Comparison of K-Means Clustering & Logistic Regression on University data to differentiate between Public and Private University

Adhien Kenya Anima Estetikha [1,*], Deden Hardan Gutama [2], Musthofa Galih Pradana [3], Dhina Puspasari Wijaya [4]

[1] Amikom University, Indonesia
[2,3,4] Alma Ata University, Indonesia
[1] kenya.akae@gmail.com *; [2] hardan@almaata.ac.id; [3] mgalihpradana@almaata.ac.id; [4] dhina.puspa@almaata.ac.id
* corresponding author

(Received February 4, 2021 Revised February 23, 2021 Accepted February 27, 2021, Available online March 1, 2021)

Abstract

The development of advances in educational methods has developed in the last few decades. especially at the higher education level such as college. The rising interest of students in pursuing their higher education has caused the sector to be split into two sectors, both private and public universities. This difference raises several questions recently about how the two types differ in carrying out the educational process. whether there is a difference in terms of cost, service or quality, we really can't tell exactly. For this study, we will try to use the K-Means Clustering & Logistic Regression to group the University into two groups, Private and Public and then compare the two model accuracy. The results of this study show that the results obtained from the K-Means clustering model (22%) are much lower than the Logistic Regression model (91%).

Keywords: K-Means, Regression, Clustering

1. Introduction

The goal of many universities is to successfully expand their education until 2021. This objective has opened the way for numerous cultural exchanges, the flow of ideas, the worldwide export / import of products and even the illicit or lawful movement of citizens across borders. Upon this topic, the ability to pursue an education was a major explanation offered to us. Furthermore, Lim et al [1] explained that there has been an increase in globalization in the higher education sector during the last few decades. The demands for university education was shown in the large flow of foreign students and the rising number of universities and colleges offering cross-border educational services. For a long time, this pattern has been around and analysis has shown it. Lately, the number of students going abroad for higher education continues to increase.

The economy also expanded at an accelerating rate in the mid-1980s, and the market for highly qualified business managers in public and private sector organizations grew. Many students interested in business education go overseas, with just six public universities in the world and admission-supporting quotas[2]. Private colleges and public institutions around the world are the subject of this initiative. The main aim of this research was to examine, evaluate and analyze the discrepancies between students from the colleges / universities affiliated with tertiary level education. As several new and upcoming colleges/universities have recently been launched in the education sector, this paper could be useful to provide new insights into the industry.

Based on the level of education they receive from the educational institution, student satisfaction is often evaluated. Service quality or quality, as others claim, is a significant criterion that allows students to decide their choice of college or university. Service organizations such as higher education institutions are undoubtedly under relentless pressure, according to[3], to outperform their rivals in the name of high quality of service. This may very well be the aspect that divides between favored and unprofitable universities. Service output will now consist of many aspects. Several other research would suggest that the choices and expectations made at a specific college by a student often depend on the lecturer's teaching success and ability. Ollin[4] argues that the degree to which lecturers are highly educated and certified will further transform the education sector's long-term development.

2. Literature Review

2.1. Difference between private & public university

Indeed, There are differences between PTS and PTN that often have the same characters, but also distinct characteristics. Tang[5] notes that 'private higher education has developed faster than the public system and can be viewed as a public higher education system which is complimentary and complementary'. This is clear from research by Middlehurst & Woodfield[6], which indicates that while higher education demand is strong in most countries, this demand can not be fulfilled by local universities. Obviously, in different countries, private universities compete with public universities. It is actually reasonable to say that all institutions, private and public, go hand in hand at this phase and evolve at the same time. More than 60 percent of students seek a bachelor's degree at public colleges, with a substantial reduction in the number of students enrolling in diploma programs. On the other hand, 40 percent of students in private schools seek certificates at the diploma level and about the same amount of bachelor's degree programs.

Abdullah and Warokka [7] find that student expectations of the teaching and learning process, teaching and learning support facilities such as libraries, computer and laboratory facilities, learning conditions such as class rooms, labs, social rooms and university buildings are the supporting factors that can influence the degree of student satisfaction. Support for facilities such as healthcare facilities, classrooms, student residences, student services and student external elements, such as banking, transport[8]. As this is the foundation of higher education, student expectations and ideas of the teaching and learning process are considered vital. In an environment conducive to studying, students can undoubtedly look to receive good teaching[9].

