ISSN 2347 - 3983

Volume 8. No. 6, June 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter76862020.pdf

https://doi.org/10.30534/ijeter/2020/76862020

Analyzing Android Users Based on Google Play Store Using K-Prototype Algorithm

Abba Suganda Girsang[1],Melva Hermayanty Saragih[2], Ezra peranginangin[3]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480, agirsang@binus.edu

[2]Management Department, BINUS Business School Undergraduate Program, Bina Nusantara University, Jakarta, Indonesia 11480,melva.saragih@binus.ac.id

[3]Ezra Peranginangin, Product Design Program, Podomoro University, Jakarta 11470, ezra.peranginangin@podomorouniversity.ac.id

ABSTRACT

As a gate applications of android application, google play store also contains the enormous potential data to analyze. With this data, the behavior of android user can be investigated to get some insights. It then can be used to serve the customer by providing appropriate information needed. This study classifies the transactions into four clusters using k-prototypes algorithm. Before grouping, the preprocessing is conducted by filtering, removing the incomplete record data. Moreover, to be more focus, six attributes from thirteen original attributes are eliminated to analyzed. Each clusters are then analyzed to get the some insight information which can be used to the appropriate users.

Key words: Play store, business, k prototypes, classification

1. INTRODUCTION

The Google Play store provides many contents such as games, applications, movies and books which is prepared third-party. In fact many developers upload their content to play store which causes a huge content in play store. Many contents is chosen only a few user. It can happen, because of the lack of robust tools for analyzing. Also, Google, as a provider play store, has not access the source code of contents. However google play store provides the comments which is useful as benchmark for the content[1][2]. The third party can analyze this data for the improvement of their business. In other side, the analysis can be conducted by collecting the information history of users who download the contents[3]. In this study, content only focuses on applications. The history data is grouped by k-prototype algorithm[4]. By grouping the history data, the specific behavior of user can be known. As a result, the specific treatment for each userscan be conducted such as specific promotion, announcement, information can be delivered.

Many methods are used to cluster data using some methods such as k-means[5][6] , ACO [7], ABC [8], Cuckoo Search [9] and so forth.

In this study, the k-prototypes algorithm is used to classify the transactions. The k-prototypes is a mixing algorithm k-means and k-modes algorithms which uses mixed numeric and categorical attributes[10]. This algorithm is chosen because the transaction data contains mixed attributes and both numerical[11]. This study generates some scenario clusters[12]. However, after analyzing, the number of clusters is four.

2. PROPOSED METHOD

This data transaction of google play store comes from Kaggle dataset. This original data consists 10841 records with 13 columns or attributes. The attributes are described as follows. App is the attribute of application name. Category is the attribute of category application such as art design, healthy, vehicle and so forth. Rating is the attribute of accumulation rating of users. Review is the attribute of number of user reviews for the app (as when scraped).Size is the attribute of size of the app (as when scraped). Install is the attribute of number of user downloads/installs for the app. Type is the attribute of paid or free. Price is the attribute of price of the application in dollars, Content rating is the attribute ofage group the application targeted. Genres is the attribute of an application can belong to multiple genres (apart from its main category), for example, a musical family game will belong to. Last update is the attribute of the date of last update application version. Current ver is the attribute of last version. Android ver is is the attribute of an android version.

Some steps of preprocess are conducted such as filtering, replacing, removing, and changing to the numbers. For this dataset, this study only uses the free app category because there is only a small amount of paid app data. While analyzing the data, there are a number of rows found that have blank values, so we filtered to remove rows that have blank values. The filter method is used because there are only 17 blank line out of a total of 10840 data lines. The focus only on analyzing apps that are not paid.

The detailed process can be seen in Table 1 and Table 2. Table 1 shows the original attributes, and Table 2 shows the attribute after preprocessing. After pre-processing, the data that will eventually be used for kmodes has 7 columns and 9545 rows of data.

