Research on Acceptance of Digital Education Platforms after the Epidemic Using Clustering Method

Yuexin Jiang[1*], Shuyu Cai[2+], Adrian Chu[3+], Yanming Chen[4]

[1]*School of Statistics, Jiangxi University of Finance and Economics, Jiangxi Nanchang, 330013, China*
[2]*Orange County, American High School, Guangdong Shenzhen, 518000, China*
[3]*Palo Alto Senior High School, Palo Alto, 94303, The United States*
[4]*The Stony Brook School, 1 Chapman Parkway, Stony Brook, New York, 11790, The United States*
**Corresponding author. Email: Wencifilia@outlook.com*
+These authors contributed equally to this work and should be considered co-second author.

ABSTRACT
In recent years, the use of e-education platforms in various stages of education has increased significantly. The outbreak of COVID-19 in 2020 disturb offline education by forcing numbers of students to stay at home. This essay use integrated methods to explore the integrated development characteristics of digital platforms. Adopting exploratory data analysis (EDA) and K-Prototype clustering methods are conducted comprehensive research on the characteristics of digital education platforms after the epidemic.

Keywords: *digital education, platforms development, exploratory data analysis (EDA), K-Prototype clustering*

1. INTRODUCTION

Schooling is a fundamental part of society's fabric. The interruption to established forms of learning in schools presented challenges to education systems across the globe [1]. In recent years, the use of e-education platforms in various stages of education has increased significantly. Digital platforms provide a good supplement to offline education, expand teaching and Possibility of learning choices. The expansive and ever-fertile frontier of online learning has become the educational learning solution designed to improve knowledge and performance [2].

The outbreak of COVID-19 in 2020 disturb offline education by forcing numbers of students to stay at home. Gives room for online education to develop [2]. Although the outbreak of the epidemic has affected the existing traditional education, it has also caused an "inflection point" in education and promoted idealization [3].

As a medium of online education, digital education platforms have naturally changed in the context of the development of digital education. And affected by different geographical factors, product brands, functions, etc., people's acceptance and preference for digital education products will also have certain differences. In order to better help this educational transformation trend, this difference has attracted the attention of many scholars. Dinu and Rodica Start [2] with universities and analysis the university's potential for continuing in order to evaluate the university potential in offering aspect, and they conclude that the quality of university equipment (such as network, computer equipment, etc.) will have a certain impact on the use of its network teaching platform. While Li and Yao [3] take the quality of education as a major factor and mentioned that the biggest advantage of online education is that it can share truly high-quality teacher resources at low cost and achieve education peace. Jonathan and Carrie [1] applied qualitative study in Scotland for generally trend of digital learning and found out people's responses to the question also indicated a linear progression within their own thinking about technology, this will largely be affected by different product suppliers and educational objects [4]. However, as a new field, few people use integrated methods to explore the integrated development characteristics of digital platforms. Judgment analysis lacks specific judgments, especially subjective analysis after direct investigation, and lacks precise and powerful support. So this work decided to adopt exploratory data analysis (EDA) and clustering methods [5][6] to conduct comprehensive research on the characteristics of digital education platforms after the epidemic.

2. DATA

2.1. Dataset explanation

This paper use the analytics competition "LearnPlatform COVID-19 Impact on Digital Learning" found on the website (1) . The data as a whole consists of three separate data sets.

1)Engagement data

As **Table 1 Engagement data**, engagement data is a time series and have five features in this dataset.

percent access: The number of students in a given school district that had loaded at least 1 page of a product.

engagement index: The total number of pages loaded per thousand students in a given school district directly relates to the product requirement.

lp id: The special code of each singal product.

district id,: The special code of school districts.

