Article

Intelligent modelling of learning patterns in tertiary education setting

Emmanuel Tuyishimire[1], Wadzanai Mabuto [2], Gatabazi Paul[2] and Sylvie Bayisingize [3]

[1] University of Cape Town; emmanuel.tuyishimire@uct.ac.za
[2] University of Johannesburg
[3] Mount Kenya University
* Correspondence: emmanuel.tuyishimire@uct.ac.za

Abstract: We are in the era where various processes need to be online. However, data from digital learning platforms are still underutilised in higher education, yet, they contain student learning patterns, whose awareness would contribute to educational development. This limits development of adaptive teaching and learning mechanisms. In this paper, a model for data exploitation to dynamically study students progress is proposed. Variables to determine current students progress are defined and are used to group students into different clusters. K-means clustering is performed on real data consisting of students from a South African tertiary institution. Cluster migration is analysed and the corresponding learning patterns are revealed.

Keywords: K-means; performance; pattern

1. Introduction

The 4th industrial revolution (4IR) has been changing existing means of production [1]. This is due to the fact that various technological models have surfaced. These include, for example, telecommunication models [2–4], advanced data collection and transportation models [5–13] together with various data and system analysis models [14–17]. Many more additional models have been developed and implemented to build various intelligent systems and this has drastically developed global economy.

The world is now in an era where intelligent systems are used to enable most activities. Additionally, 2020 marked a historical year when the COVID-19 pandemic imposed social distancing on the world, and this created a need for digitizing various processes for community interaction. This is changing the model of communication among people, yet various communication means need to be facilitated by various mediating digital mechanisms.

Measures to respond to such novel and critical conditions have begun to surface. For example, the United Nations Educational, Scientific and Cultural Organisation (UNESCO) launched an initiative to (i) help countries in mobilizing resources and implementing innovative and context-appropriate solutions to provide education remotely, leveraging hi-tech, low-tech and no-tech approaches; (ii) seek equitable solutions and universal access; (iii) ensure coordinated responses and avoid overlapping efforts; and (iv) facilitate the return of students to school when they re-open to avoid an upsurge in dropout rates [18]. This indicates that the educational systems need to be re-visited to adopt the critical conditions created in a technology-driven era.

This loud call for changes in educational systems need to be supplemented with advanced students evaluation models. It is important to periodically and timely assess students performance in order for adjusting teaching and learning strategies.

Models for students performance evaluation have been done in various settings. For nursing students, performance evaluation has been done in [19]. This has been done by following competency-based education approach [20], i.e. students are assessed

doi:10.20944/preprints202109.0392.v1

based on real-world professional performance. In this study, it has been shown that this model of assessing students outperforms other traditional models of assessments such as Grade Point Average (GPA) referred to as the most popular traditional quantitative indicator to assess academic performance. However, this model ranks students in terms of how they could professionally perform and does not predict how students would perform in further studies. This model does not provide room for instructor to mitigate poor performances which might impact on future learning processes.

To make more efficient the teaching-learning processes, a model to determine the relationship between teacher performance scores and student achievement has been proposed in [21]. Here, It has been found that there were a significant correlation between teacher's performance and the whole class performance. However, this does not determine those critical students which might need additional assistance.

On the other hand, a student centred model for evaluating student performance in laboratory applications has been proposed using fuzzy logic [22]. This Set Theory approach has been found to outperform classical models of performance evaluation,i.e. the models for performance evaluation based on exam results which is evaluated only as success or failure. Performance levels/compartments have been extended from two (success or failure) to 5 (Very Unsuccessful, Unsuccessful, Average, Successful, Very Successful). Given the performance of a student based on a list of assessments, there set of logic rules ([23]) to determine which current level a student may be ranked in. However this model of performance evaluation, requires a consensus of all involved educators on the rules to be adopted. Beside, this qualitative way of ranking students does not show how much they have improved or lowered. Furthermore, this model use only one performance score to rank students and this might provide confusing insights on the improvement or lowering of students.

