CS3DS19 - Data Science Algorithms and Tools

Major Coursework

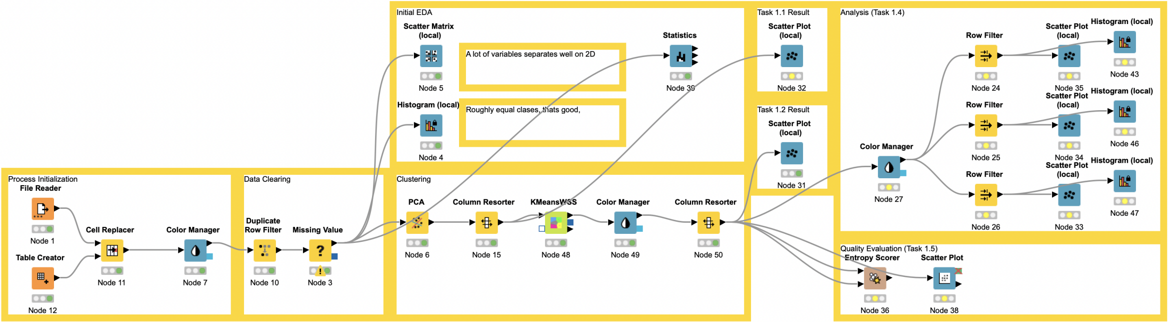
2021/2022

Student ID: 27020363

February 7, 2022

1 Task 1

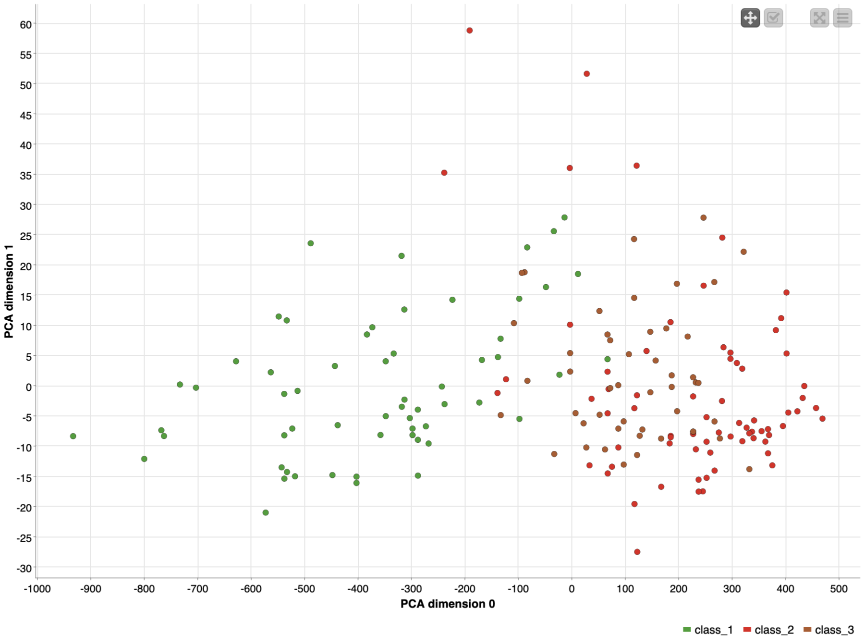
In this Section, the main parts of the workflow will be explained.



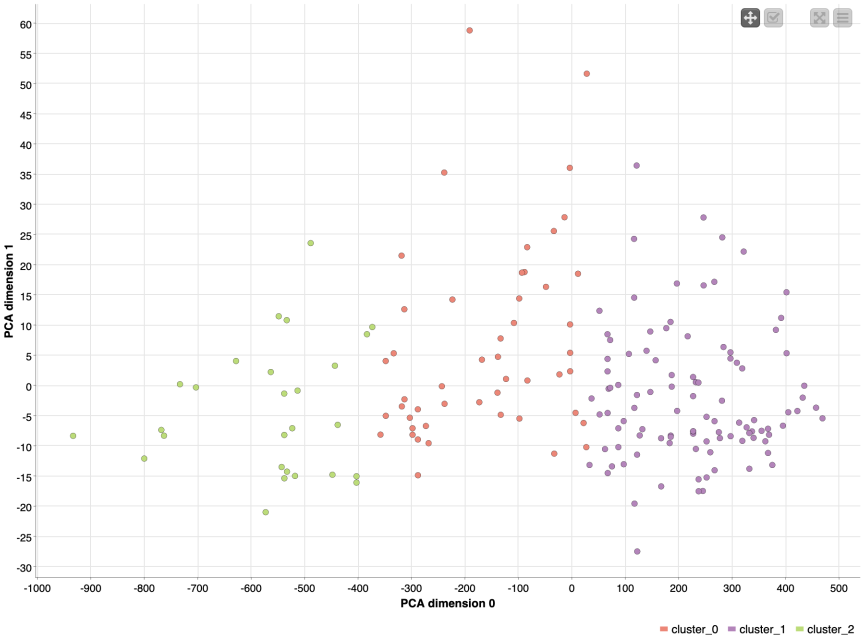
The workflow is splitted into 7 Stages. Those are