Table 1 Engagement data

	time	lp id	pct_access	engagement_data	district id
1	2020-01-01	99792.0	0.02	2.26	1044
2	2020-01-01	80493.0	0.01	0.17	1044
3	2020-01-01	16164.0	0.03	0.35	1044

2)District information data

district information consists seven region attributes described by **Table 2 District information data**.

state: The state of the district.

locale: The locale type of the district.

pct_black/Hispanic: The percentage of black or hispanic students within the district

pct_free/reduced lunch: Percentage of students eligible for free or reduced cost for lunch.

Table 2 District information data

	state	locale	pct black/hispanic	pct free/reduced	pre-pupil total
1	Illionis	Suburb	0.1	0.1	15000.0
2	Utoh	Suburb	0.1	0.3	7000.0
3	Wisconsin	Suburb	0.1	0.1	11000.0

3)Product information data

Table 3 describes the product information, which consists of:

provider/company name: The name of provider and the company.

sectors: The main education target of the product.

primary essential ***function:*** The main function of the product,

The columns of data holding the url and product name were dropped since they were unnecessary.

Table 3 Product information data

	company name	category	sectors
1	StudyPad Inc.	LC	PreK-12
2	Age of learning, Inc	LC	PreK-12
3	ABCya.cim LLC	LC	PreK-12

2.2 Exploratory data analysis(EDA)

EDA offers a better understanding of the dataset. Preliminary analysis of the relationship between variables and data processing helps a more reliable result.

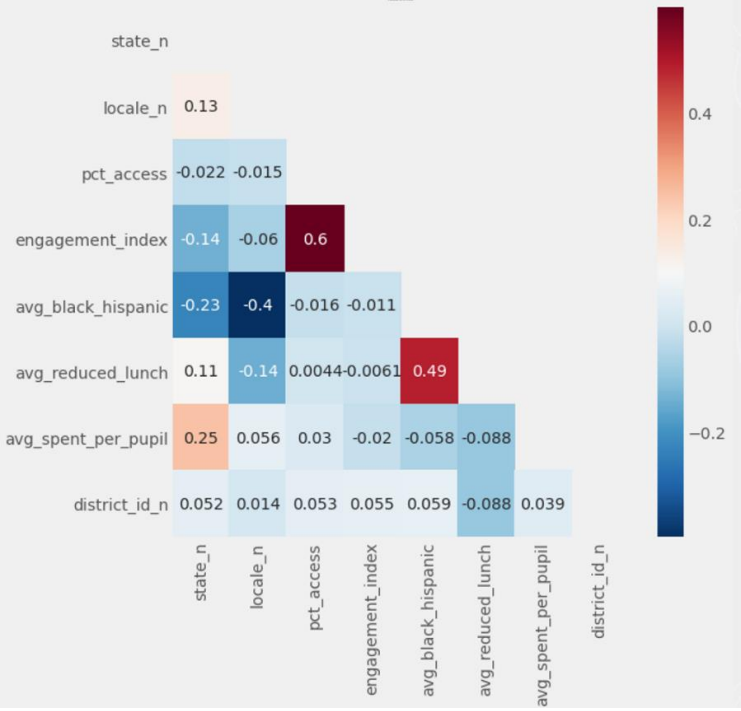

Figure 1: Relational Heat Map of Digital Variables

Heatmap(**Figure 1**) visualise the correlation between each parameter in the data set. A simple label encoding being applied. According to the blocks with deeper colour, the correlation between pct access and engagement index is the highest.