Furthermore, the factors related to student performance in a distance-learning setting have been evaluated in [24**?**], for business communication course, and for Medical students, the factors have been revealed in [25]. For private colleges, performance factors have been shown in [26]. This kind of factors are Courses related and may not necessarily be applicable to generally study student performance in any subject.

There have been several data mining models to predict students performance [27–32]. However none of them focuses on students continuous development.

This paper focuses on the application of intelligent systems for eduction development. The motivation comes from the fact that poor understanding of student learning patterns usually limits development of adaptive teaching and learning processes. It is known that keeping university students consistently engaged with their academic studies and taking ownership of their learning, is a widely recognised challenge for most educators [33]. The deficiencies in learner behaviour, account for a significant portion of academic failure; yet there has been limited research in mining data useful for understanding student learning patterns which can be used to predict academic performance or to optimise the teaching process. Without relying on objective evidence of student learning behaviours, educators are unable to differentiate between challenges that relate to the larger class from those specific to individual students. Therefore, some students are at the risk of being left behind or being overlooked in the learning process. This is particularly relevant in the context of traditionally disadvantaged institutions, which enrol students of wide-ranging aptitude; these students are at increased risk of sub-optimal academic achievement in the absence of evidence-informed adaptive learning and teaching processes.

On the other hand, data from digital learning platforms and Machine Learning methods are underutilised in higher education. It is known that, the advent of machine-learning methods and the progressive use of digital learning platforms at institutions of higher learning, has created opportunities for educators to understand and modify the learning behaviours of students. However, myths around the complexity of machine learning and slow adoption of teaching technologies, often result in missed opportunities to improve efficiencies in teaching and learning processes.

In this paper, a model for studying the patterns of students performance is proposed. This is achieved by analysing each student 's marks distribution and calculate related distribution parameters, on basis of which a dynamic k-means clustering is performed. The current performance lever of students in the same cluster is determined by the underlying cluster heads. Inter-clusters migration is analysed to evaluate students improvement or lowering.

It is to the best of our knowledge that this is a first attempt to study dynamic students patterns and hence this model is not statistically compared with any other. The model is rather evaluated using real data from Second year students from a South African tertiary institution.

The rest of this papers is organised as follows. The proposed performance model is proposed in Section 1 and related experiment results are discussed in Section 3. Lastly the paper is Concluded in Section 4.

2. Proposed model for students performance evaluation

The proposed model for learning performance patterns is described in Figure 1.

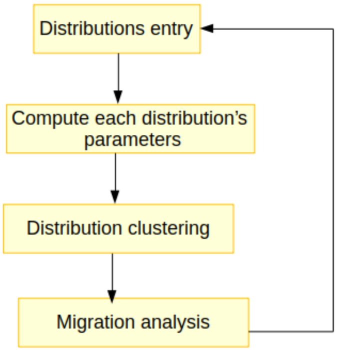


Figure 1. The proposed model.

The process consist of repetitively start from the top to the bottom as described in Figure 1. The steps are described as follows.

Distribution entry

Each student may be described as the schema *Student* below. In fact, a student is determined by his/her student number n and his sequence of marks obtained in his previous quizzes if any.

Student

$n : ID$

$Q : Seq\mathbb{R}^{+}$

$\#Q \in \mathbb{Z}^{+}$

Note that each student's number might reflect more static details such as demographic or any other recorded data which might bring insight on how the students conducts him/herself in the whole learning process. Each time a new quiz/test is given, the schema above my be updated as follows.

Update
$\Delta Student$
$q? : \mathbb{Z}^+$

$Q' = Q ^\frown \langle q? \rangle$
$n' = n$

Keeping the student's number, the student's quizzes sequence is updated by adding (concatenating) the new quiz marks.

Compute each distribution parameters

The performance evolution for each student may be determined by some parameters of his/her marks distribution. In this paper we choose two linearly independent parameters namely :

- Current mean: this is to determine the expected mark for the student and this can be used to determine whether a student may have a passing or fail marks.
- Current Standard deviation: it would determine how far the student's mark may be different from hist current average marks. This would help in showing improving or lowering students.

