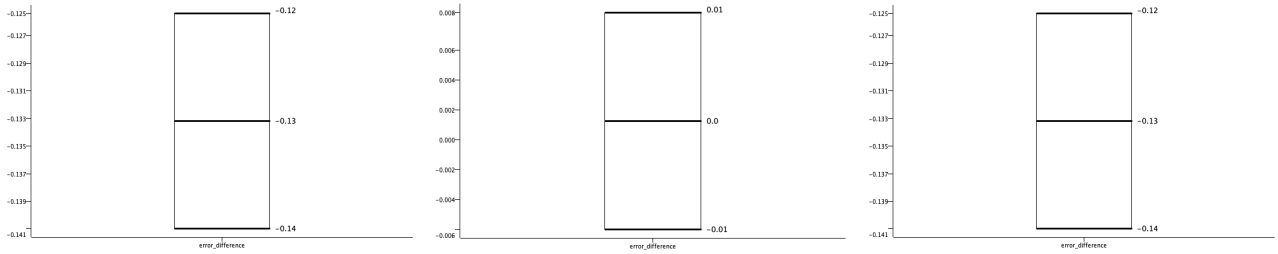


Fig. 14. Confusion matrix of the three models



(a) J48 vs SVM Confidence Interval (b) J48 vs NN Confidence Interval (c) NN vs SVM Confidence Interval

Fig. 15. Paired error confidence interval comparison

Candidate			
	Precision	**Recall**	**F-Measure**
J48	*0.458*	*0.387*	*0.424*
NN	*0.491*	*0.312*	*0.382*
SVM	*0.5*	*0*	*0.001*
False Positive			
	Precision	**Recall**	**F-Measure**
J48	*0.799*	*0.832*	*0.815*
NN	*0.788*	*0.863*	*0.824*
SVM	*0.572*	*0.95*	*0.714*
Confirmed			
	Precision	**Recall**	**F-Measure**
J48	*0.631*	*0.68*	*0.654*
NN	*0.603*	*0.693*	*0.645*
SVM	*0.481*	*0.256*	*0.335*

(a) Table 3. Performance measures on different classes for the three classifiers

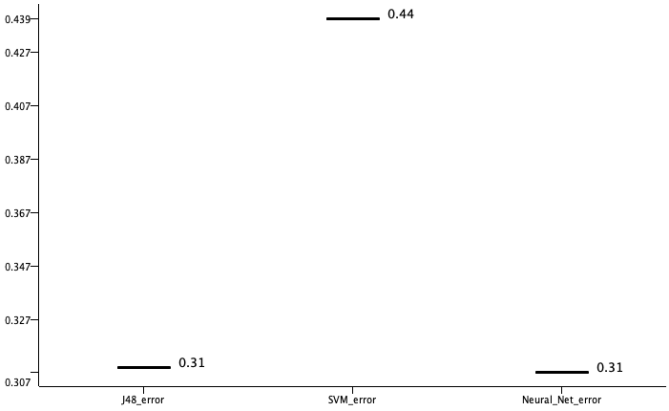


Fig. 16. Models errors values comparison, SVM presents the higher error

The worst performing class between the three models seems to be Candidate. As explained above, this kind of misclassification does not have a real weight by itself. While SVM does not seem able to correctly classify the class, J48 and the NN do find a meager number of examples that do not fit neither the Confirmed and False Positive classes. There seems to be a tendency to classify Candidates more as False Positive than Confirmed, which is actually quite likely, given the distribution of stellar features in 7, showing how Candidates (blue points) appear mostly in the region of False Positive exoplanets (orange points). The Confirmed class achieves good performances for J48 and NN, but does show a slight tendecy to be misclassified as Candidate - in the feature space, some candidates lying on the boundary between the class seem to be misclassified. False Positives show otherwise the best performance of the three classes. It's also the majority class of the dataset, and for the same reasons mentioned beforehand, seems to be quite distinguishable in the feature space than other classes. The other performance measures are self-explanatory, and summarize the considerations made on the confusion matrix. Of note, SVM does present the better recall for the False Positive class, however this seems to be justified by the tendency of the model to label everything as False Positive, as showed by its low precision values.

7. CONCLUSIONS

The analysis conducted raised some interesting elements of work. Future developments may be:

- **Domain expert** appraisal of the data analysis and clustering results, which may confirm the validity of some empirical findings and suggest different ways the data could be analyzed

- **Comparison** of exoplanets found by Kepler with the Transit Method to other exoplanets found by other discovery methods and present in other datasets.
- **In-depth analysis** of each exoplanet candidate using additional raw data from NASA [2009a]
- **Compare other models** using the standardized pipeline developed to compare the performances of different approaches

The context chosen is certainly not one of the simplest but it is one of the most interesting. Approaching the subject has made necessary a research on astronomy and exoplanetology to fully understand the scope of the data collected, and how each attribute related to the classification of exoplanets. There is still a long way to go regarding space exploration.