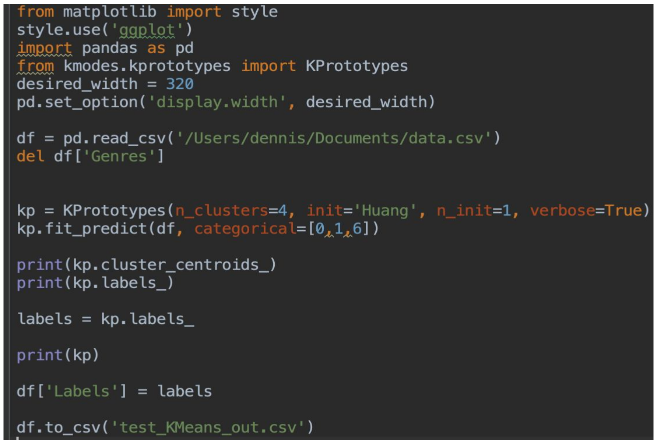
Table 1: The original column

Column Name	process
App	Removed
Category	
Rating	
Reviews	
Size	
Installs	
Type	Removed
Price	Removed
Content Rating	
Genres	
Last Updated	Removed
Current Ver	Removed
Android Ver	

Table 2: The attribute data after preprocessing

Column Name
Category
Rating
Reviews
Size
Installs
Content Rating
Genres
Android Ver

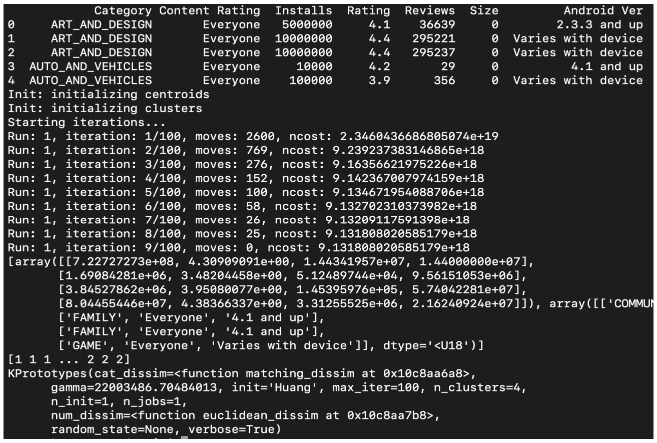
The k prototype with python language is used for clustering the data as shown Figure 1.



```python
from matplotlib import style
style.use('ggplot')
import pandas as pd
from kmodes.kprototypes import KPrototypes
desired_width = 320
pd.set_option('display.width', desired_width)

df = pd.read_csv('/Users/dennis/Documents/data.csv')
del df['Genres']


kp = KPrototypes(n_clusters=4, init='Huang', n_init=1, verbose=True)
kp.fit_predict(df, categorical=[0,1,6])

print(kp.cluster_centroids_)
print(kp.labels_)

labels = kp.labels_

print(kp)

df['Labels'] = labels

df.to_csv('test_KMeans_out.csv')
```

Figure 1: The K-Prototype algorithm

In the code above, our main goal is to divide the dataset into 4 clusters to analyze the app group based on the number of installs, ratings, reviews and sizes. The process can be seen in Figure 2.