Distribution clustering

Now that we have two independent parameters, a distance function may be defined on them. Here, we considered the Euclidean distance. i.e. given that $s_1(\lambda_1, \sigma_1)$ and $s_2(\lambda_2, \sigma_2)$ two students and their respective current averages λ_1 and λ_2; and current standard deviations σ_1 and σ_2. The considered distance is expressed as follows.

$$d(s_1, s_2) = \sqrt{(\lambda_2 - \lambda_1)^2 + (\sigma_2 - \sigma_1)^2} \tag{1}$$

The distance function shown in Equation 1 expresses the difference between two student's performances. It can then be used to group students based on how their perform. This setting (two independent parameters and a distance function) allow the use of K-means algorithm [34], the famous clustering model for this context. It is recognised that in common practices the optimum number k of clusters is first computed but this is beyond this paper's work.

Migration analysis

After students clustering, it is important to determine whether any two students are differently performing (students in the different clusters) or are performing in the same way (students in the same cluster). Moreover, based on cluster heads, two clusters may be compared in terms of underlying students' performance.

Each time a new quiz is given to student the above mentioned process is repeated. The new clustering can then be compared with the previous clustering (if any). New marks entry cleary change the new values of parameters. Consequently, some students would eventually change their level of performance and hence clusters (migration). His current cluster and previous cluster may be compared to determine whether a student is improving or not.

3. Experimental results

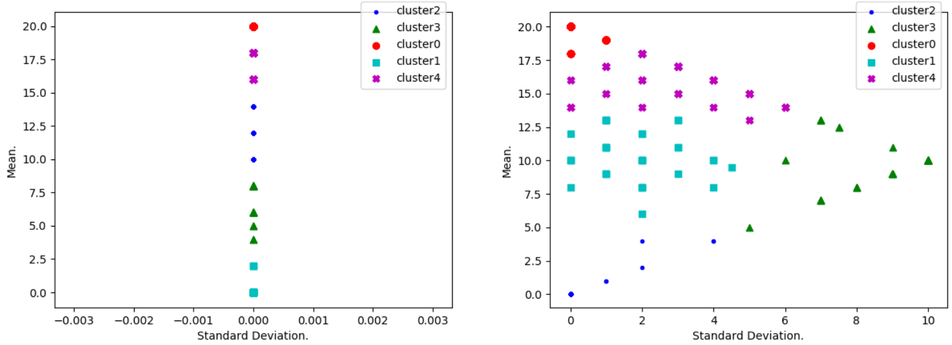
The proposed model has been applied to real data from second year students at the University of Johannesburg, whose faculty and department are omitted in this paper. We consider a class of 703 students. Marks for 9 consecutive quizzes (out of 20 marks each) have been recorded for each student. We compute the mean and standard deviation for each student and these two parameters have been used as independent dimensions for clustering the students.

Students performance have been done in three major steps:

- **Performance evolution** A K-means algorithm has been used to group students according to their mean and standard deviation. Students with a highest mean and least standard deviation are considered to be consistently motivated, those with a lowest mean and least standard deviation are the ones who are consistently discouraged. High standard deviation shows that the corresponding student have diverse marks due to either encouragement or discouragement.
- **Performance distribution.** At this stage, students have been grouped in five performance clusters. For each of the clusters Qualification codes have been interpreted. Theres are B1CEMQ, B3A17Q, B3AE7Q, B3F17Q, BC1413, BCG014, BCGE14 and None.
- **Consistency.** We study the relevance of the quizzes-based performance and succeeding in written tests, for fully engaged students. Test 1 has been written after the first four quizzes and Test 2 after the next/last four. Here, the coefficient of variation has been employed to measure each student's performance based to a series of marks got in the considered quizzes.

3.1. Performance evolution

After each quiz, 5 clusters have been computed. The following figures show how students have been dynamically grouped in performance clusters.



