Practical Machine Learning Project

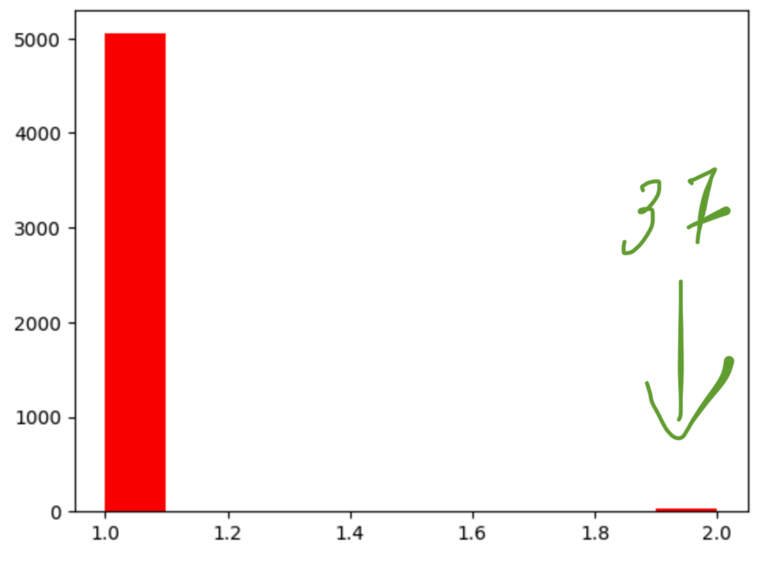
1. Introduction

This project aims to classify the potential star systems that have at least 1 exoplanet in it's orbit (exoplanet is a planet that is outside of our solar system). Why is this usefull? In a distant future, we will need a way to classify stars, meaning if some stars have planets orbiting them or not in hopes of finding a suitable planet to sustain human life.

The dataset is composed of aproximatly 5000 observations and 3200 features. Each features represents the intensity of the star a a specific point in time. In this dataset there are 5050 confirmed non-exoplanet star systems and 37 confirmed exoplanet systems. The set is taken from kaggle.com, but the entirety of the data is colected by NASA using the Kepler Space Telescope.

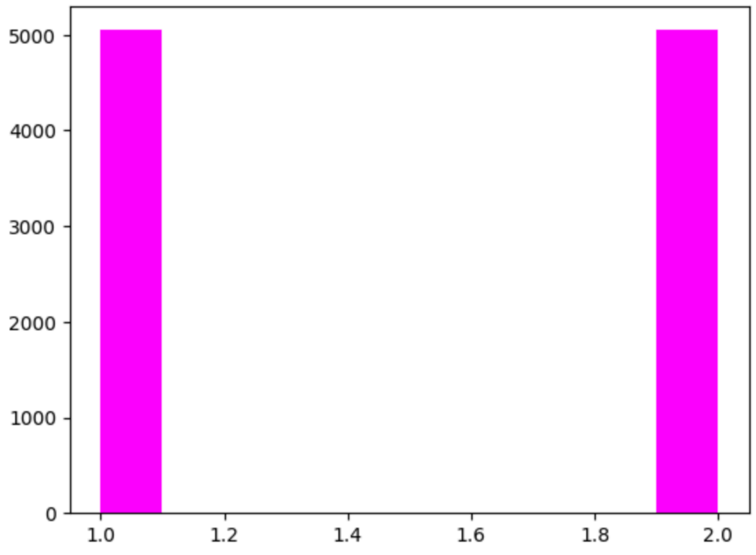
2. Exploratory Data Analysis

This dataset is fortunate enough to not have any missing values or duplicates. Unfortunately there is a big inbalance regarding the classes.



