भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data Science(EP4130) Project

Classification of FRBs using Unsupervised ML algorithms

PRATYAKSH RAJ
EP20BTECH11017

Raijade Omkar Sunil
EP20BTECH11020

Contents

1 Abstract

Fast radio bursts (FRBs) are one of the strangest astronomical transients. They are highly energetic Radio signal or Burst of very high energy in radio frequencies. Their origination machenism is not yet very well known. Observationally, they can be classified into repeaters and apparently non-repeaters. However, due to the lack of continuous observations, some apparently repeaters may have been incorrectly recognized as non-repeaters as we saw them only once. We solve this problem of classification using machine learning. We tried to use unsupervised machine learning methods-Kmeans Clustering and HDBSCAN. We will try to make clusters and will then assign the cluster as repeating or non repeating based on the thresholds we define for the ratios of repeating to non repeating.Then the non repeating candidates in repeating clusters can be identified as hidden repeaters. But before applying these Clustering algorithms we will also do the dimensionality reduction using PCA and t-SNE on the data and will also try to find best value of no. of clusters for eg. using silhouttee analysis for KMeans to get k. After both the clusters have been build, we will also provide a list of Strongest repeating candidates (that are initially non repeating but may repeat in future) and also the ranking of most important features for differentiating the clusters.

2 Introduction

Fast radio bursts (FRBs) are a type of millisecond luminous cosmic bursts in the radio wavelength whose origins have yet been brought to light. No model is able to explain all FRB phenomenon. Although most FRBs are a onetime burst but a small fraction like FRB 20121102 shows repeated burst. These repeating FRBs eliminated the possibility of cataclysmic event models being applied to all FRBs. However we can still argue that mechanism of repeating FRB is different and they are not a one time event. Therefore to test a physical model we first must be able to distinguish the characteristics of repeaters from non repeaters. There are some studies on the differences between repeating and non-repeating FRBs, pointing out some difference in parameter distributions of repeating and non-repeating FRBs such as burst width, bandwidth, spectral shape, energy and brightness temperature. These comparasions were done seperately on each parameter. Therefore here we use unsupervised ML algorithms to perform clustering on Data as they are able to cluster similar data together without the output labels. Therefore we can get better insights on characteristic of repeating and non-repeating FRBs.

3 CHIME

The Canadian Hydrogen Intensity Mapping Experiment (CHIME) is a novel transit radio telescope operating across the 400–800 MHz band. CHIME is composed of four 20 m × 100 m semicylindrical paraboloid reflectors, each of which has 256 dual polarization feeds suspended along its axis, giving it a 200 deg2 field of view. This, combined with wide bandwidth, high sensitivity, and a powerful correlator, makes CHIME an excellent instrument for the detection of fast radio bursts (FRBs). It has a total of 536 FRBs observed between July 25, 2018 and July 1, 2019.

4 Catalog

There are 474 non-repeating bursts, and 62 bursts from 18 repeating FRB sources. We treat each sub-burst as an independent burst, and we excluded the burst without flux measurements. This leaves us with 594 individual bursts including sub-bursts, consisting of 500 bursts from the apparently non-repeating category and 94 bursts from the repeating category. The input features to the machine learning algorithms in this study can be classified into two types: primary features that are drawn directly from the CHIME catalog, and secondary features that are calculated from primary features. The distributions of the input features are shown in Fig. 1, with the details discussed below:

4.1 Primary Features

4.1.1 Box-car width (s)

This width represents a rough measure of the duration of the entire burst combining all sub-bursts. The CHIME catalog also provides a more sophisticated fitburst width, which we utilize to calculate rest widths later in this section.

4.1.2 Flux (Jy)

The flux reported by the CHIME catalog is the peak flux averaged across the frequency band. For bursts with sub bursts, the CHIME catalog provides one single flux and we adopt the same value for all of the sub-bursts.

4.1.3 Fluence (Jy ms)

The Fluence reported by the CHIME catalog is the integral of the flux time series across the burst extent and averaged across the frequency band. This value is also the same for sub-bursts in a burst as provided by the CHIME catalog.