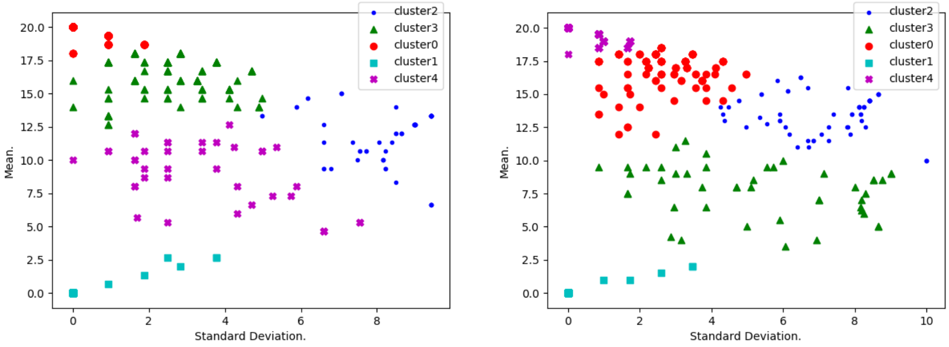
(a) Clustering after Quiz 1.

(b) Clustering after Quiz 2.

Figure 2. Clustering after Quiz 1 and 2.

Figure 2a shows that after the first quiz, all 703 students have one of 12 different shown marks. The figure shows that only one mark indicates a highest performance level and two marks indicate a least performance level. As it is the very first quiz, the standard deviation for each student is zero.

Figure 2b reveals a case where the standard deviation may differ from zero. It shows an instance where a student could deviate from his/her previous marks (his/her current mean). This high deviation may be caused to the fact that the student(s) might have missed the previous quiz (and thus has 0/20) and when s/he showed up for the second quiz, s/he gets very high mark. Apart from this, the student might be highly encouraged or discouraged after the first quiz. In this later case, such students would be assisted by commenting on their performance.

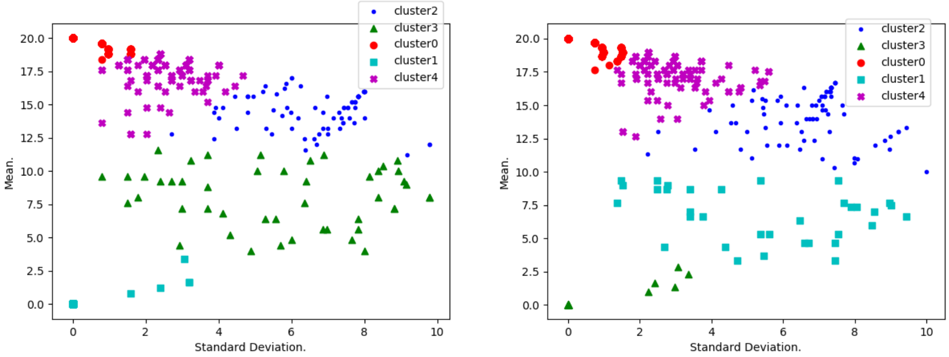


(a) Clustering after Quiz 3.

(b) Clustering after Quiz 4.

Figure 3. Clustering after Quiz 3 and 4.