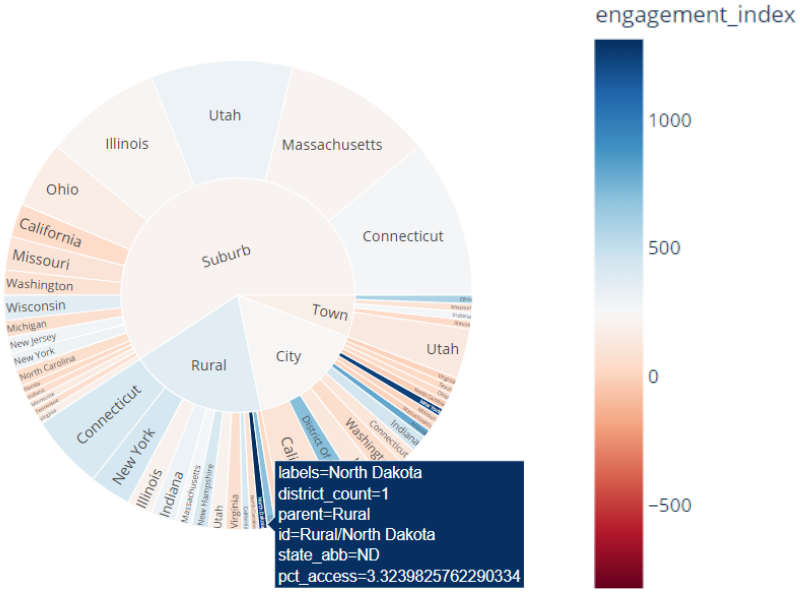

Figure 2: Sun graph of locale and states via engagement_index

Figure 2 shows regional information. The school district in each state were classified into Suburb, city, town and rural. School district in most States belongs to suburb with low engagement index. And the highest engagement and access all happened in city area with a small number of school districts.

3. CLUSTERING

In order to better analyse the distinction between school areas from a higher dimension, this work apply clustering method to handle variables.

3.1. Data pre-processing

The work focus on the general product selection and usage. Date pre-processing includes removing logic

abnormal data and interval transformation by median. **Table 4.** for details.

Table 4 Handling logical error values.

types	target	The range to be removed
Abnormal values	Pct_access(PCT)	PCT=0&ENG>0
	Engagement_index(ENG)	
Interval variables	avg_black_hispanic	(upper limit+lower limit)*1/2
	avg_reduced_lunch	
	avg_spent_per_pupil	

Moreover, this work used 10% in both tails to clean up outliers.

3.2. Methodology

1）Principle Component Analysis

Principle Component Analysis is a lossy compression method, which is widely used for feature dimensionality reduction in data analysis. It helps to extract the main components and improves the model's quality.

2) K-Prototype Algorithm

K-Prototype is a clustering method based on partitioning and proposed by ZHEXUE HUANG in terms of an improvement of K-Means and K-Mode algorithm[4]. It clustering process is use the k-modes approach to updating the categorical attributes.K-Prototype can handle categorical or mixed dataset with highly efficiency.

3.3 Clustering

3.3.1 Prepare for clustering

PCA is applied to the dataset to avoid the dimensional problem, the results shows in **Figure 3** that most of the information can be explained with 3 component.