4.1.4 Excess dispersion measure (pc cm3)

Excess dispersion measure (DM) is the DM of the FRB excluding the galaxy disk component. We use the NE2001 electron-density model ([**?**])values provided by the CHIME collaboration. This value is the same for sub-bursts.

4.1.5 Peak frequency (MHz)

Peak frequency is the frequency corresponding to the highest flux density. This value is different for different sub-bursts.

4.2 Secondary Features

4.2.1 Redshift

we use the observed DMs to estimate z following the standard procedure. In general, the observed DM of an FRB can be broken down into four components.

DM = DM(MW) + DM(Halo) + DM(IGM) + DM(Host)/(1 + z),

where DMMW is the contribution from the Milky way disk, DMHalo is the contribution from the Milky way halo, DMIGM is the contribution from inter-galactic medium (IGM), and DMHost is the contribution from the FRB host galaxy.DMIGM is directly tied to the redshift of the FRB given by

$$\mathrm{DM}_{\mathrm{IGM}}(z) = \frac{3cH_0\Omega_b f_{\mathrm{IGM}}}{8Gm_p}\int_0^z \frac{\chi(z)(1+z)}{\sqrt{\Omega_m(1+z)^3+\Omega_\Lambda}}\,\mathrm{d}z \tag{4.1}$$

We assume DM(Halo) to be a constant of 30 pc cm^{-3} , DMHost $=$ 70 pc $cm^{^3}$ and f$_{\mathrm{IGM}}$ $= 0.83$([?])

4.2.2 Brightness temperature (K)

For an FRB with peak specific flux S_ν, duration Δt and observed central frequency ν, and redshift z, the full expression of brightness temperature should be (

4.2.3 Rest frame width

The rest-frame width is calculated by correcting the time-dilation effect, i.e.

$$\Delta t_r = \frac{\Delta t}{1+z}, \tag{4.2}$$

where Δt is the observed sub-burst width given by `fitburst`. This value is fitted with an FRB profile model and is different for different sub-bursts in the same burst.

4.3 Luminosity L(erg/s)

$$L = 4\pi D_L^2 \mathcal{S}_{\nu,p}\nu_c \tag{4.3}$$

Where $\mathcal{S}_{\nu,p}$ is the specific peak flux and ν_c is the observed peak frequency of the FRB as above. We take their logarithmic values.

4.4 Rest-frame frequency bandwidth

(MHz)

$$\Delta\nu = (\nu_{\max} - \nu_{\min})(1+z) \tag{4.4}$$

where $\nu_{\max}$ and $\nu_{\min}$ are the highest and lowest observed frequencies of the burst reported in the CHIME catalog in units of MHz.

4.5 Burst energy

$$E = 4D_{\rm L}^2 F\nu_c/(1+z) \tag{4.5}$$

Where F is the fluence of the FRB, and ν_c is peak frequency. Adopting ν_c rather than the bandwidth is more appropriate for wide-spectra FRBs such as the majority of the non-repeating ones. It overestimates the energy of FRBs if the spectra have narrow bands. For consistency, we adopt ν_c for all the bursts since the majority of the bursts are apparently non-repeating ones.