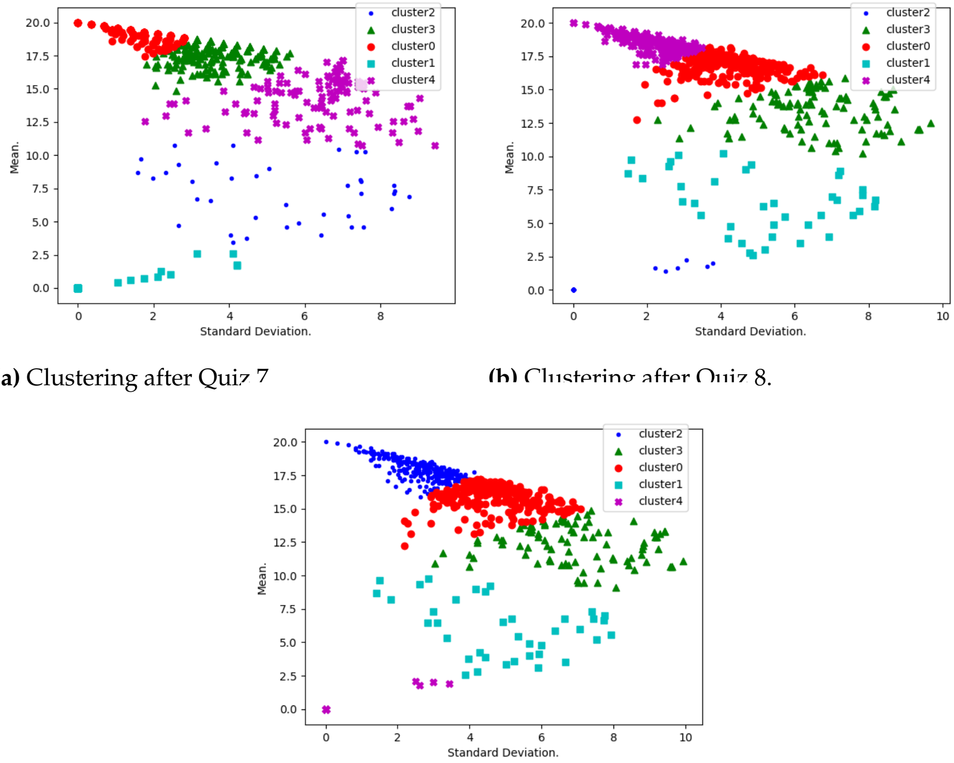
Figures 3a and **??** show the performance statuses after the third and fourth quizzes, respectively. They both show that students statuses get more diverse as the number of quizzes increases (more data points). This is happening in each cluster except the cluster of under performing students (Cluster 1 for both figures). Students in this cluster are the most discouraged and need some encouragement.



(a) Clustering after Quiz 5.

(b) Clustering after Quiz 6.

Figure 4. Clustering after Quiz 5 and 6.



(a) Clustering after Quiz 7

(b) Clustering after Quiz 8.

(c) Clustering after Quiz 9.

Figure 5. Clustering after Quiz 7, 8, and 9.

Figures 4a, 4b, 5a, 5b and 5c show the same trend as Figures 3a and 3b as discussed above. It is important to note that the distribution of the data points after each quiz show a poor correlation between the mean and standard deviation, and this highlights the fact that the two variables are indeed independent. The least performing students are very isolated (the cluster closest to (0,0)) from other clusters and such students need to be subjected to some encouragement measures.

3.2. *Performance distribution*

After the 10^{th} quiz, the clustered distribution is represented in Figure 6

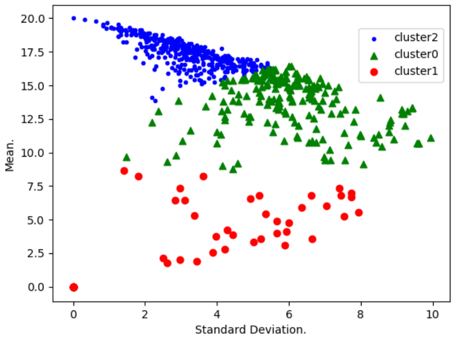
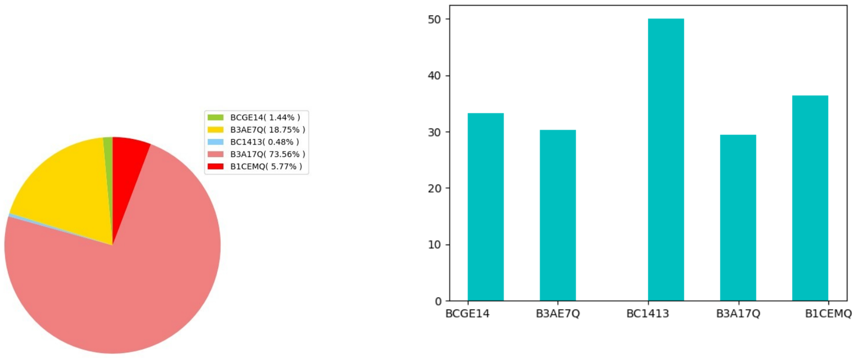