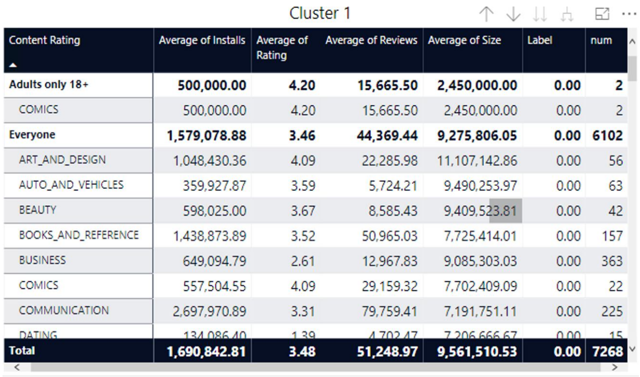
```
             Category Content Rating  Installs  Rating  Reviews  Size         Android Ver
0      ART_AND_DESIGN       Everyone   5000000     4.1    36639     0        2.3.3 and up
1      ART_AND_DESIGN       Everyone  10000000     4.4   295221     0  Varies with device
2      ART_AND_DESIGN       Everyone  10000000     4.4   295237     0  Varies with device
3   AUTO_AND_VEHICLES       Everyone     10000     4.2       29     0          4.1 and up
4   AUTO_AND_VEHICLES       Everyone    100000     3.9      356     0  Varies with device
Init: initializing centroids
Init: initializing clusters
Starting iterations...
Run: 1, iteration: 1/100, moves: 2600, ncost: 2.3460436686805074e+19
Run: 1, iteration: 2/100, moves: 769, ncost: 9.239237383146865e+18
Run: 1, iteration: 3/100, moves: 276, ncost: 9.16356621975226e+18
Run: 1, iteration: 4/100, moves: 152, ncost: 9.142367007974159e+18
Run: 1, iteration: 5/100, moves: 100, ncost: 9.134671954088706e+18
Run: 1, iteration: 6/100, moves: 58, ncost: 9.132702310373982e+18
Run: 1, iteration: 7/100, moves: 26, ncost: 9.13209117591398e+18
Run: 1, iteration: 8/100, moves: 25, ncost: 9.131808020585179e+18
Run: 1, iteration: 9/100, moves: 0, ncost: 9.131808020585179e+18
[array([[7.22727273e+08, 4.30909091e+00, 1.44341957e+07, 1.44000000e+07],
       [1.69084281e+06, 3.48204458e+00, 5.12489744e+04, 9.56151053e+06],
       [3.84527862e+06, 3.95080077e+00, 1.45395976e+05, 5.74042281e+07],
       [8.04455446e+07, 4.38366337e+00, 3.31255525e+06, 2.16240924e+07]]), array([['COMMUN
       ['FAMILY', 'Everyone', '4.1 and up'],
       ['FAMILY', 'Everyone', '4.1 and up'],
       ['GAME', 'Everyone', 'Varies with device']], dtype='<U18')]
[1 1 1 ... 2 2 2]
KPrototypes(cat_dissim=<function matching_dissim at 0x10c8aa6a8>,
      gamma=22003486.70484013, init='Huang', max_iter=100, n_clusters=4,
      n_init=1, n_jobs=1,
      num_dissim=<function euclidean_dissim at 0x10c8aa7b8>,
      random_state=None, verbose=True)
```

Figure 2: Process iteration

3. ANALYSIS RESULT

This study uses some scenario with the different number clustering with three, four and five group. However, the number cluster chosen is four. Each cluster is described as follows.

Cluster 1 (Clusterfamily): found that members of this cluster have the lowest average application downloads and the lowest rating as shown Figure 3, this cluster is dominated by the Family application category by 16.33% as shown Figure 4. This segment is dominated by content rating with values "everyone". This result gives the insight. Mostly family applications are used in all of type age. In this group, the members give the bad average rating(only 2,61) in everyone age. However the percentage of number category business is too high (5, 19%). This data also shows
the business category is disliked by the all type age.



Cluster 1

Content Rating	Average of Installs	Average of Rating	Average of Reviews	Average of Size	Label	num
Adults only 18+	**500,000.00**	**4.20**	**15,665.50**	**2,450,000.00**	**0.00**	**2**
COMICS	500,000.00	4.20	15,665.50	2,450,000.00	0.00	2
Everyone	**1,579,078.88**	**3.46**	**44,369.44**	**9,275,806.05**	**0.00**	**6102**
ART_AND_DESIGN	1,048,430.36	4.09	22,285.98	11,107,142.86	0.00	56
AUTO_AND_VEHICLES	359,927.87	3.59	5,724.21	9,490,253.97	0.00	63
BEAUTY	598,025.00	3.67	8,585.43	9,409,523.81	0.00	42
BOOKS_AND_REFERENCE	1,438,873.89	3.52	50,965.03	7,725,414.01	0.00	157
BUSINESS	649,094.79	2.61	12,967.83	9,085,303.03	0.00	363
COMICS	557,504.55	4.09	29,159.32	7,702,409.09	0.00	22
COMMUNICATION	2,697,970.89	3.31	79,759.41	7,191,751.11	0.00	225
DATING	134,086.40	1.39	4,702.47	7,206,666.67	0.00	15
Total	**1,690,842.81**	**3.48**	**51,248.97**	**9,561,510.53**	**0.00**	**7268**