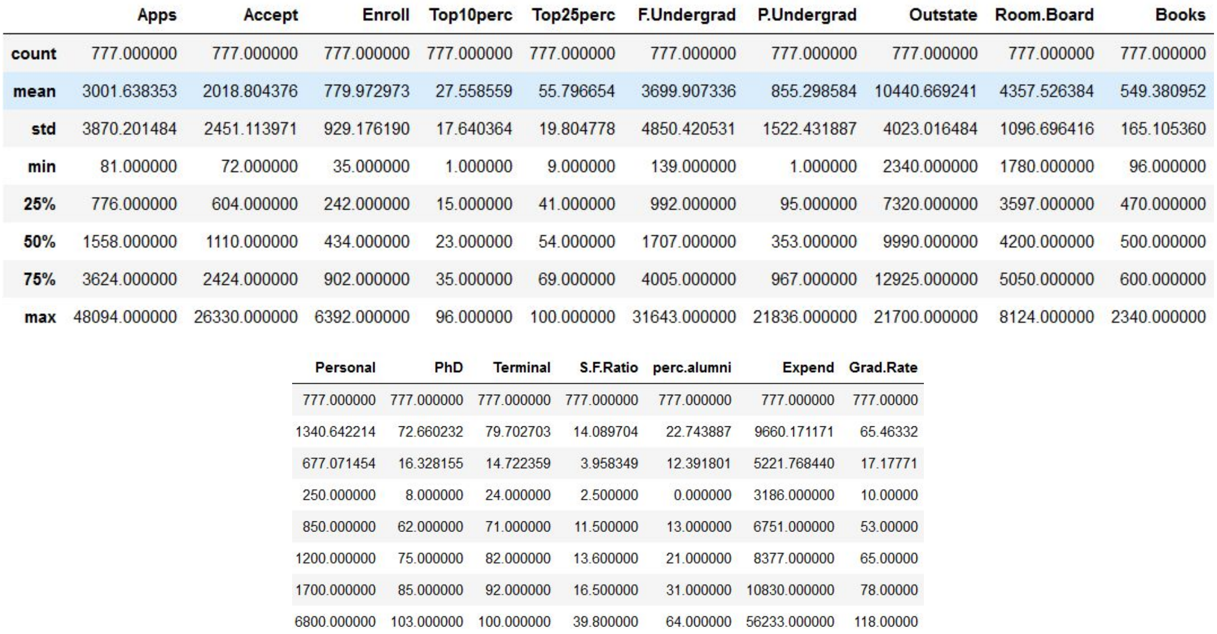
It is generally accepted that any measure of the efficiency and consistency of the provision of education[10] is often provided by the availability and quality of physical inputs. The cost and cost study of teaching reveals that public universities are spending more on classrooms and libraries, whereas private colleges are spending more on laboratories and computers[11,12]. Public universities, however, usually have good classrooms and library services, while private universities have superior labs and electronic facilities. Based on these considerations, students will select one of these institutions.

2.2. Quality of learning at higher education levels

As higher education institutions compete with each other, consistency has arisen as an adopted subject. It is not difficult to understand that for the growth of a college or university, quality is important[13]. One of the criteria that makes an organization distinct from others has always been quality. For them, having this specificity is a strategic advantage which makes them special. The quality of the faculty, academic credibility, skilled academics, foreign reach, adequate use of capital, partnerships and networks are the ten features of a world-class university described by Saunus[14]. It embraces many fields, is technologically trained, and practices the art of effective leadership. Equally significant for the sustainability and growth of the university are all the conditions listed above. On the other hand, in Zakaria, Ahmad, and Norzaidi, Abubakar[15] states that a world-class university must have twelve features. The twelve features contain or consist of separate variables, ranging from lecturers, teachers, administrative personnel, and all facets of the growth of universities. The features include: government-accredited niche initiatives, cross-border partnerships in science and research, accessibility of employees and student mobility programs, registration and number of foreign students enrolled, international honors from international organisations, good governance and global appreciation of graduates. For all these significant parameters, it is obvious that consistency is broadly defined and consists of different items. U. Suuroja[16], Lehtinen and J.R Lehtinen note that service quality comes in three dimensions, namely physical quality, digital quality and the quality of the business. Another well-known service efficiency model was introduced in 1982 by Gronroos 1982 in Tolpa[17]. It distinguishes two kinds of standard of service: technological and practical.

3. Methodology

The case of this research is a tertiary institution which is one of the pillars that supports all educational activities, ranging from new student admissions, academic service administration systems, financial information systems, personnel information systems, e-libraries, e-learning, assembly. registration system, Helpdesk system, and various other information systems and applications. The research activity is depicted in Figure 1, starting from visualizing the dataset used and explaining each of its features, then analyzing exploratory data to find & select the features that will be used to create the model, and ending with model evaluation.