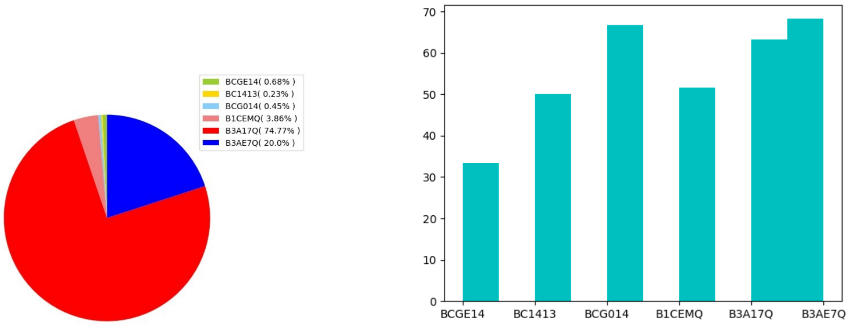
Figure 6. Clustering after Quiz 10.

We categorise 703 students in three categories/clusters (see Figure 6): More Encouraged Students (MESs) refereed as Cluster2, Encouraged Students (ESs) refereed as Cluster0 and discouraged or Less Encouraged Students (LESs) refereed as Cluster1. We study the clusters based distribution of qualification codes and this is complemented by the study of proportional distribution to study the expectedness of code distributions. Qualification codes



(b) Proportional distribution of qualification
(a) Qualification code distribution for Cluster2. code for Cluster2.

Figure **??** shows that the majority of MESs are the ones whose qualification code is B3A17Q and the minority corresponds to BC1413. However 50*percentageonFigure***??***)ofBC*1413*stude*
the cluster. This means BC1413 students are the most expectedly more encouraged. This is due to the fact that the total number of students whose qualification code is BC1413, is relatively small.



(a) Cluster0 **(b)** Cluster 0

Figure 8. Qualification codes for Cluster 1

Figure **??** shows that 74.77% of ESs are B3A17Q , 20% B2AE7Q and clearly these code qualifications represent the majority of Cluster0. On the other hand Figure **??** shows that BCG014 and B3AE7Q students are the most expectedly encouraged.

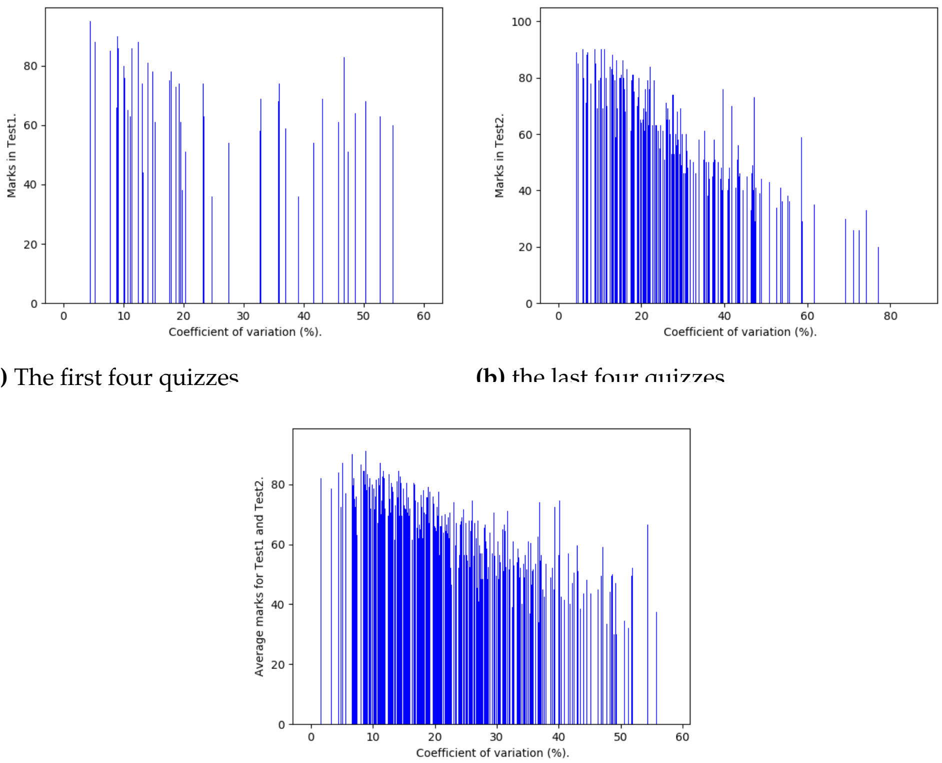
Figure **??** shows that most discouraged students are B3A17Q and Figure **??**