REFERENCES

Thomas Barclay, Jason F. Rowe, Jack J. Lissauer, Daniel Huber, François Fressin, Steve B. Howell, Stephen T. Bryson, William J. Chaplin, Jean-Michel Désert, Eric D. Lopez, and et al. A sub-mercury-sized exoplanet. *Nature*, 494(7438):452–454, Feb 2013. ISSN 1476-4687. doi: 10.1038/nature11914. URL `http://dx.doi.org/10.1038/nature11914`.

Natalie M. Batalha, Jason F. Rowe, Stephen T. Bryson, Thomas Barclay, Christopher J. Burke, Douglas A. Caldwell, Jessie L. Christiansen, Fergal Mullally, Susan E. Thompson, Timothy M. Brown, and et al. Planetary candidates observed bykepler. iii. analysis of the first 16 months of data. *The Astrophysical Journal Supplement Series*, 204(2):24, Feb 2013. ISSN 1538-4365. doi: 10.1088/0067-0049/204/2/24. URL `http://dx.doi.org/10.1088/0067-0049/204/2/24`.

William J. Borucki, David G. Koch, Gibor Basri, Natalie Batalha, Timothy M. Brown, Stephen T. Bryson, Douglas Caldwell, Jørgen Christensen-Dalsgaard, William D. Cochran, Edna DeVore, and et al. Characteristics of planetary candidates observed bykepler. ii. analysis of the first four months of data. *The Astrophysical Journal*, 736(1):19, Jun 2011. ISSN 1538-4357. doi: 10.1088/0004-637x/736/1/19. URL `http://dx.doi.org/10.1088/0004-637X/736/1/19`.

Jeffrey L. Coughlin, F. Mullally, Susan E. Thompson, Jason F. Rowe, Christopher J. Burke, David W. Latham, Natalie M. Batalha, Aviv Ofir, Billy L. Quarles, Christopher E. Henze, and et al. Planetary candidates observed bykepler. vii. the first fully uniform catalog based on the entire 48-month data set (q1–q17 dr24). *The Astrophysical Journal Supplement Series*, 224(1):12, May 2016. ISSN 1538-4365. doi: 10.3847/0067-0049/224/1/12. URL `http://dx.doi.org/10.3847/0067-0049/224/1/12`.

Google Brain Team. Open sourcing the hunt for exoplanets, 2018. URL `https://ai.googleblog.com/2018/03/open-sourcing-hunt-for-exoplanets.html`.

Heinrich Braun Martin Riedmiller. A direct adaptive method for faster backpropagation learning: The rprop algorithm, 1993. URL `http://www.neuro.nigmatec.ru/materials/themeid_17/riedmiller93direct.pdf`.

Sean D. McCauliff, Jon M. Jenkins, Joseph Catanzarite, Christopher J. Burke, Jeffrey L. Coughlin, Joseph D. Twicken, Peter Tenenbaum, Shawn Seader, Jie Li, and Miles Cote. Automatic classification ofkepler-planetary transit candidates. *The Astrophysical Journal*, 806(1):6, Jun 2015. ISSN 1538-4357. doi: 10.1088/0004-637x/806/1/6. URL `http://dx.doi.org/10.1088/0004-637X/806/1/6`.

NASA. Nasa exoplanet archive, 2009a. URL `https://exoplanetarchive.ipac.caltech.edu/docs/data.html`.

NASA. National aeronautics and space administration, the kepler mission star field. 2009b. URL `https://www.nasa.gov/pdf/189566main_Kepler_Mission.pdf`.

NASA. Kepler exoplanet search result, 2017. URL `https://www.kaggle.com/nasa/kepler-exoplanet-search-results`.

Kyle A. Pearson, Leon Palafox, and Caitlin A. Griffith. Searching for exoplanets using artificial intelligence. *Monthly Notices of the Royal Astronomical Society*, 474 (1):478–491, Oct 2017. ISSN 1365-2966. doi: 10.1093/mnras/stx2761. URL `http://dx.doi.org/10.1093/mnras/stx2761`.

John C. Platt. A fast algorithm for training support vector machines, 1998. URL `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf`.

Steven L. Salzerg. C4.5: Programs for machine learning, 1993. URL `https://link.springer.com/content/pdf/10.1007%2FBF00993309.pdf`.

Johanna Teske, David Ciardi, Steve Howell, Lea Hirsch, and Rachel Johnson. The effects of stellar companions on exoplanet radius distributions. 04 2018.