Figure 3: Cluster family data

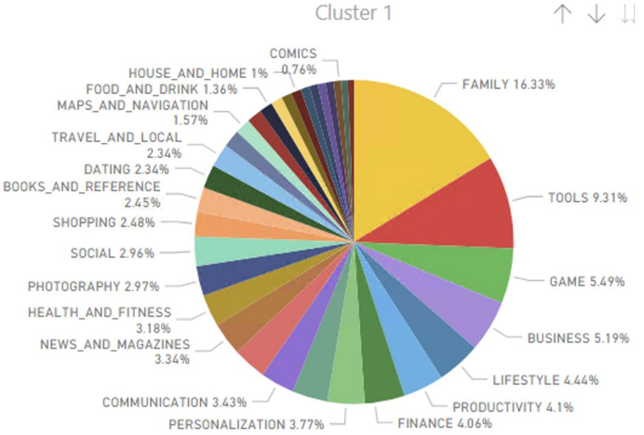
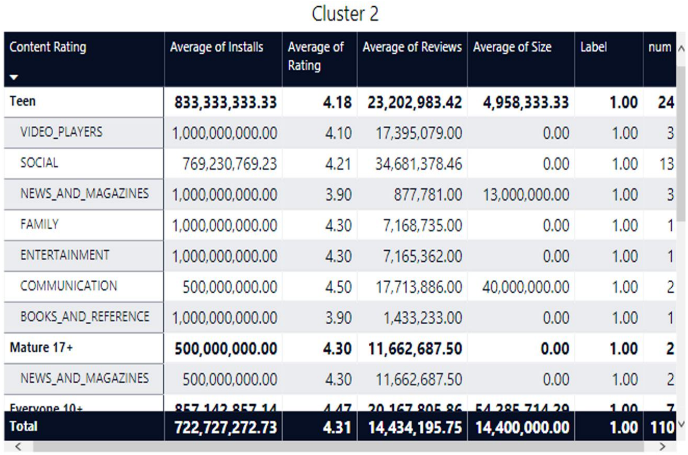


Figure 4: Percentage cluster family based on category

Cluster 2 (cluster communication): found that the members of cluster 2 have the highest average number of downloads as shown Figure 5, this cluster is dominated by the communication application category by 22.73% as shown Fig 6. In this group, the number of each category is almosst same.


Cluster 2

Content Rating	Average of Installs	Average of Rating	Average of Reviews	Average of Size	Label	num
Teen	**833,333,333.33**	**4.18**	**23,202,983.42**	**4,958,333.33**	**1.00**	**24**
VIDEO_PLAYERS	1,000,000,000.00	4.10	17,395,079.00	0.00	1.00	3
SOCIAL	769,230,769.23	4.21	34,681,378.46	0.00	1.00	13
NEWS_AND_MAGAZINES	1,000,000,000.00	3.90	877,781.00	13,000,000.00	1.00	3
FAMILY	1,000,000,000.00	4.30	7,168,735.00	0.00	1.00	1
ENTERTAINMENT	1,000,000,000.00	4.30	7,165,362.00	0.00	1.00	1
COMMUNICATION	500,000,000.00	4.50	17,713,886.00	40,000,000.00	1.00	2
BOOKS_AND_REFERENCE	1,000,000,000.00	3.90	1,433,233.00	0.00	1.00	1
Mature 17+	**500,000,000.00**	**4.30**	**11,662,687.50**	**0.00**	**1.00**	**2**
NEWS_AND_MAGAZINES	500,000,000.00	4.30	11,662,687.50	0.00	1.00	2
Everyone 10+	**857,142,857.14**	**4.47**	**20,167,805.86**	**54,285,714.29**	**1.00**	**7**
Total	**722,727,272.73**	**4.31**	**14,434,195.75**	**14,400,000.00**	**1.00**	**110**