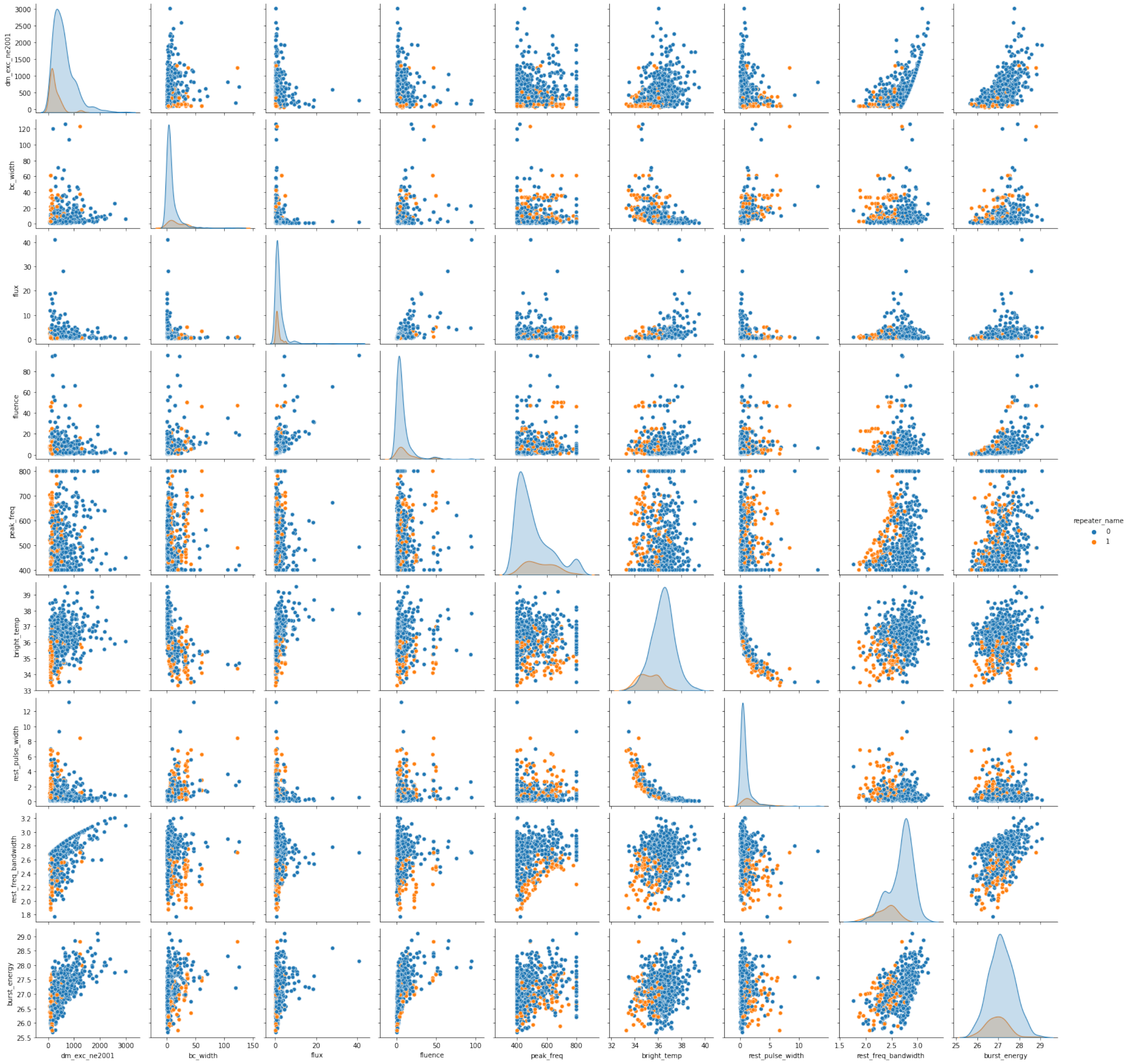


Figure 1. corner plots of all features

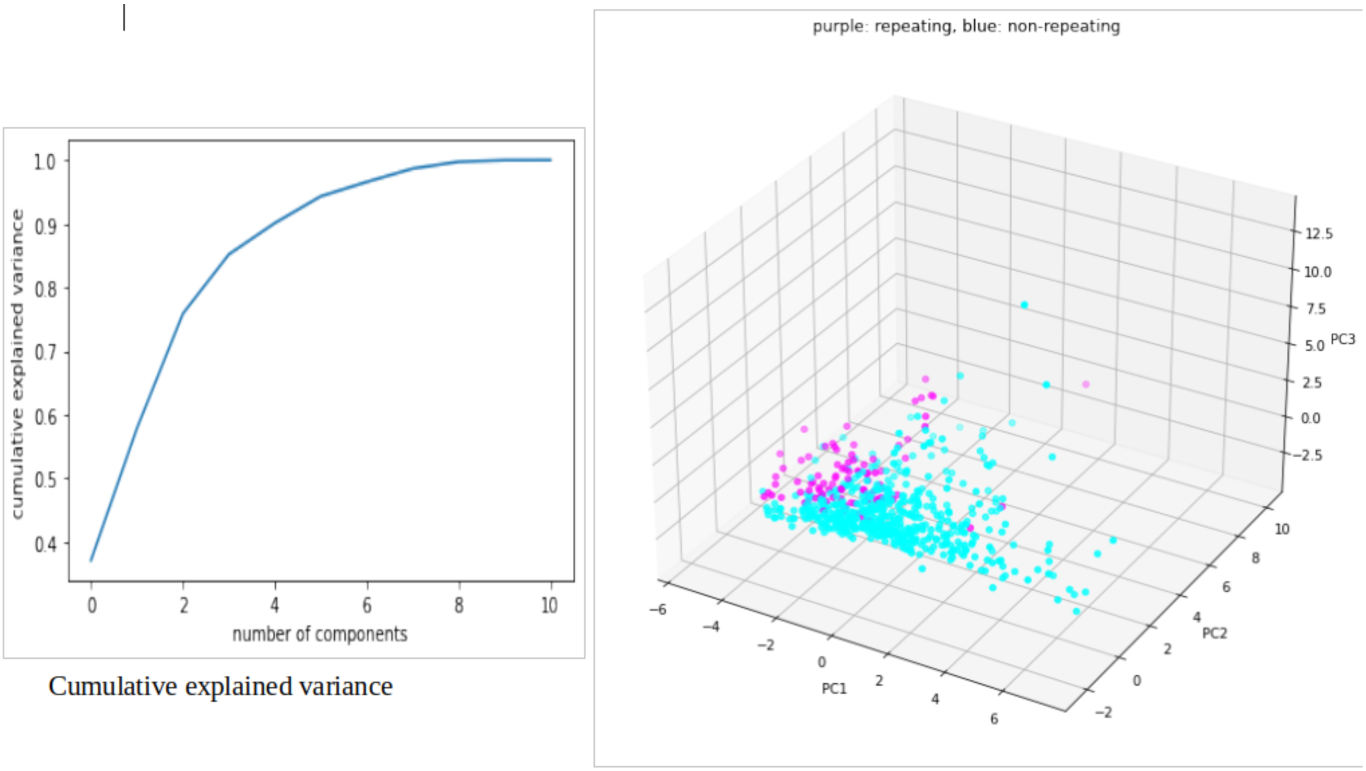
5 Machine Learning Algorithms

Two kinds of unsupervised machine learning methods, clustering and dimensionality reduction, are used in this paper. Dimensionality reduction algorithms learn high-dimensional data sets and automatically transform them into a lowdimensional space. Clustering algorithms group a set of data points into clusters based on their similarities. In practice, high-dimension data are usually first visualized by dimensionality reduction to a lower dimension. Then, clustering methods are used to identify clusters of the data points, after which we can label them manually.
We utilize two kinds of dimensionality reduction algorithms, linear and manifold-based ones which are PCA and t-SNE respectively. Then we will use KMeans on features reduced by PCA and HDBSCAN on features reduced by t-SNE. Reson for this is given in their resp. sections of clustering algos.

6 Dimensionality Reduction

6.1 Principal component analysis (PCA)

Principal Components Analysis (PCA) is a algorithms for linear dimensionality reduction. Given a data set, the PCA algorithm finds the directions (vectors) along which the data has a maximum variance and deduces the relative importance of these directions. Then, the algorithm keeps the most principal vectors as the principal components according to their importance. The number of reserved vectors is decided by the hyperparameter n_components.
Here we used Cumulative explained variance to get best value of n_components. Since PCA is a linear method based on variances, we preprocess the data to standardize features by removing the mean and scaling to unit variance and then putting them into the PCA algorithm. From cumulative explained variance we get no. of principal components= 3. Therefore now we will PCA with n_components = 3. Results are shown in fig2. We can see a clear boundary between Repeating and non repeating FRBs.