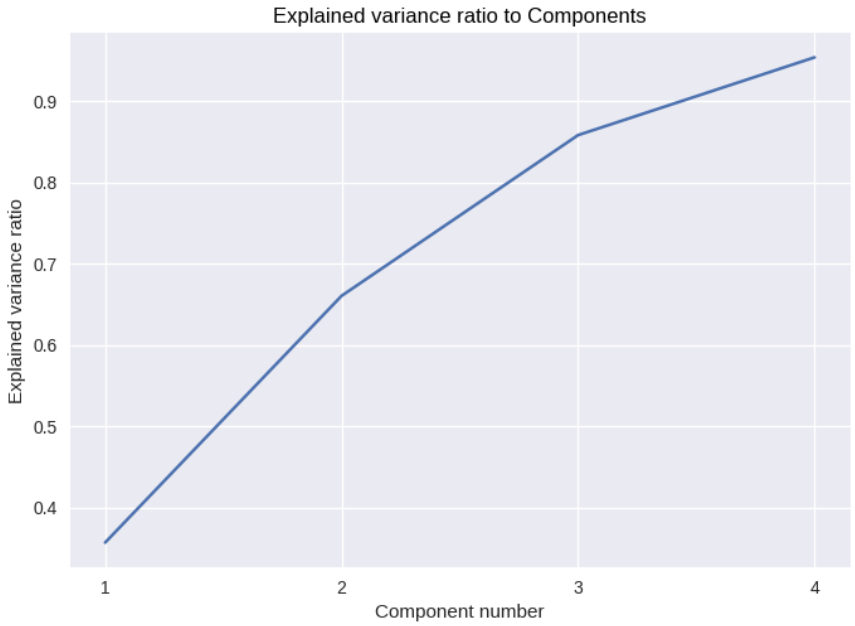

Figure 3 Component selection using Principal Component Analysis(PCA)

The components interpretation as follows:

***The first component*:** Includes the proportion of people's number in each product and the usage rate. Which shows the **acceptance of the platform**.

***The second component*:** The regional race ratio and lunch fee reduction/exemption ratio. Regards of quality **of life and diversity**,

The third component: Indicates **education quality** as tuition related to education quality.

3.3.1 Cluster number selection

Elbow method[5][6] is adopted to select the most appropriate number of clusters, as shown in **Figure 4**.

The curve gradually flattened out after the distortion being greatly reduced in 4 clusters. So 4 clusters has been applied.

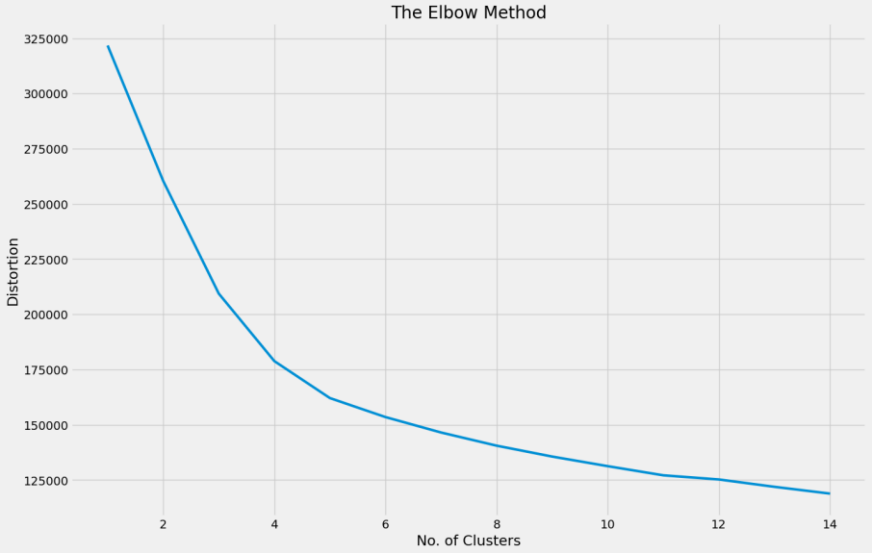

Figure 4 Elbow Line

3.4 Result and cluster interpretation

Figure 5 shows the spatial visualization results of three principal components after clustering. While colors represent different clusters.