Figure 5: Cluster commuication data

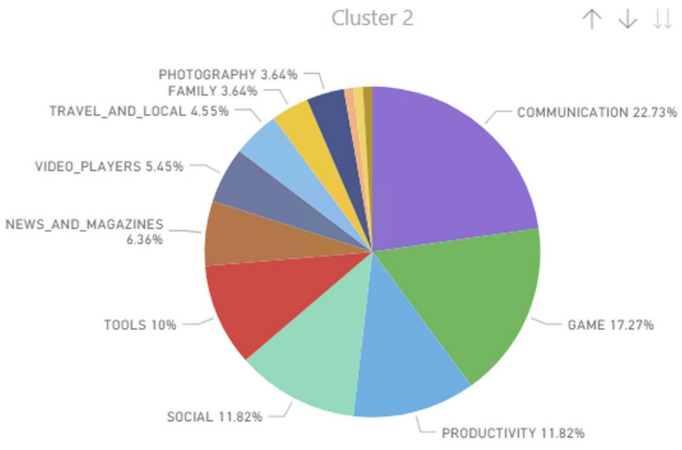
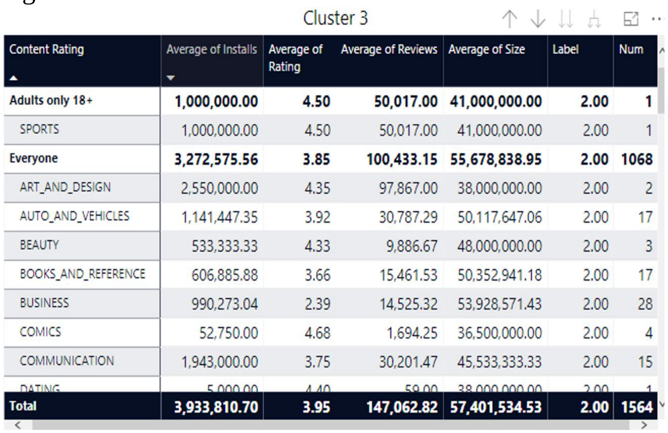


Figure 6: Percentage cluster communication based on category

Cluster 3 (Cluster game-family). This member consists the the second lowest download as shown Fig 7. The rating in this cluster also ranks the second lowest. This cluster is in the application category of family and game category as shown Figure 8.


Cluster 3

Content Rating	Average of Installs	Average of Rating	Average of Reviews	Average of Size	Label	Num
Adults only 18+	**1,000,000.00**	**4.50**	**50,017.00**	**41,000,000.00**	**2.00**	**1**
SPORTS	1,000,000.00	4.50	50,017.00	41,000,000.00	2.00	1
Everyone	**3,272,575.56**	**3.85**	**100,433.15**	**55,678,838.95**	**2.00**	**1068**
ART_AND_DESIGN	2,550,000.00	4.35	97,867.00	38,000,000.00	2.00	2
AUTO_AND_VEHICLES	1,141,447.35	3.92	30,787.29	50,117,647.06	2.00	17
BEAUTY	533,333.33	4.33	9,886.67	48,000,000.00	2.00	3
BOOKS_AND_REFERENCE	606,885.88	3.66	15,461.53	50,352,941.18	2.00	17
BUSINESS	990,273.04	2.39	14,525.32	53,928,571.43	2.00	28
COMICS	52,750.00	4.68	1,694.25	36,500,000.00	2.00	4
COMMUNICATION	1,943,000.00	3.75	30,201.47	45,533,333.33	2.00	15
DATING	5,000.00	4.40	59.00	38,000,000.00	2.00	1
Total	**3,933,810.70**	**3.95**	**147,062.82**	**57,401,534.53**	**2.00**	**1564**