Cumulative explained variance

Dimensionality reduction result from PCA

Figure 2. from cumulative explained variance we get no. of principal components= 3 as optimum value and in right fig. data points are plotted in feature space of 3 principal components using PCA

6.2 t-distributed Stochastic Neighbor Embedding (t-SNE)

T-distributed Stochastic Neighbor Embedding (t-SNE) is a manifold dimensionality reduction algorithm. It is based on Stochastic Neighbor Embedding (SNE). Stochastic Neighbor Embedding (SNE) first converts the high-dimensional Euclidean distances into conditional probabilities. But t-SNE modifies the Gaussian distribution used in the probability of SNE to a Student-t distribution. Then, SNE algorithms minimize the sum of Kullback-Leibler divergences by optimizing the cost function with gradient descent.
Here we apply t-SNE to reduce features to two Principal components. plot is shown in fig3

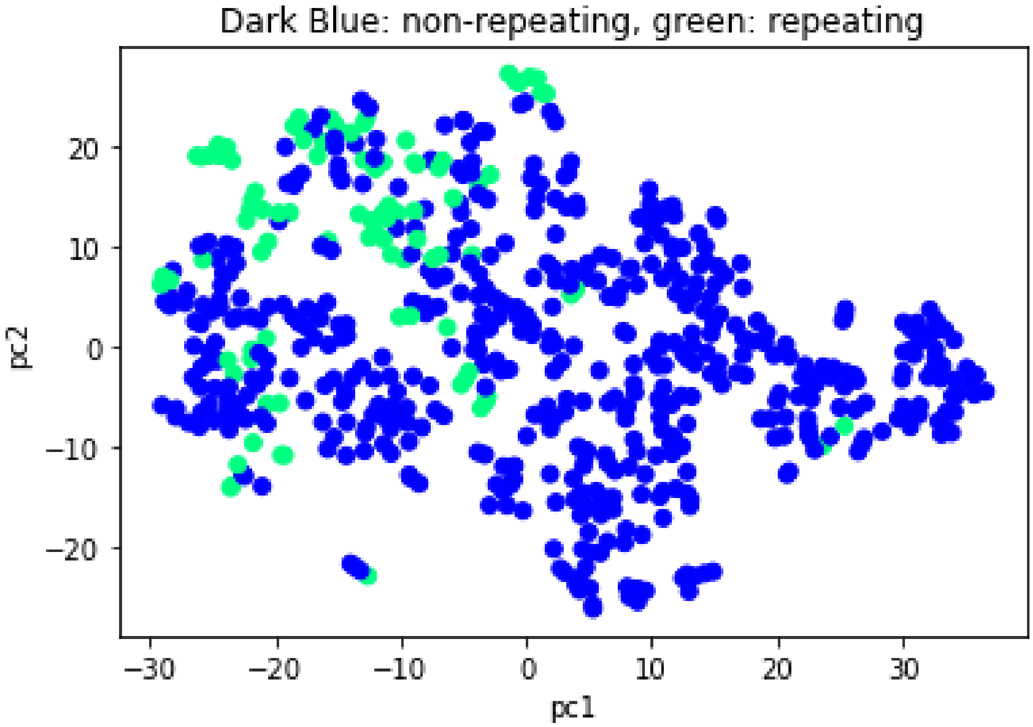


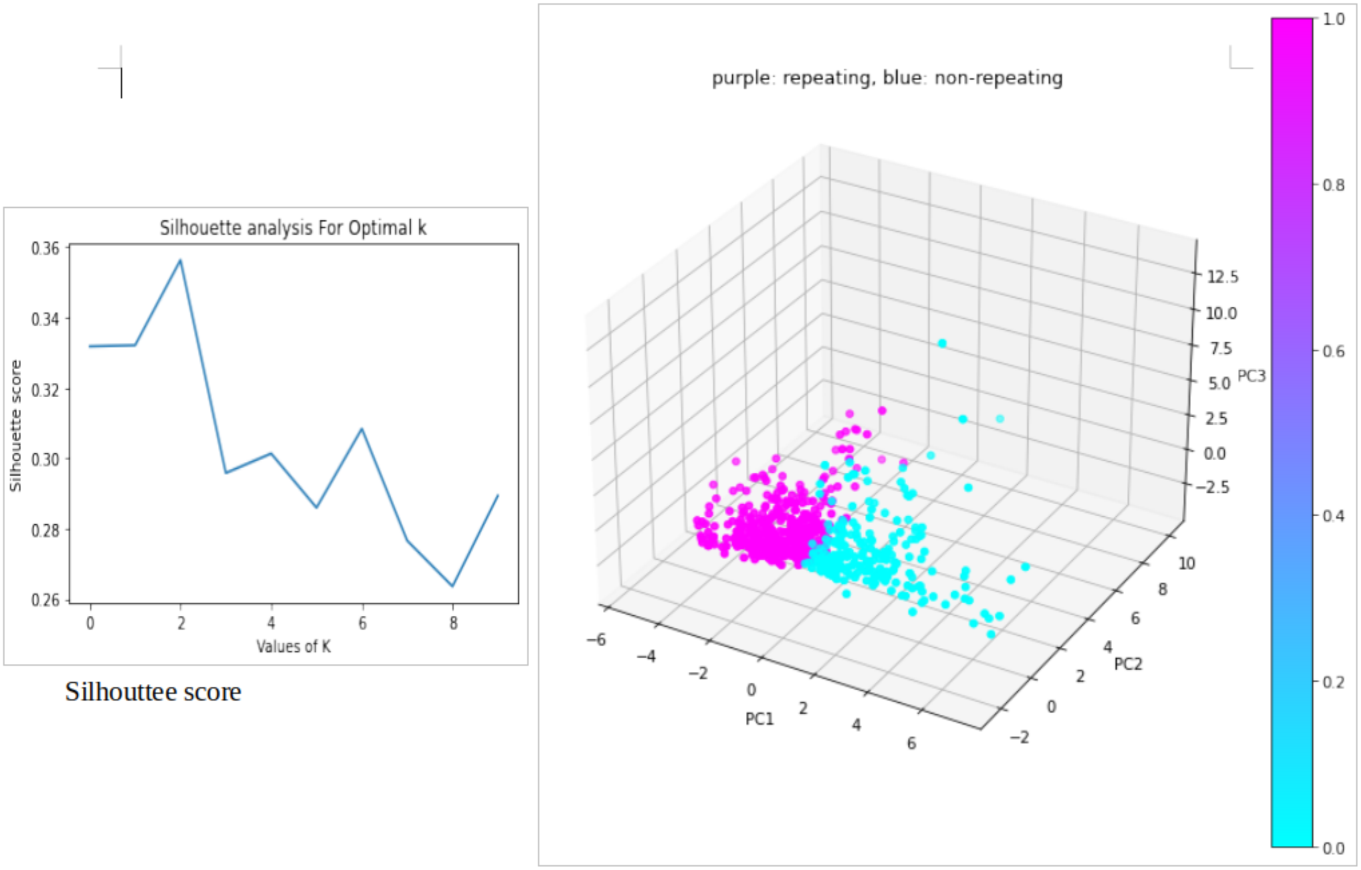
Figure 3. Dimensionality reduction result from t-sne, only two Principal components

7 Clustering

7.1 KMeans

The k-means algorithm groups the data points into clusters by minimizing the sum of squares of Euclidean distances between the geometric points and their centroids. It first initializes k points as cluster centers and then optimizes their positions until they reach the real centers of each cluster. Because k-means is based on the distance from each point to their centers, it performs well in circularlike clusters but fails to identify clusters with strange shapes such as curved shapes. Therefore, we use k-means in dealing with linear-based dimensionality reduction algorithms PCA. A vital hyperparameter in k-means is n_clusters, which refers to how many cluster centers are present in the model. In this paper, we calculate the silhouette coefficient of k-means with different n_clusters to determine the best number of clusters. Silhouette coefficient has long been used to evaluate the quality of clustering. It ranges from -1 to 1, and a higher value stands for more coherent clusters.

we can see in fig4 that silhouttee score is highest for K= 2. Therefore we applied KMeans on PCA space with K=2 in fig4. Purple cluster contains high amount of repeating FRB than in Blue cluster. Therefore purple cluster is assigned as Repeating Cluster and Blue as non repeating. Now the non repeating candidates in purple cluster can be considered as hidden repeaters.



Silhouttee score

KMeans clustering

Figure 4. We can see that silhouette score is highest for k=2. Then right fig. is k-means clustering result in PCA space. The purple cluster, containing a higher ratio of repeaters than in blue, is identified as the repeater cluster, while the blue one is identified as the non-repeater cluster.