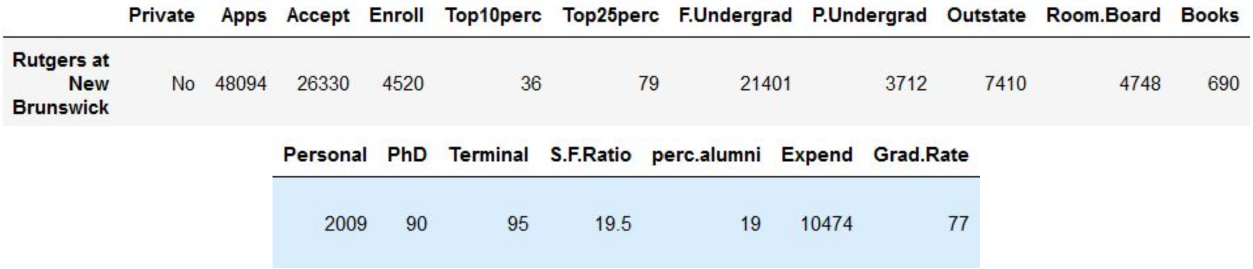


	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000

Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
1340.642214	72.660232	79.702703	14.089704	22.743887	9660.171171	65.46332
677.071454	16.328155	14.722359	3.958349	12.391801	5221.768440	17.17771
250.000000	8.000000	24.000000	2.500000	0.000000	3186.000000	10.00000
850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.00000
1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.00000
1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.00000
6800.000000	103.000000	100.000000	39.800000	64.000000	56233.000000	118.00000

Fig. 1. Basic Statistics of Datasets

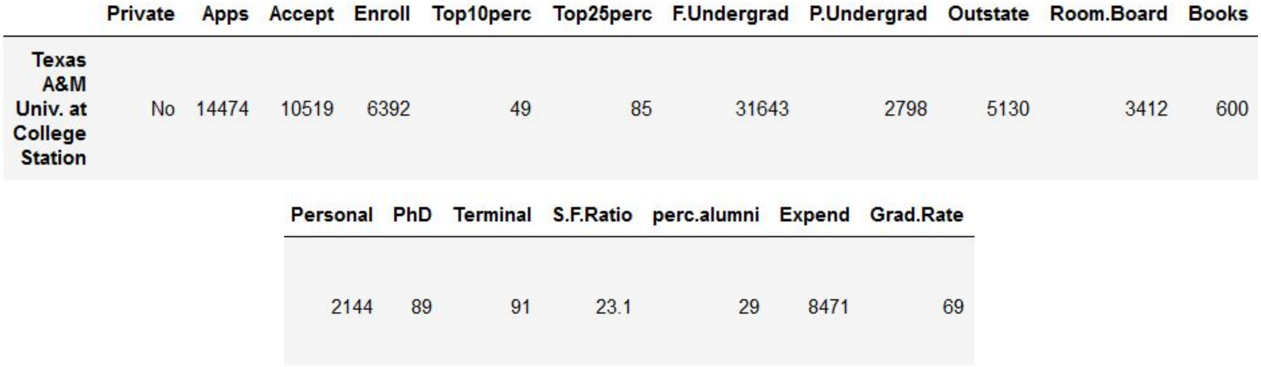
Just looking at the maximum value of the image above (figure 1), Typically high-end private colleges will bill up to $60 thousand dollars/year, but the actual out-of-state tuition (maximum expenditure) is on average. There are many factors that stand out. The sum was just $21,700, so that we can reasonably presume that these tuition fees are reported instead of every year on a semester basis. Then, there are institutions with PhD degrees that have more than 100 percent of the professors. This sounds suspicious, but the handful of high-achieving faculty that may have dual PhD degrees who have double-counted may maybe justify it. and lastly, we also have a university with a 118% graduation rate. We will discuss this anomaly later. the next thing we will look for is the university with the highest number of applicants, the highest admission, and the most students.



	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
Rutgers at New Brunswick	No	48094	26330	4520	36	79	21401	3712	7410	4748	690

Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
2009	90	95	19.5	19	10474	77

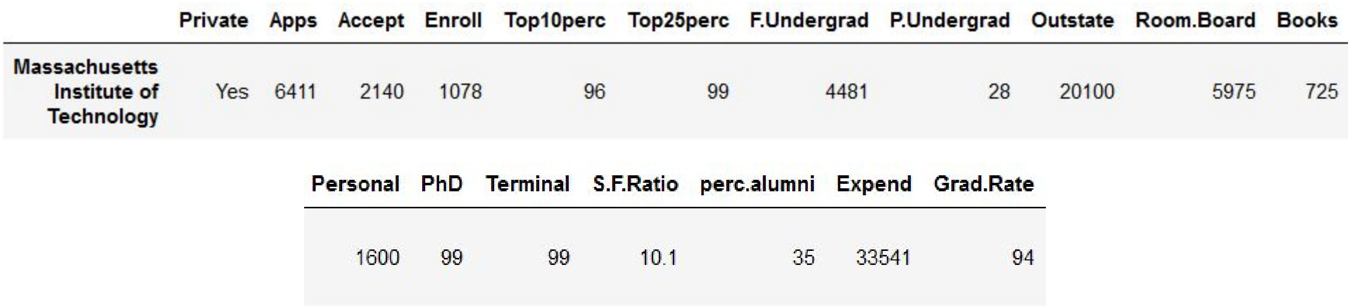
Fig. 2. Highest applicant



	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
Texas A&M Univ. at College Station	No	14474	10519	6392	49	85	31643	2798	5130	3412	600