Figure 7: Cluster game-family data

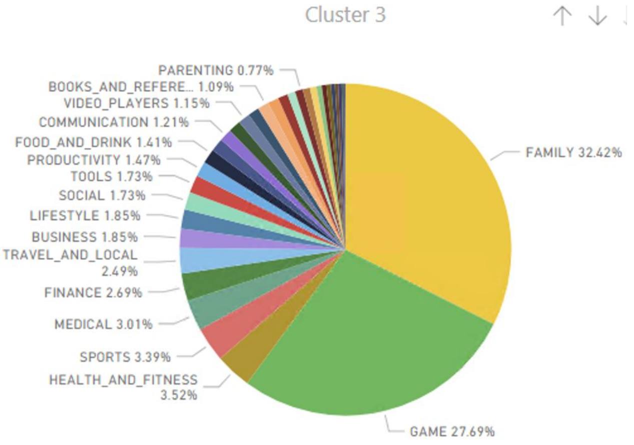
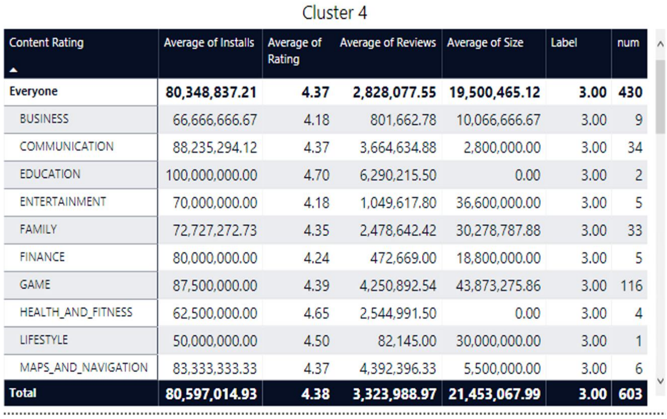


Figure 8: Percentage cluster game-family based on category

Cluster 4 (cluster game-photography)has the second highest level of application installs and the second highest rating as shown Fig. 9 . This cluster is dominated by the game and photography category as shown Figure10.


Cluster 4

Content Rating	Average of Installs	Average of Rating	Average of Reviews	Average of Size	Label	num
Everyone	**80,348,837.21**	**4.37**	**2,828,077.55**	**19,500,465.12**	**3.00**	**430**
BUSINESS	66,666,666.67	4.18	801,662.78	10,066,666.67	3.00	9
COMMUNICATION	88,235,294.12	4.37	3,664,634.88	2,800,000.00	3.00	34
EDUCATION	100,000,000.00	4.70	6,290,215.50	0.00	3.00	2
ENTERTAINMENT	70,000,000.00	4.18	1,049,617.80	36,600,000.00	3.00	5
FAMILY	72,727,272.73	4.35	2,478,642.42	30,278,787.88	3.00	33
FINANCE	80,000,000.00	4.24	472,669.00	18,800,000.00	3.00	5
GAME	87,500,000.00	4.39	4,250,892.54	43,873,275.86	3.00	116
HEALTH_AND_FITNESS	62,500,000.00	4.65	2,544,991.50	0.00	3.00	4
LIFESTYLE	50,000,000.00	4.50	82,145.00	30,000,000.00	3.00	1
MAPS_AND_NAVIGATION	83,333,333.33	4.37	4,392,396.33	5,500,000.00	3.00	6
Total	**80,597,014.93**	**4.38**	**3,323,988.97**	**21,453,067.99**	**3.00**	**603**