shows that all students corresponding to that qualification code are least encouraged.

3.3. Consistency

Figure 9a shows the correspondence between the performance in the first four quizzes and the first test. The correlation coefficient has been calculated to be $r = -0.03$ which is small. This means that good performing in the first four quizzes would not significantly imply good performance in Test1.

However, Figure 9b shows that the performance in the last four tests is highly correlated $(r = -0.84)$ with the performance in Test2. This is why Figure 9c shows that the overall performance in all quizzes has high correlation with the performance in the average of Test1 and Test2 (r=-0.75).



(a) The first four quizzes

(b) the last four quizzes

(c) Performance after all the quizzes.

It is important to note that for the first four quizzes, students have same sequence of marks. This shown by the fact 579 students are distributed in only 59 different sequences of marks (coefficient of variations). This number is significantly small and the reason behind might be a potential justification of the poor correlation between the first quizzes and test. For the last four quizzes 399 different values of coefficient of variation have been found, and this shows that students are expected to have significantly different sequence of marks.

4. Conclusion

This paper has proposed a novel model for evaluating students performance. This has been achieved by introducing the performance distance which is measured using the mean and standard deviation of each student's marks distribution. This has been accompanied by using the K-means clustering to group students in performance groups and late the cluster migration has been analysed. This model has been experimented on real data collected from the University of Johannesburg.

References

1. Schwab, K. *The fourth industrial revolution*; Currency, 2017.
2. Tuyishimire, E. Routing in Mobile Networks **2013**.
3. Tuyishimire, E. Internet of things: least interference beaconing algorithms. Master's thesis, University of Cape Town, 2014.
4. Tuyishimire, E.; Bagula, B.A. A novel management model for dynamic sensor networks using diffusion sets. 2020 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2020, pp. 1–6.
5. **Emmanuel Tuyishimire**.; Adiel, I.; Rekhis, S.; Bagula, B.A.; Boudriga, N. Internet of Things in Motion: A Cooperative Data Muling Model Under Revisit Constraints. Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences. IEEE, 2016, pp. 1123–1130.
6. Bagula, A.; **Emmanuel Tuyishimire**.; Wadepoel, J.; Boudriga, N.; Rekhis, S. Internet-of-Things in Motion: A Cooperative Data Muling Model for Public Safety. Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scal-

able Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences. IEEE, 2016, pp. 17–24.