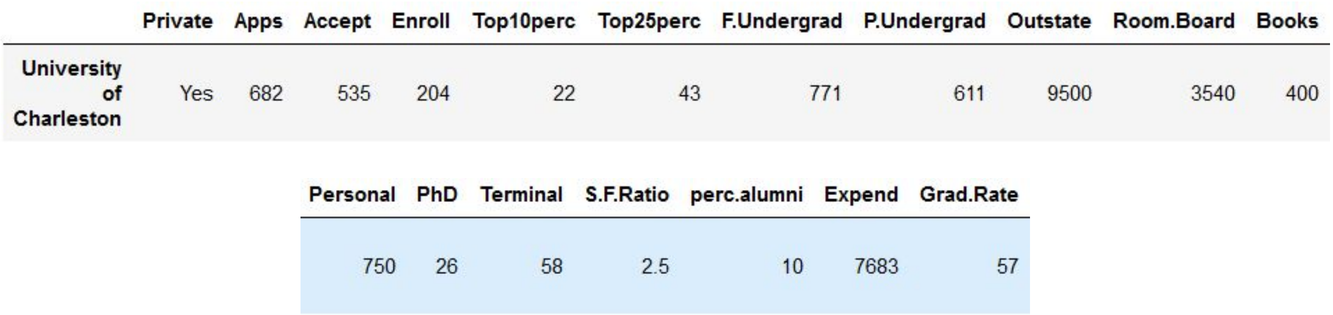
Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
2144	89	91	23.1	29	8471	69

Fig. 3. Highest enrollment



	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Massachusetts Institute of Technology	Yes	6411	2140	1078	96	99	4481	28	20100	5975	725	1600	99	99	10.1	35	33541	94

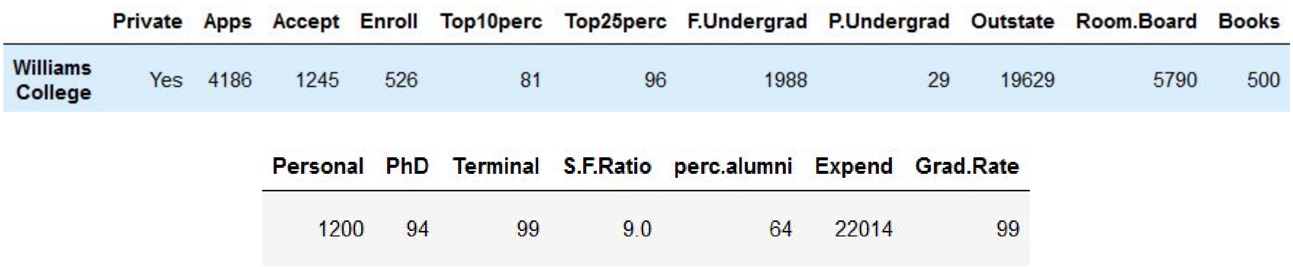
Fig. 4. Highest Top 10 Percentage



	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
University of Charleston	Yes	682	535	204	22	43	771	611	9500	3540	400	750	26	58	2.5	10	7683	57

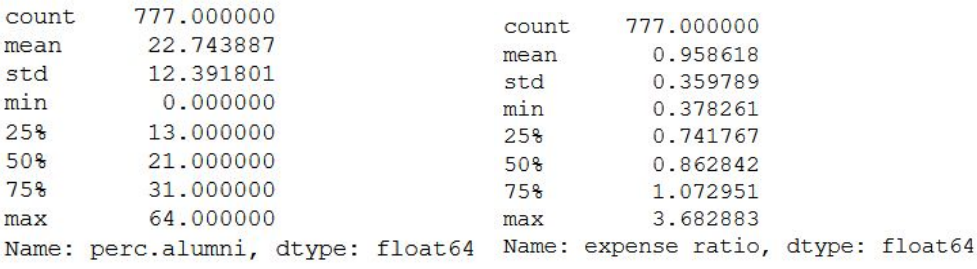
Fig. 5. Highest Student/faculty ratio

It turns out, on the basis of the data received, that Rutgers has the highest number of registrants and intakes, but not the highest number. Your student-faculty ratio is another thing that high schools and colleges boast about: the smaller the better. Texas A&M managed to have the largest attendance out of the 777 universities in this data collection. In figure 4, on the other hand, it is obvious that MIT has the number of students in the top 10% rank. Another thing that high schools and universities boast about is their student-faculty ratio: the lower the better. The winner here is the University of Charleston, with a fairly low ratio of 2.5, which is practically an intimate teaching environment such as homeschooling or private tutoring. Consider that the 25th percentile of the student-faculty ratio is 11.5. The next question we will examine is which university has the highest alumni donation rate? It is not surprising that Williams has the largest number of donations as it is a small liberal arts college (with ~ 2,000 total students) serving mostly wealthy families with strong connections. Calling .describe () previously told us that the median donation percentage was very low, namely 21%, hence the abundance of job opportunities on campus as student callers.