7.2 Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)

HDBSCAN is refined from DBSCAN in simplifying the density measurement and combining with hierarchical clustering while scanning the densities. It chooses the clusters based on the hierarchical structure from such steps. HDBSCAN performs well after manifold dimensionality reduction over k-means due to the strange shapes of the groups. Those groups usually are not in circles or follow Gaussian distribution. k-means or other clustering algorithms fail to apply. Therefore, we utilize HDBSCAN after the manifold algorithms t-SNE.

We first input data into the t-SNE algorithms to reduce the dimension and then utilize the HDBSCAN algorithm to label the clusters. The clusters that only include non-repeaters are recognized as "non-repeater clusters"-black and Green. The clusters containing more than a single-digit number (we defined 15% as the criterion) of repeaters are classified as "repeater clusters"-red, yellow and brown. Non-repeaters in repeater clusters can then be treated as hidden repeaters. As for the clusters that only have a few repeaters and the ratio is not enough for 15%, they are labeled as "other clusters"-blue. The single-digit repeaters in these clusters may be miss classified due to the inaccuracies of their input features, or indicate that the clusters are repeater ones. Further observation is required. see fig5

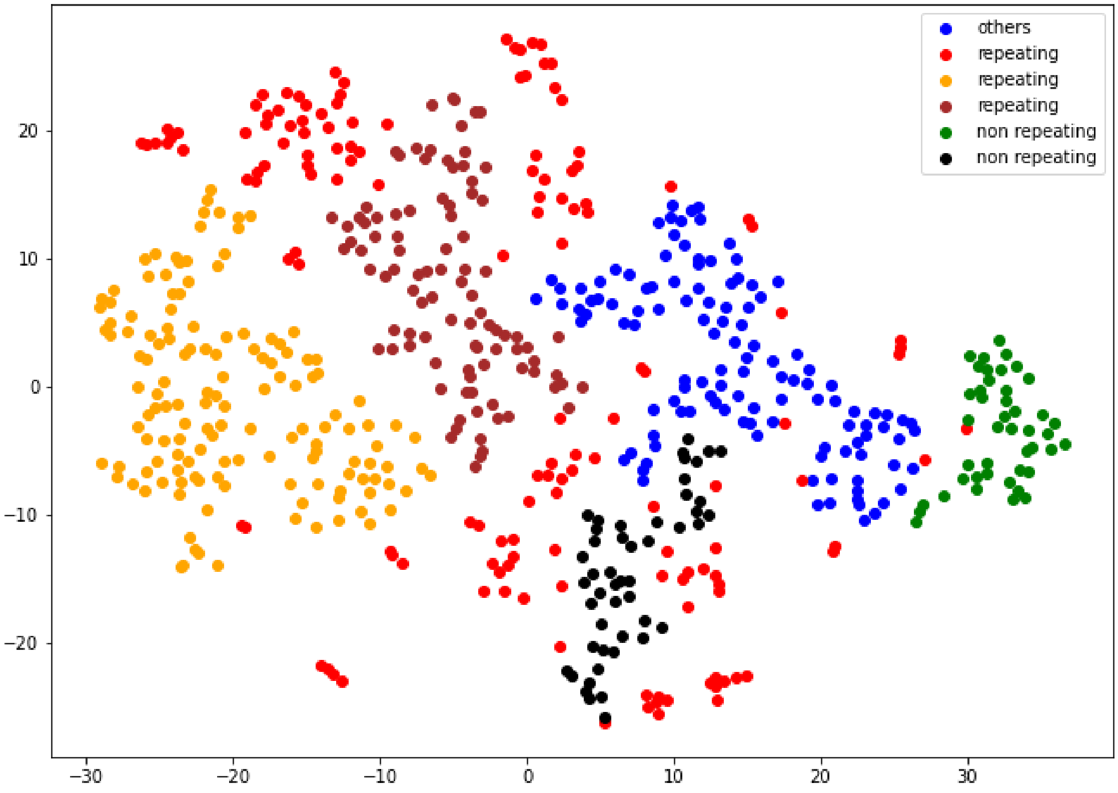
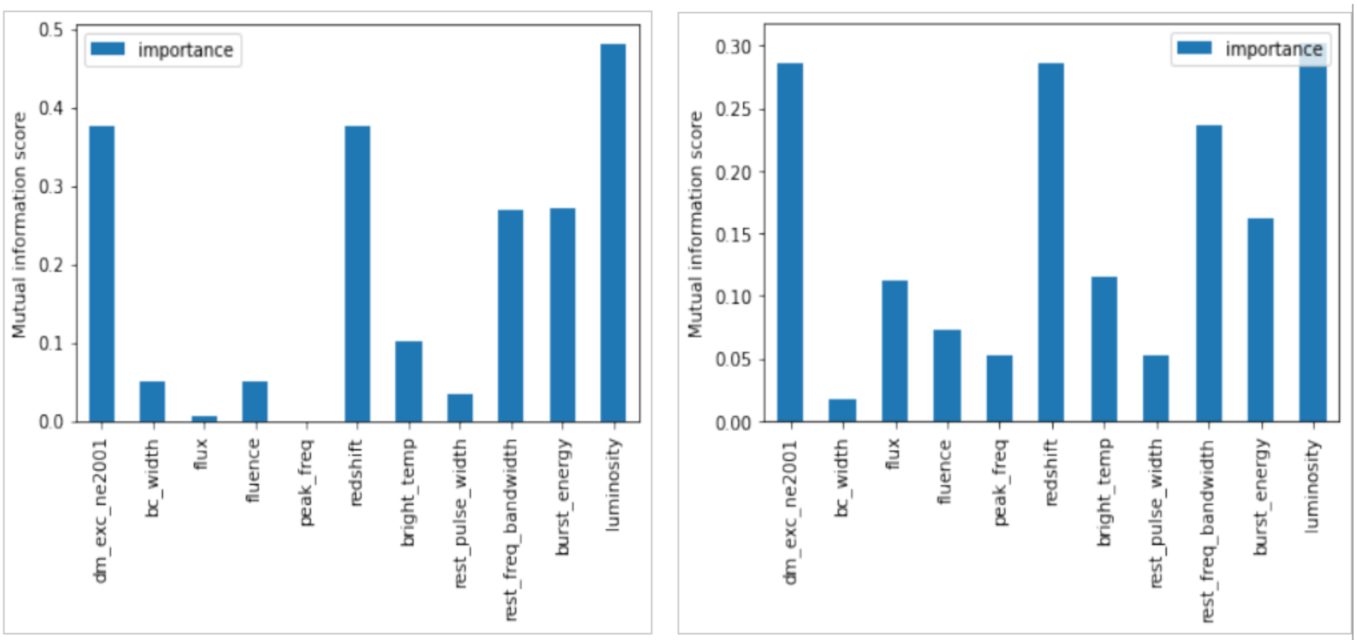