- Process Initialization : Read files, and transform the type of the class into string(was number). Finally we assign the initial colours for the subsequent plots, based on the classes field given by the dataset
- Data Clearing : Standard data clearing techniques, Drop duplicate data and missing values(by dropping all the rows involving missing data).
- Initial EDA : Some useful statistics that i used to get myself familliar with the data and their properties
- Clustering : PCA and KMeans Nodes(Futher Explaned below)
- Analysis : Evaluation of error, Classes distribution per cluster and histograms
- Quality Evaluation : Entropy scorer(Entropy/purity)

Task 1.1 : plot1



Task 1.2 : plot2

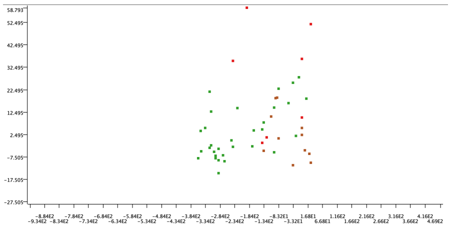


Task 1.3 : compare, discuss and explain the differences between plot1 and plot2

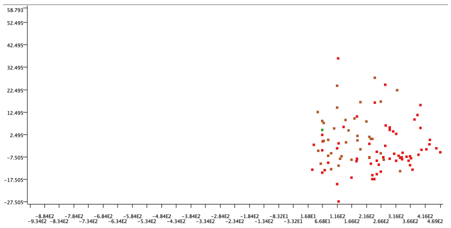
The main point of interest in the figures above, is the fact that, the clustering algorithm(K-Means, k=3) failed to appropriately partition the data, and recognise the classes provided by the dataset. We can see that there are multiple errors(Detailed analysis per-class on Task 1.4). This phenomenon can be explained due to the absence of the standardization process, something that will be explained in the next Task. The exact scale of the problem cannot be appropriately assessed with only those two plots though, we will need to examine the distribution of the classes in each cluster, something that will be done on the next task(Task 1.4)

Task 1.4 : plot3a,plot3b,plot3c

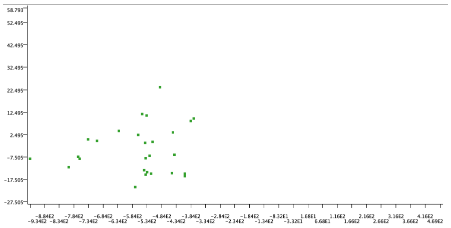
The 3 requested plots are shown below



(a) plot3a

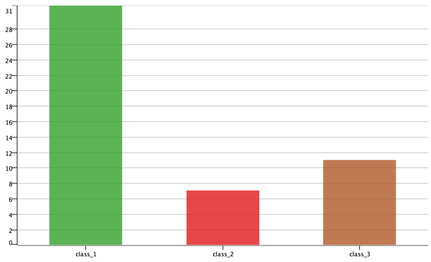


(b) plot3b

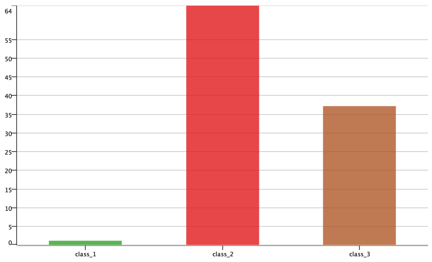


(c) plot3c

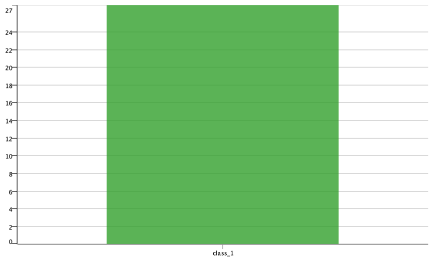
Here, we can see clearly each cluster and the classes of each datapoint that fails under those clusters. With use of histograms, we can learn the distribution of of classes in each cluster. Under ideal conditions(a good clustering), each cluster will contain the majority of the datapoints for some class, with a few missed points, unfortunately, this is not the case



(a) plot3a distribution



(b) plot3b distribution



(c) plot3c distribution

This is the result of the lack of a normalization/standarization process, something that we will explain on the next task.

References

[1] *En.wikipedia.org. 2021. United States Free Speech Exceptions. [online] Available at: ¡https://en.wikipedia.org/wiki/United_States_free_speech_exceptions¿ [Accessed 10 January 2021].*