Initial Distribution of classes

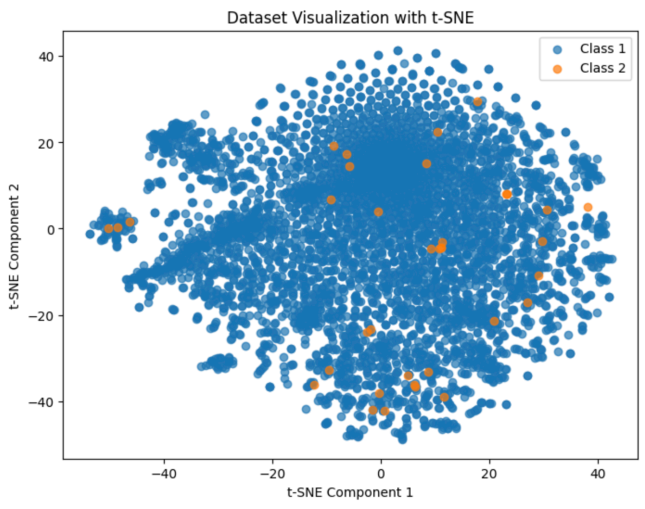
To address this issue, a resampling algorithm was used from the library called

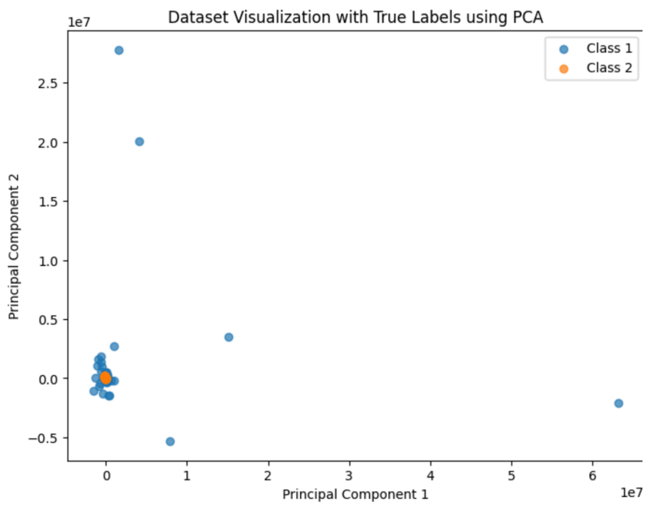


Distribution post resampling

SMOTE and in the end the plot was balanced with 5050 non-exoplanet star systems and 5050 exoplanet star systems.

Using PCA and t-SNE on the bare set, these plots were observed:





3. Preprocessing

In the preprocessing state, some data scalling was performed using MinMaxScaler and StandardScaler. Why both? Well because the self oriented maps(SOM) are working better with the features scaled between [0,1] and also because MinMaxScaler doesn't infer a normal distribution. For Mean-Shift was used the StandardScaler as the Mean-Shift algorithm assumes the distribution to be a normal one but also because it can improve numerical stability in iterative algorithms, especially when dealing with more features then usually.

For extracting meaningfull features in an unsupervised manner IncrementalPCA was used because of the sheer amount of data the set had. I took the time to experiment with the optimal number of components and the batch size leaving with 100 components for SOM and a batch size of 404, for Mean-Shift 35 compoments and a batch size of 100.

In this state, I took the time to extract some meaningfull features to compare IncrementalPCA with them. Those features include: the mean, standard deviation, the skew of the distribution and the kurtosis.

4. Clustering Algorithms Used

For clustering it was used the Self-Organizing-Maps (SOM) and Mean-Shift.

How do Self-Organizing-Maps work? Their fundamental way of working is by mapping high-dimentional data onto a 2D space while preserving the topological structure of the data.

The key parameters for this algorithm are:

1. Grid-size - determins how the points are distributed into clusters.
2. Learning rate - controls the speed of adaptability .
3. Neighborhood function - what function is responsible in the influence of neighboring nodes in the adjustment process.

How does Mean-Shift works? By ajusting the nodes iteratively and moving them closer towards higher density regions it identifies clusters. It is also a non-paramertric clustering algorithm which means it does not need to have the number of clusters specified in advance.

The key parameters for this algorithm are:

1. Bandwidth - it's the radius for the kernel function
2. Kernel function - the function used to determine/estimate the density

5. Hyperparameter tuning

Tunning was done using for loops iterating and scoring the algorithm:

For SOM:

1. grid_sizes = [5, 10, 15]
2. learning_rates = [0.01, 0.05, 0.25]
3. neighborhood_function = 'bubble' and in previos iterations, the 'gaussian' was used

Best results were achieved using this combination of parameters: grid_size 5, learning_rate 0.25, neighborhood_function 'bubble'.