	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Williams College	Yes	4186	1245	526	81	96	1988	29	19629	5790	500	1200	94	99	9.0	64	22014	99

Fig. 6. Highest Alumni Donation Percentage



```
count    777.000000
mean      22.743887
std       12.391801
min        0.000000
25%       13.000000
50%       21.000000
75%       31.000000
max       64.000000
Name: perc.alumni, dtype: float64
```

```
count    777.000000
mean       0.958618
std        0.359789
min        0.378261
25%        0.741767
50%        0.862842
75%        1.072951
max        3.682883
Name: expense ratio, dtype: float64
```

Fig. 7. Expense Ratio Output

So who really gets their money from university education at the ratio of school expenses for each student to other tuition fees. Show in Figure 7 above, that 69 percent of colleges spend more money in deciding to what they owe for their students. In other words, as seen from tuition fees alone, almost 69 percent of colleges lost their money to their students. One would ask if any of the Ivy League schools are potentially the schools with the best ratio. However, with a fee ratio of 3.68, it turned out not to be the University of Alabama at Birmingham. However, if UAB pays about 4 times the tuition costs above the semester fees on each undergraduate, a more pessimistic way of looking at it is, then the corporation probably won't last much longer.

3.1. Visualization

We will examine how tuition rates differ with the number of students and whether the school is public or private after doing some further study. For private schools, we suspect the school costs are higher because they do not accept government funds as public schools. As more students means more jobs and more houses, the number of full-time students will depress student figures, although it is likely that universities can benefit from economies of scale with large numbers of students to create programs.

The first hypothesis we made was that private universities tend to have higher tuition fees and higher rooms and dormitories because they have generally better facilities.

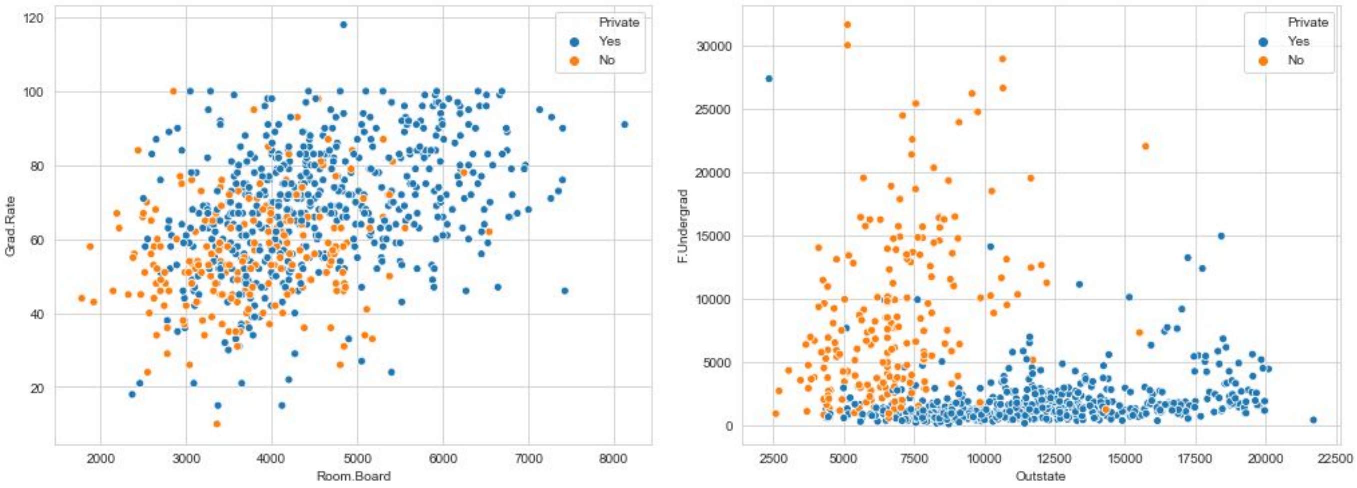


Fig. 8. The number of full-time undergraduates vs out-of-state tuition

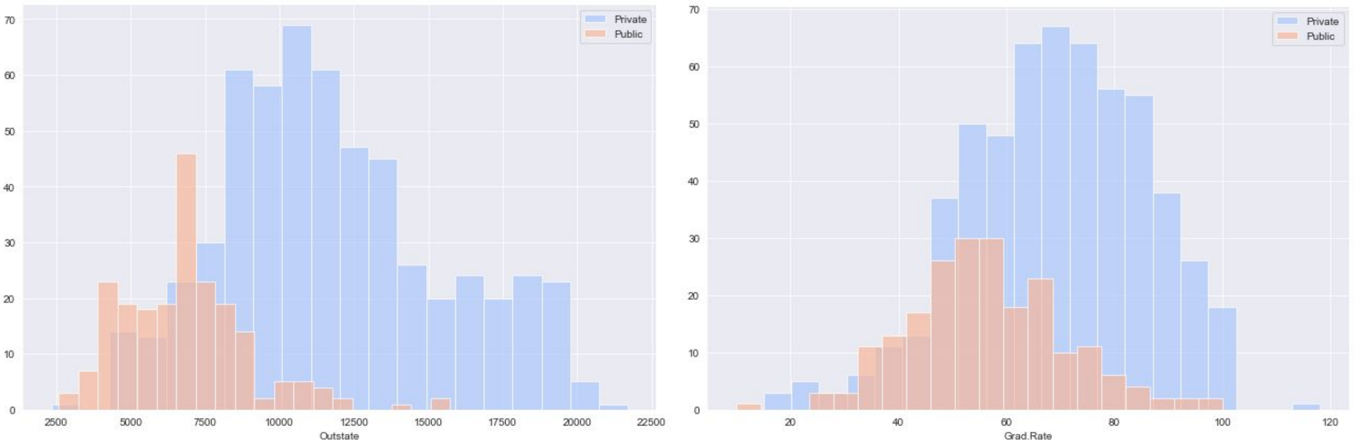


Fig. 9. For private and public schools, out-of-state fees

As one would imagine, for private colleges, the expense of public schools is higher, and so are the graduation rates, maybe because if a pupil spends so much money in their education, parents are more likely to allow their children to complete college. Today, returning to the school with a graduation record of over 100%, it turns out that Cazenovia University is the culprit. Perhaps this school pulled the easy and counted the dual majors that graduated twice, which is obviously very ridiculous.

3.2. Model Creation

In this study, KMeans will be imported from Scikit-learn, a Python machine learning program. To help us understand how the K-Means modeling operates, the number of clusters is the following: first. The basic problem domain defines whether we prefer 2 or 4 or 10. We sorted public and private schools in this situation, so we can pick the number of clusters to be 2. If you examine genetic variance in a population and recognize that 7 variations are identified in advance, so you can select 7. After that, distribute each data point to groups at random.

In each group, take the data centre point. This is a multivariate math word for "the average of all the data in each cluster." for those of you who think the centroid is a kind of cool-sounding asteroid. For 1-D data, you already understand this intuitively: a food item's average price gives you a number such as $ 52. If you calculate the average price and the average amount of ingredients (2 dimensions), you get two figures. Such as $52 and 2 objects. We have 18 characteristics in this dataset, so each centroid corresponds to an 18-D range of coordinates. The most important thing is that, until there is no further category adjustment, we need to reassign each data point to the category corresponding to the closest center of mass.

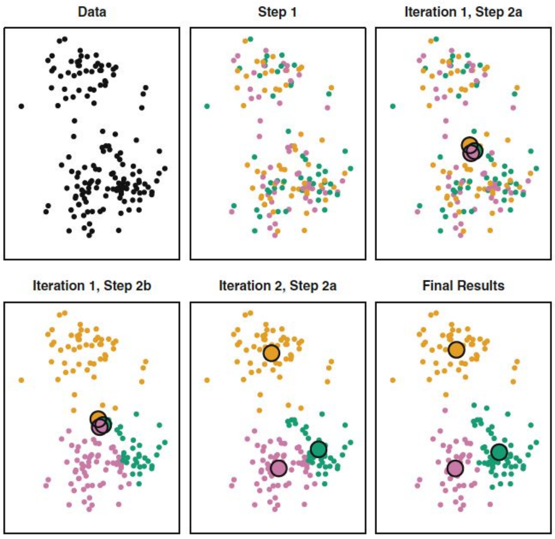
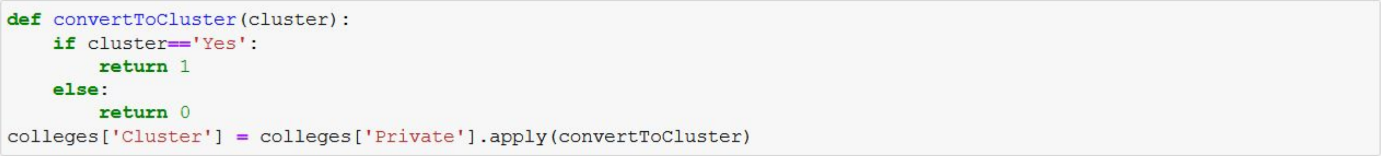


Fig. 10. Step by step K-means grouping

There's no perfect way to evaluate groupings if you don't have labels. Before evaluating grouping performance, it is very important to make sure that the data set we are working with has a clustering tendency and does not contain points that are uniformly distributed. If the data does not contain clustering trends, then the clusters identified by any sophisticated clustering algorithm may not be relevant. The distribution of points that are not uniform in the data set is important in grouping. In this case, however, the college data set tells us whether each school is public or private, so we can cross-validate our K-means model with this label to compare the performance of generally supervised and unsupervised models. First, we need to change the "Private: Yes or No" column to 0 and 1 which can be understood by the K-means model.

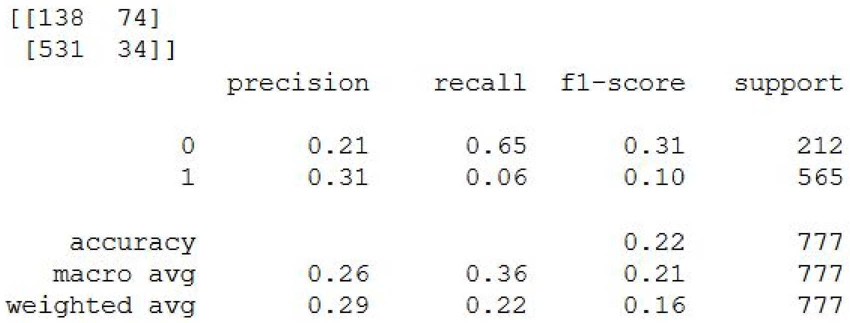


```
def convertToCluster(cluster):
    if cluster=='Yes':
        return 1
    else:
        return 0
colleges['Cluster'] = colleges['Private'].apply(convertToCluster)
```

Fig. 11. Converting K-means Model

4. Results and Discussion

Two fast methods to test a machine learning model's efficiency are to look at the confusion matrix and classification reports. Classification is the task of presenting a group of findings on the basis of certain parameters. Classification is part of supervised learning in machine learning, meaning that the data used to train the model has a label which identifies each category. An significant step in a machine learning model's life cycle is its assessment of results. The confusion matrix and classification reports are two methods used to test classification models.



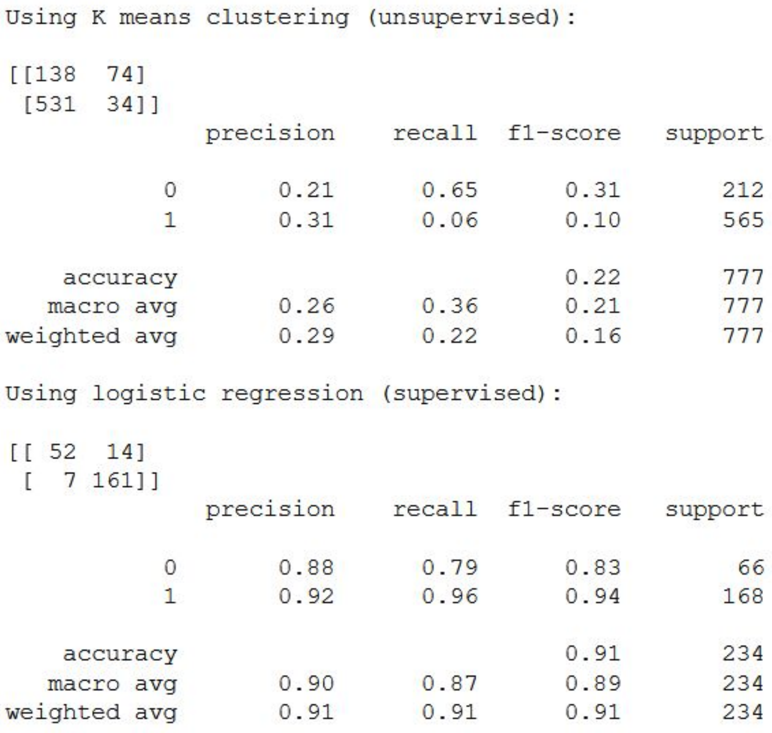
```
[[138  74]
 [531  34]]
              precision    recall  f1-score   support

           0       0.21      0.65      0.31       212
           1       0.31      0.06      0.10       565

    accuracy                           0.22       777
   macro avg       0.26      0.36      0.21       777
weighted avg       0.29      0.22      0.16       777
```

Fig. 12. The unsupervised K-means model, uncertainty matrix and classification report from Scikit-learn

As you can see from the model findings above, the accuracy achieved by using K-Means clustering is not high enough to be considered an effective model, i.e. 22%. But note the concept of unsupervised: this model tries to understand, without marks, the mess of 18 features that we have. To test its efficiency, let's now try to equate it with a supervised model. As you can see in Figure 13 below, the precision is lifted from 22 percent to 91 percent using basic logistic regression, a supervised learning model, which can be boosted by the option of other models.



```
Using K means clustering (unsupervised):

[[138  74]
 [531  34]]
              precision    recall  f1-score   support

           0       0.21      0.65      0.31       212
           1       0.31      0.06      0.10       565

    accuracy                           0.22       777
   macro avg       0.26      0.36      0.21       777
weighted avg       0.29      0.22      0.16       777

Using logistic regression (supervised):

[[ 52  14]
 [  7 161]]
              precision    recall  f1-score   support

           0       0.88      0.79      0.83        66
           1       0.92      0.96      0.94       168

    accuracy                           0.91       234
   macro avg       0.90      0.87      0.89       234
weighted avg       0.91      0.91      0.91       234
```

Fig. 13. Performance analysis for unsupervised and supervised learning models (logistic regression)

5. Conclusion

By knowing, the results of the model accuracy obtained are 91% percent for supervised learning with logistic regression and 22% for unsupervised learning with K-Means clustering. It can be concluded that the supervised learning model can be more reliable in this study. However, we must remember the unsupervised definition: the model aims to understand the 18 features in the database, without labels & This is not an easy task, so it might be unexpected if the results are accurate. what he got was far from the expected standard. With the results obtained, we hope that this study can be used as a good basis for educators, students, or other researchers in future research.

