Analyzing Education Statistics from NYC Public Schools Using K-Means Clustering

***Abstract*—This project is an exercise in working with datasets and gaining insight into them using data mining techniques. The scope of this project includes data visualization, data manipulation, and classification modelling.**

1. INTRODUCTION

PASSNYC is a nonprofit organization which aims to improve education for New York City’s talented, underserved students [1]. The organization takes an analytical approach to its mission, using data from different sources to see where and how New York City schools can improve their education experiences, and achieve measurable results in the process.

PASSNYC currently has a dataset posted on the public dataset repository website, Kaggle. It contains information on education and demographics from over 1200 New York City schools [2]. Such information includes racial makeups, attendance rates, and test registration statistics [2]. The organization wants users of the website to focus on data for SHSAT registration, a test which is used to place students into specialized high schools, which the organization says has been a proven path for successful students.

II. PREVIOUS RESEARCH

There are many ways to tackle this problem, but I have chosen to pursue a k-means clustering algorithm to group together needy schools and classify them as such. A k-means algorithm aims to distribute a set of datapoints into k-clusters [3]. A distance based metric is used to measure similarity of datapoints in a d-dimensional feature space [3]. K randomly chosen datapoints are used as centers for each cluster, and these centers are iteratively moved until they converge to optimal points that correctly divides the data [3].

This algorithm may be useful for this project. We are working with a large dataset, and this method isn’t too costly to use at such a scale. Also, the data comes unlabelled, so traditional classification algorithms, such as decision trees, may not be as useful [4].

III. NEW METHODS

First, the data must be visualized and analyzed to determine which features are the most useful to look at. We may find that some variables have no correlation to SHSAT registration numbers. Then, we may have to manipulate the data in order to normalize it, or discard an egregious outliers that may affect our clustering algorithm.

We must also tailor the algorithm to our specific needs in this problem. We can do this perhaps through carefully choosing our distance metric, or using a variation of the k-means algorithm, like the fuzzy k-means algorithm [4]. A functioning model of this data can only be completed with multiple rounds of testing and updating the algorithm to find an optimal configuration.

REFERENCES

1 PASSNYC. (2019). *PASSNYC*. [online] Available at: http://www.passnyc.org/.

2 Kaggle.com. (2019). *Data Science for Good: PASSNYC*. [online] Available at: https://www.kaggle.com/passnyc/data-science-for-good.

3 Hot, E. and Popovic-Bugarin, V. (2015). Soil data clustering by using K-means and fuzzy K-means algorithm. *2015 23rd Telecommunications Forum Telfor (TELFOR)*, [online] pp.890 - 893. Available at: https://ieeexplore-ieee-org.ezproxy.gsu.edu/document/7377608.

4 Banerjee, S., Choudhary, A. and Pal, S. (2015). Empirical evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means clustering algorithms. *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, [online] pp.168 - 172. Available at: https://ieeexplore-ieee-org.ezproxy.gsu.edu/document/7443889.