For Mean-Shift:

1. bandwidth = [2.86, 6.5, 8, 320]

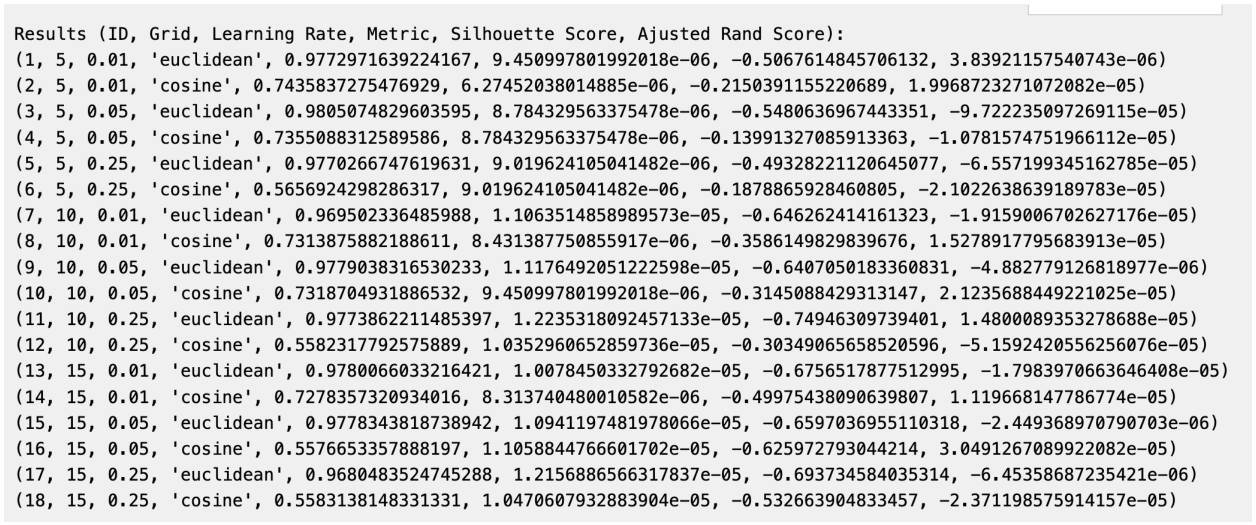
Best results were achieved using the biggest bandwidth meaning 320 for which it makes sens when talking about galaxies.

6. Results and metrics

The results and metrics will be displayed as plots:

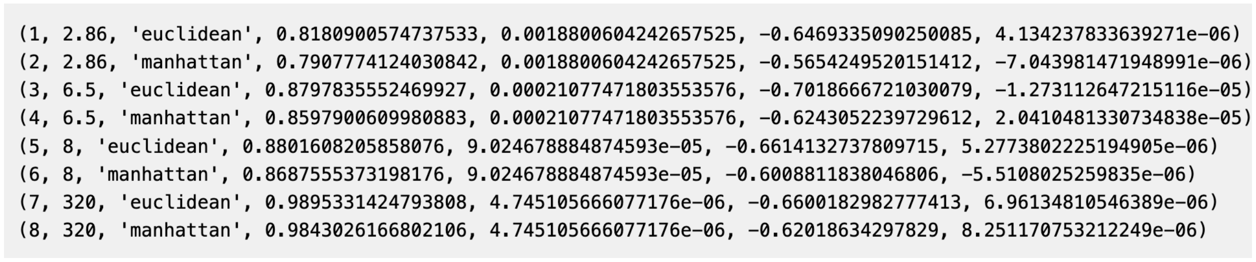
Each combination of parameters with a respectiv metric compared have been ploted based on the id it has, for simplicity i will attach an image with the respective output:

SOM:



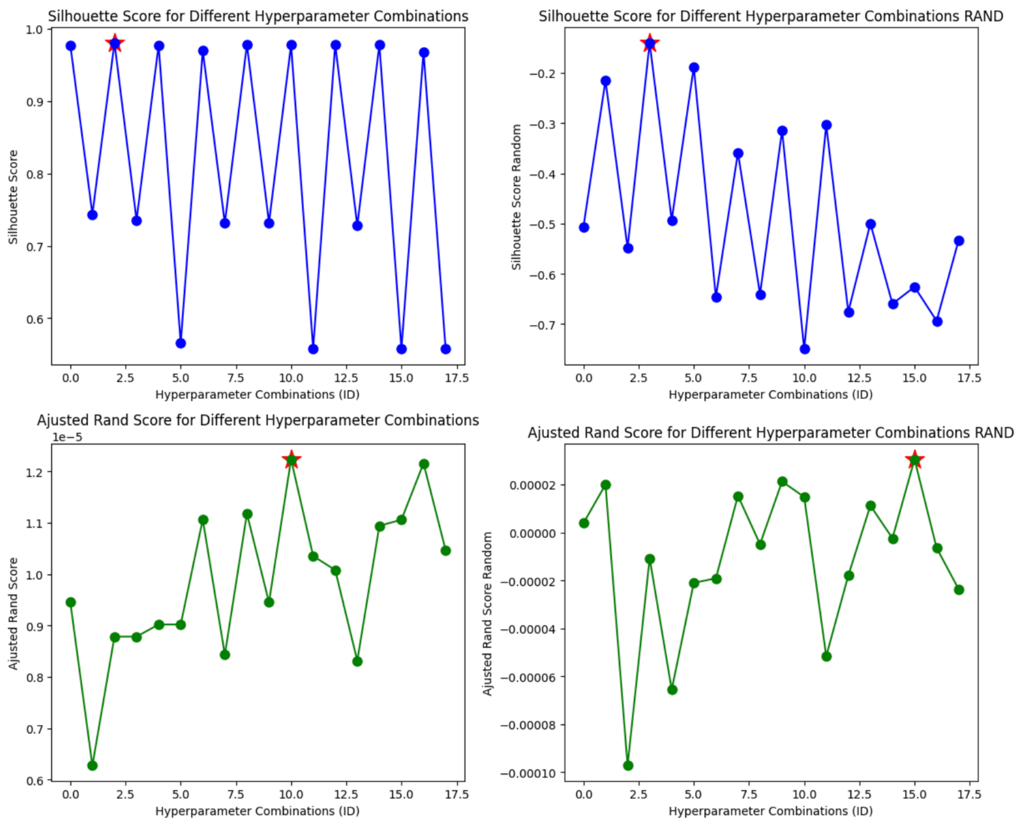
```
Results (ID, Grid, Learning Rate, Metric, Silhouette Score, Ajusted Rand Score):
(1, 5, 0.01, 'euclidean', 0.9772971639224167, 9.450997801992018e-06, -0.5067614845706132, 3.83921157540743e-06)
(2, 5, 0.01, 'cosine', 0.7435837275476929, 6.27452038014885e-06, -0.2150391155220689, 1.9968723271072082e-05)
(3, 5, 0.05, 'euclidean', 0.9805074829603595, 8.784329563375478e-06, -0.5480636967443351, -9.722235097269115e-05)
(4, 5, 0.05, 'cosine', 0.7355088312589586, 8.784329563375478e-06, -0.13991327085913363, -1.0781574751966112e-05)
(5, 5, 0.25, 'euclidean', 0.9770266747619631, 9.019624105041482e-06, -0.49328221120645077, -6.557199345162785e-05)
(6, 5, 0.25, 'cosine', 0.5656924298286317, 9.019624105041482e-06, -0.1878865928460805, -2.1022638639189783e-05)
(7, 10, 0.01, 'euclidean', 0.969502336485988, 1.1063514858989573e-05, -0.646262414161323, -1.9159006702627176e-05)
(8, 10, 0.01, 'cosine', 0.7313875882188611, 8.431387750855917e-06, -0.3586149829839676, 1.5278917795683913e-05)
(9, 10, 0.05, 'euclidean', 0.9779038316530233, 1.1176492051222598e-05, -0.6407050183360831, -4.882779126818977e-06)
(10, 10, 0.05, 'cosine', 0.7318704931886532, 9.450997801992018e-06, -0.3145088429313147, 2.1235688449221025e-05)
(11, 10, 0.25, 'euclidean', 0.9773862211485397, 1.2235318092457133e-05, -0.74946309739401, 1.4800089353278688e-05)
(12, 10, 0.25, 'cosine', 0.5582317792575889, 1.0352960652859736e-05, -0.30349065658520596, -5.1592420556256076e-05)
(13, 15, 0.01, 'euclidean', 0.9780066033216421, 1.0078450332792682e-05, -0.6756517877512995, -1.7983970663646408e-05)
(14, 15, 0.01, 'cosine', 0.7278357320934016, 8.313740480010582e-06, -0.49975438090639807, 1.119668147786774e-05)
(15, 15, 0.05, 'euclidean', 0.9778343818738942, 1.0941197481978066e-05, -0.6597036955110318, -2.449368970790703e-06)
(16, 15, 0.05, 'cosine', 0.5576653357888197, 1.1058844766601702e-05, -0.625972793044214, 3.0491267089922082e-05)
(17, 15, 0.25, 'euclidean', 0.9680483524745288, 1.2156886566317837e-05, -0.693734584035314, -6.45358687235421e-06)
(18, 15, 0.25, 'cosine', 0.5583138148331331, 1.0470607932883904e-05, -0.532663904833457, -2.371198575914157e-05)
```

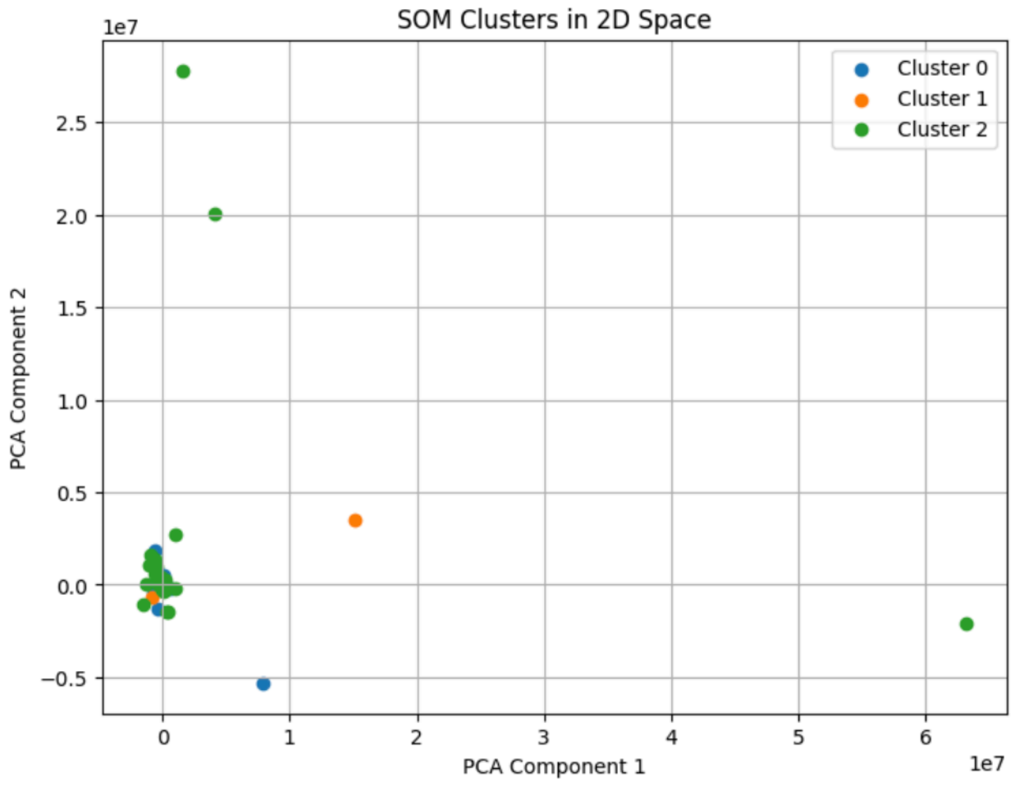
Mean-Shift:



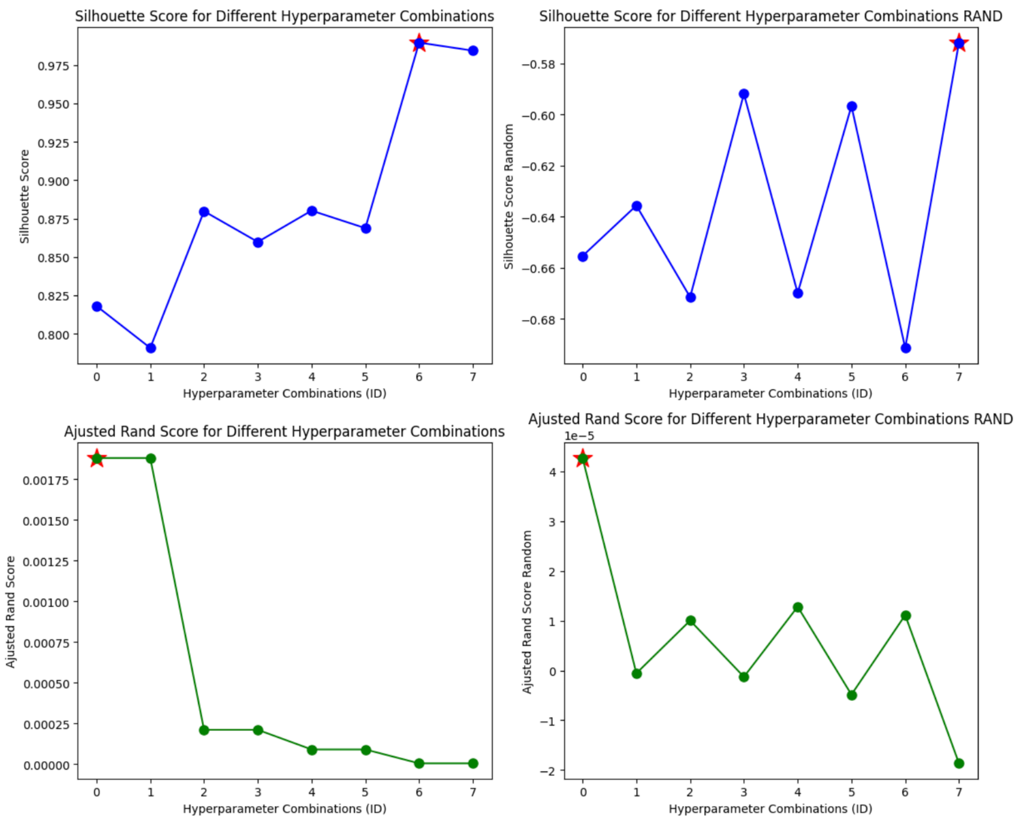
```
(1, 2.86, 'euclidean', 0.8180900574737533, 0.0018800604242657525, -0.6469335090250085, 4.134237833639271e-06)
(2, 2.86, 'manhattan', 0.7907774124030842, 0.0018800604242657525, -0.5654249520151412, -7.043981471948991e-06)
(3, 6.5, 'euclidean', 0.8797835552469927, 0.00021077471803553576, -0.7018666721030079, -1.273112647215116e-05)
(4, 6.5, 'manhattan', 0.8597900609980883, 0.00021077471803553576, -0.6243052239729612, 2.0410481330734838e-05)
(5, 8, 'euclidean', 0.8801608205858076, 9.024678884874593e-05, -0.6614132737809715, 5.2773802225194905e-06)
(6, 8, 'manhattan', 0.8687555373198176, 9.024678884874593e-05, -0.6008811838046806, -5.510802525983 5e-06)
(7, 320, 'euclidean', 0.9895331424793808, 4.745105666077176e-06, -0.6600182982777413, 6.96134810546389e-06)
(8, 320, 'manhattan', 0.9843026166802106, 4.745105666077176e-06, -0.62018634297829, 8.251170753212249e-06)
```