References

[1] Y. M. Lim, C. S. Yap, and T. H. Lee, "Destination choice, service quality, satisfaction, and consumerism: International students in Malaysian institutions of higher education," *African J. Bus. Manag.*, vol. 5, no. 5, pp. 1691–1702, 2011, doi: 10.5897/AJBM10.610.

[2] J. Eckert, M. Luqmani, S. Newell, Z. Quraeshi, and B. Wagner, "Developing Short-Term Study Abroad Programs: Achieving Successful International Student Experiences," *Am. J. Bus. Educ.*, vol. 6, no. 4, pp. 439–458, 2013, doi: 10.19030/ajbe.v6i4.7943.

[3] A. Shekarchizadeh, A. Rasli, and H. Hon-Tat, "SERVQUAL in Malaysian universities: Perspectives of international students," *Bus. Process Manag. J.*, vol. 17, no. 1, pp. 67–81, 2011, doi: 10.1108/14637151111105580.

[4] R. Ollin, "Learning from industry: Human resource development and the quality of lecturing staff in further education," *Qual. Assur. Educ.*, vol. 4, no. 4, pp. 29–36, 1996, doi: 10.1108/09684889610146172.

[5] B. T. L. Tang and T. L. Tang, "By Thomas Li-Ping Tang, PhD, and Theresa Li-Na Tang," vol. 41, no. 1, pp. 97–126, 2012.

[6] R. Middlehurst and S. Woodfield, "The role of transnational, private, and for-profit provision in meeting global demand for tertiary education: mapping, regulation and impact: Summary Report," 2004, [Online]. Available: http://www.col.org/colweb/site/pid/3108%5Cnhttp://eprints.kingston.ac.uk/6910/1/Middlehurst-R-6910.pdf%5Cnhttp://eprints.kingston.ac.uk/1728/%5Cnhttp://eprints.kingston.ac.uk/1727/.

[7] J. Hanaysha, H. Abdullah, and A. Warokka, "Service Quality and Students' Satisfaction at Higher Learning Institutions: The Competing Dimensions of Malaysian Universities' Competitiveness," *J. Southeast Asian Res.*, vol. 2011, pp. 1–10, 2011, doi: 10.5171/2011.855931.

[8] L. Harvey and P. Barr, "STUDENT FEEDBACK A report to the Higher Education Funding Council for England," *Quality*, no. October, 2001.

[9] J. Mogan and J. E. Knox, "Characteristics of 'best' and 'worst' clinical teachers as perceived by university nursing faculty and students," *J. Adv. Nurs.*, vol. 12, no. 3, pp. 331–337, 1987, doi: 10.1111/j.1365-2648.1987.tb01339.x.

[10] M. E. Lockheed *et al.*, "Educational Evaluation and Policy How Textbooks Affect Achievement in Developing Countries: Evidence From Thailand," *Anal. Winter*, vol. 8, no. 4, pp. 379–392, 1986, [Online]. Available: http://eepa.aera.net.

[11] J. Johnes, "Measuring teaching efficiency in higher education: An application of data envelopment analysis to economics graduates from UK Universities 1993," *Eur. J. Oper. Res.*, vol. 174, no. 1, pp. 443–456, 2006, doi: 10.1016/j.ejor.2005.02.044.

[12] R. Wilkinson and I. Yussof, "Public and private provision of higher education in Malaysia: A comparative analysis," *High. Educ.*, vol. 50, no. 3, pp. 361–386, 2005, doi: 10.1007/s10734-004-6354-0.

[13] M. Sadiq Sohail, J. Rajadurai, and nor Azlin Abdul Rahman, "Managing quality in higher education: A Malaysian case study," *Int. J. Educ. Manag.*, vol. 17, no. 4, pp. 141–146, 2003, doi: 10.1108/09513540310474365.

[14] J. M. Saunus *et al.*, "Integrated genomic and transcriptomic analysis of human brain metastases identifies alterations of potential clinical significance," *J. Pathol.*, vol. 237, no. 3, pp. 363–378, 2015, doi: 10.1002/path.4583.

[15] Z. Bin Zakaria, A. Bin Ahmad, and M. Daud Norzaidi, "Determining World Class University from the Evaluation of Service Quality and Students Satisfaction Level: An Empirical Study in Malaysia," *Int. J. Sci. Res. Educ.*, vol. 2, no. 22, p. 5966, 2009.

[16] M. Suuroja, “Service Quality - Main Conceptualizations and Critique,” *SSRN Electron. J.*, vol. 6, no. 23, 2011, doi: 10.2139/ssrn.486947.

[17] E. Tolpa, “Measuring Customer Expectations of Service Quality : case Airline Industry Measuring Customer Expectations of Service Quality : case Airline Industry,” 2012.