Figure 5. Clustering results of HDBSCAN in the t-SNE plane. here red, yellow, brown clusters are considered as repeating as the contain more than 15% of repeating FRBs and black,Green are taken as non-repeating as they donot contain any repeating FRB and blue is identified as other cluster as it contain less than 15% repeating FRB

8 Feature Ranking: Mutual Information

Now we want to know the correlation between input features and final output.Mutual Information regression in the scikit-learn library is used to estimate the Mutual Information between the features and the coordinates.MI is calculated from the Kullback–Leibler divergence between the joint distribution and the product of the marginal distributions of two variables. Higher MI score implies a higher dependence and importance. Therefore from the plots in fig6 we can see that Luminosity, rest frame frequency bandwidth and redshift/excess dispersion measure are the most important features in differentiating repeaters from non-repeaters.



PCA+KMeans

T-SNE+HDBSCAN

Figure 6. From the Mutual information of Both algos. its is clear that Luminosity, rest frame frequency bandwidth and excess dispersion measure/redshift are the most important features in differentiating a repeating from non repeating FRB

9 Results and Conclusion

Now after doing feature reduction, clustering by both algorithms and getting the ranking of features via mutual information score, we want to get the most probable repeating candidates that may repeat in future. Therefore, we take the intersection of repeating clusters by both algorithms and got 230 candidates that have high probability of repeating in future. Therefore we can conclude from this study that many of the observed FRBs may repeat and continuous observation is required to get more data, with more focus on these 230 repeating candiodates.

10 Link to Code

`https://github.com/PratyakshRaj/DSA_Project_Unsupervised_ML/blob/main/DSA_project_code.ipynb`