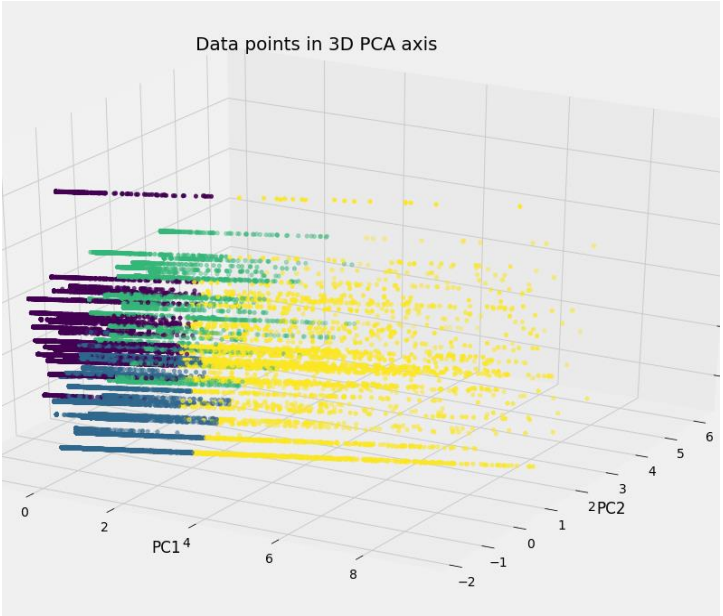

Figure 5 The spatial visualization results

Variables are not well distinguished in the graph. This work select a few representative characteristics from geographical and products aspect to interoperate further.

3.4.1 overall distribution

Figure 6 shows the distribution of each clusters in percentage.

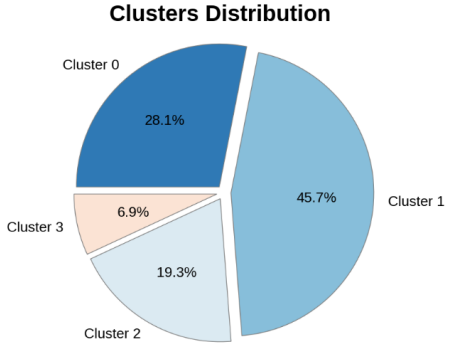

Figure 6 Clusters distribution

Cluster1 occupies the vast majority of the samples, while cluster3 has the smallest. The summarize result as blow:

Cluster0： The school area with high education quality and low welfare benefits and diversity. Digital learning platform is well accepted.

Cluster1： The quality of education is considerably lower, with a medium life welfare and diversity. The online learning platform is not that well accepted in some of the districts in this cluster.

Cluster2： Mainly school area with high education quality, high welfare and diversity, digital learning platform are not that well accepted around these areas.

Cluster3： A school area with good quality of education, high life welfare and diversity, and accept digital learning platform pretty well

4. CONCLUSIONS

Clustering result demonstrated the main characteristics of the acceptance:

♦The situation of the school area will affect the acceptance of regional products even though the education level of these schools is higher.

♦Schools with similar education levels have lower welfare and diversity actually have a higher acceptance.

♦The teaching quality of a school has a great influence on general, and it will be much higher than other schools under the same conditions.

REFERENCES

[1] Brown, J., McLennan, C., Daniela Mercieca, D., (2021) Technology as Thirdspace: Teachers in Scottish Schools Engaging with and Being Challenged by Digital Technology in First COVID-19 Lockdown. education science., 136: 1-16.

[2] Țurcanu, D., Siminiuc, R., Bostan, V., (2020) The Impact of the COVID-19 Pandemic on the Use of Digital Technologies in Ensuring the Efficient e-Learning Process at the Technical University of Moldova. Creative Education., 11:2116-2131

[3] Di Palma, D., Belfiore, P., (2020). Technology and Didactic Innovation in School at the Time of COVID-19: An Evaluation of the Educational Effectiveness in the Student Perspectiv. Council of the European Union.,20:1-8.

[4] Wang, Z., Fan, H., (2021). Spatio temporal cluster analysis and socio-economic influencing factors of COVID-19 epidemic situation in Hubei Province. Engineering Journal of Wuhan University., 54: 868-870.

[5]Zi-qi, J., Ling, S., (2020). K-prototypes Clustering Algorithm for Mixed Data Clustering. Journal of Chinese Computer Systems., 9:1845-1848.

[6] Fu, L., Wu, S., (2019). A new internal clustering validation index for categorical data based on concentration of attribute values. Chinese Journal of Engineering., 41: 682-684.

[7]LearnPlatform. (2020) LearnPlatform COVID-19 Impact on Digital Learning https://www.kaggle.com/c/learnplatform-covid19-impact-on-digital-learning/data?select=products_info.csv