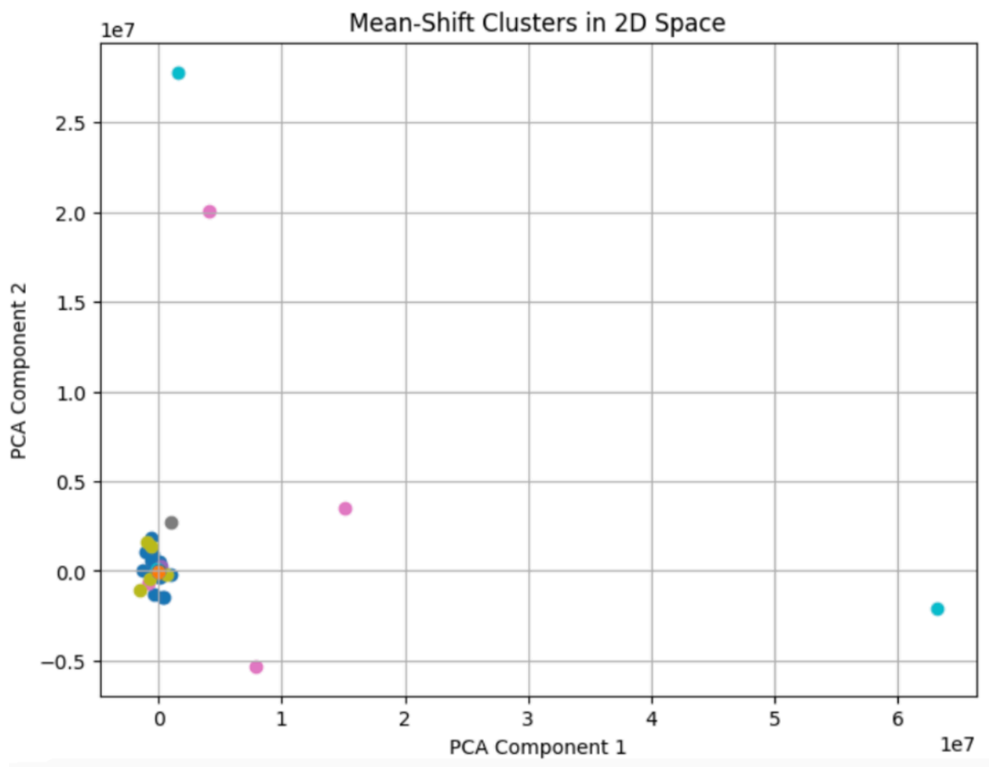
SOM:


Silhouette Score for Different Hyperparameter Combinations
Silhouette Score
Hyperparameter Combinations (ID)
Silhouette Score for Different Hyperparameter Combinations RAND
Silhouette Score Random
Hyperparameter Combinations (ID)
Ajusted Rand Score for Different Hyperparameter Combinations
1e−5
Ajusted Rand Score
Hyperparameter Combinations (ID)
Ajusted Rand Score for Different Hyperparameter Combinations RAND
Ajusted Rand Score Random
Hyperparameter Combinations (ID)


SOM Clusters in 2D Space
1e7
Cluster 0
Cluster 1
Cluster 2
PCA Component 2
PCA Component 1
1e7

Mean-Shift:


Silhouette Score for Different Hyperparameter Combinations
Silhouette Score
Hyperparameter Combinations (ID)
Silhouette Score for Different Hyperparameter Combinations RAND
Silhouette Score Random
Hyperparameter Combinations (ID)
Ajusted Rand Score for Different Hyperparameter Combinations
Ajusted Rand Score
Hyperparameter Combinations (ID)
Ajusted Rand Score for Different Hyperparameter Combinations RAND
1e−5
Ajusted Rand Score Random
Hyperparameter Combinations (ID)


Mean-Shift Clusters in 2D Space
1e7
PCA Component 2
PCA Component 1
1e7