7. **Emmanuel Tuyishimire**.; Bagula, A.; Rekhis, S.; Boudriga, N. Cooperative Data Muffling from Ground Sensors to Base Stations Using UAVs **2017**.
8. Ismail, A.; Bagula, B.A.; Tuyishimire, E. Internet-of-things in motion: A uav coalition model for remote sensing in smart cities. *Sensors* **2018**, *18*, 2184.
9. Tuyishimire, E.; Bagula, B.A.; Ismail, A. Optimal clustering for efficient data muling in the internet-of-things in motion. International Symposium on Ubiquitous Networking. Springer, 2018, pp. 359–371.
10. Ismail, A.; Tuyishimire, E.; Bagula, A. Generating dubins path for fixed wing uavs in search missions. International Symposium on Ubiquitous Networking. Springer, 2018, pp. 347–358.
11. Mauwa, H.; Bagula, A.; Tuyishimire, E.; Ngqondi, T. An optimal spectrum allocation strategy for dynamic spectrum markets. 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2019, pp. 1–8.
12. Tuyishimire, E. Cooperative data muling using a team of unmanned aerial vehicles **2019**.
13. Mauwa, H.; Bagula, A.; Tuyishimire, E.; Ngqondi, T. Community healthcare mesh network engineering in white space frequencies. 2019 ITU Kaleidoscope: ICT for Health: Networks, Standards and Innovation (ITU K). IEEE, 2019, pp. 1–8.
14. Antoine, B.; Emmanuel, T.; Olasupo, A. Cyber physical systems (cps) surveillance using an epidemic model. *arXiv preprint arXiv:1912.07479* **2019**.
15. Tuyishimire, E.; Bagula, A.; Ismail, A. Clustered data muling in the internet of things in motion. *Sensors* **2019**, *19*, 484.
16. Tuyishimire, E.; Bagula, B.A. Modelling and analysis of interference diffusion in the internet of things: an epidemic model. 2020 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2020, pp. 1–6.
17. Tuyishimire, E.; Bagula, B.A. A Formal and Efficient Routing Model for Persistent Traffics in the Internet of Things. 2020 Conference on Information Communications Technology and Society (ICTAS). IEEE, 2020, pp. 1–6.
18. Mohamed, A. UNESCO rallies international organizations, civil society and private sector partners in a broad coalition to ensure# LearningNeverStops, 2020.
19. Fan, J.Y.; Wang, Y.H.; Chao, L.F.; Jane, S.W.; Hsu, L.L. Performance evaluation of nursing students following competency-based education. *Nurse Education Today* **2015**, *35*, 97–103. doi:https://doi.org/10.1016/j.nedt.2014.07.002.
20. Anema, M.; McCoy, J. *Competency based nursing education: guide to achieving outstanding learner outcomes*; springer publishing company, 2009.
21. Milanowski, A. The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *peabody Journal of Education* **2004**, *79*, 33–53.
22. Gokmen, G.; Akinci, T.Ç.; Tektaş, M.; Onat, N.; Kocyigit, G.; Tektaş, N. Evaluation of student performance in laboratory applications using fuzzy logic. *Procedia-Social and Behavioral Sciences* **2010**, *2*, 902–909.
23. Yen, J.; Langari, R.; Zadeh, L.A. *Industrial applications of fuzzy logic and intelligent systems*; IEEE press, 1995.
24. Cheung, L.L.; Kan, A.C. Evaluation of factors related to student performance in a distance-learning business communication course. *Journal of Education for Business* **2002**, *77*, 257–263.
25. Pulito, A.R.; Donnelly, M.B.; Plymale, M. Factors in faculty evaluation of medical students' performance. *Medical Education* **2007**, *41*, 667–675.
26. Mortada, L.; Bolbol, J.; Kadry, S. Factors Affecting Students' Performance a Case of Private Colleges in Lebanon. *J Math Stat Anal* **2018**, *1*, 105.
27. Cortez, P.; Silva, A.M.G. Using data mining to predict secondary school student performance **2008**.
28. Osmanbegovic, E.; Suljic, M. Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business* **2012**, *10*, 3–12.
29. Kabakchieva, D. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies* **2013**, *13*, 61–72.
30. Ramesh, V.; Parkavi, P.; Ramar, K. Predicting student performance: a statistical and data mining approach. *International journal of computer applications* **2013**, *63*.
31. Kabakchieva, D. Student performance prediction by using data mining classification algorithms. *International journal of computer science and management research* **2012**, *1*, 686–690.
32. Mengash, H.A. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access* **2020**, *8*, 55462–55470.
33. Wanner, T.; