CS 7641 Team Project

Team 24

Georgia Institute of Technology

Exoplanet Classification/Clustering

Contents

[Background 2](#_Toc52014567)

[Problem Definition 2](#_Toc52014568)

[Methods 2](#_Toc52014569)

[Potential results 3](#_Toc52014570)

[Discussion 3](#_Toc52014571)

[References 3](#_Toc52014572)

# Background

NASA built and launched the Kepler Space Observatory satellite in 2009. The primary objective of the telescope was to find other habitable planets besides our own. As of May 2016, Kepler verified nearly 1300 new exoplanets, and is continuing to collect data on more. NASA published the data that Kepler collected on Kaggle, where we are sourcing this dataset.

# Problem Definition

Using the data that Kepler has collected, NASA is trying to classify exoplanets as candidates for habitation, or as false positives. As there are a nearly infinite number of “objects of interest” in the galaxy let alone, the universe, the amount of time to classify these objects of interest as probable habitable planets versus non-habitable would be too costly and time-intensive for humans to do. Therefore, we are creating a machine learning algorithm to predict/classify whether an object of interest detected by Kepler is a candidate for life or not, to reduce labor costs and human effort.

# Methods

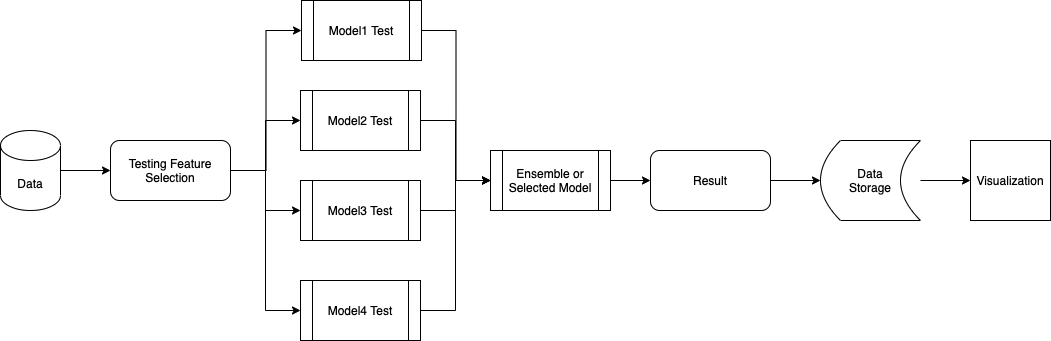
The diagram below visualizes the project design flow we intend to employ. For this problem, our team will be testing both unsupervised and classification models in order to determine what ensemble and features are best fit.

**Unsupervised Approach:**

Our proposed unsupervised approach is clustering. Although the data is already labeled, new data will not have a label; we will cluster on the other attributes of the exoplanet to group planets based on their physical characteristics. For this approach, we can either use a hard assignment algorithm like k-means, or a soft assignment algorithm like Gaussian mixture models.

**Supervised Approach:**

The goal of the supervised learning in this project will be to classify the exoplanets after clustering them based on their physical characteristics. The two approaches we are considering are decision trees and Naïve Bayes. We could create our decision tree based on the characteristics that were most useful for clustering, or use Naïve Bayes with an assumption that are features are independent of one another.



# Potential Results

From the results of our model tests, we will need to determine which features and models to use in our overall project design. Based on the classification research we have identified [4], we believe that the using various clustering algorithms will aid in structuring our data and identifying defining features. We will then be able to use these identified features in classification. This idea of an unsupervised approach aiding a supervised approach has been applied in other contexts, including text classification [4].

# Discussion

Classifying exoplanets is not a completely new problem, and a few papers have addressed it [2, 3]. Our project differs from these methods because of our dataset and our approach. Combining an unsupervised and supervised approach on this specific dataset has not been conducted.

# References

[1] <https://www.kaggle.com/nasa/kepler-exoplanet-search-results>

[2] https://arxiv.org/pdf/1708.00605.pdf

[3] <https://arxiv.org/pdf/1805.08810.pdf>

[4] https://dl.acm.org/doi/pdf/10.1145/1277741.1277918?casa\_token=18GxWvWRax0AAAAA:ZOomsYWlbmgK-lXDCUGzJYoAZgcytAhi\_4BvW2WYSnTpcmpA1CAK7XzLfgVAuyb\_0Q0KTfoFbbw