Figure 9: Cluster game-photography data

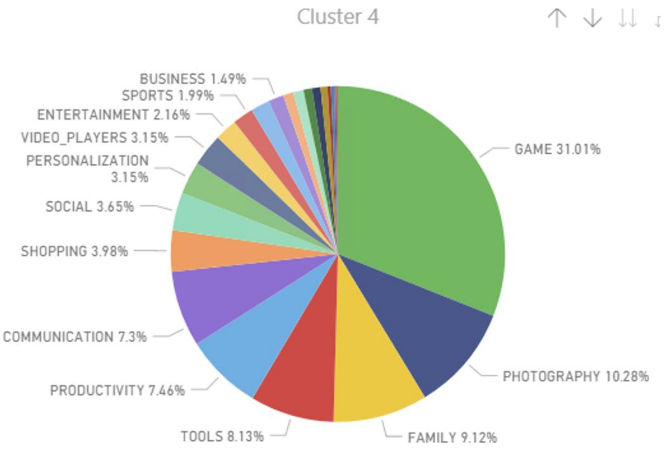


Figure 10: Percentage cluster game-family based on category

5. CONCLUSION

Based on the analysis of the dataset using the k-modes prototype method, insight is obtained in the form of information that can provide an overview for developers who will create or develop new applications by referring to the type of application that is the most successful at the moment and the genre most favored by the majority of application users. mobile. This will be quite helpful in providing ideas / ideas that are fundamental, for example for gaming applications having diverse genres such as Adventure, Role Play Games, Multiplayer Games, Action, Fighting, Strategy, and so on. That way developers will have a clearer vision in making applications that have the potential to obtain greater profits.

REFERENCES

[1] S. McIlroy, N. Ali, and A. E. Hassan, “Fresh apps: an empirical study of frequently-updated mobile apps in the Google play store,” *Empir. Softw. Eng.*, vol. 21, no. 3, pp. 1346–1370, 2016. https://doi.org/10.1007/s10664-015-9388-2

[2] N. Viennot, E. Garcia, and J. Nieh, “A measurement study of google play,” in *The 2014 ACM international conference on Measurement and modeling of computer systems*, 2014, pp. 221–233. https://doi.org/10.1145/2637364.2592003

[3] I. Malavolta, S. Ruberto, T. Soru, and V. Terragni, “End users’ perception of hybrid mobile apps in the google play store,” in *2015 IEEE International Conference on Mobile Services*, 2015, pp. 25–32.

[4] R. S. Sangam and H. Om, “An equi-biased k-prototypes algorithm for clustering mixed-type data,” *Sadhana - Acad. Proc. Eng. Sci.*, vol. 43, no. 3, pp. 1–12, 2018. https://doi.org/10.1007/s12046-018-0823-0

[5] G. Gan and M. K. P. Ng, “K-Means Clustering With Outlier Removal,” *Pattern Recognit. Lett.*, vol. 90, pp. 8–14, 2017.

[6] S. Sitohang, A. S. Girsang, and Suharjito, “Prediction of the number of airport passengers using fuzzy C-means and adaptive neuro fuzzy inference system,” *Int. Rev. Autom. Control*, vol. 10, no. 3, 2017.

[7] A. S. Girsang, T. W. Cenggoro, and K.-W. Huang, “Fast ant colony optimization for clustering,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 12, no. 1, pp. 78–86, 2018.

[8] C. Zhang, D. Ouyang, and J. Ning, “An artificial bee colony approach for clustering,” *Expert Syst. Appl.*, vol. 37, no. 7, pp. 4761–4767, 2010. https://doi.org/10.1016/j.eswa.2009.11.003

[9] A. S. Girsang, A. Yunanto, and A. H. Aslamiah, “A hybrid cuckoo search and K-means for clustering problem,” in *ICECOS 2017 - Proceeding of 2017 International Conference on Electrical Engineering and Computer Science: Sustaining the Cultural Heritage Toward the Smart Environment for Better Future*, 2017.

[10] J. Ji, T. Bai, C. Zhou, C. Ma, and Z. Wang, “An improved k-prototypes clustering algorithm for mixed numeric and categorical data,” *Neurocomputing*, vol. 120, pp. 590–596, 2013. https://doi.org/10.1016/j.neucom.2013.04.011

[11] P. Arora, Deepali, and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm for Big Data,” *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016.

[12] K. S. Ranti, K. Salim, and A. S. Girsang, “Clustering Steam User Behavior Data using K-Prototypes Algorithm,” in *Journal of Physics: Conference Series*, 2019, vol. 1367, no. 1. https://doi.org/10.1088/1742-6596/1367/1/012018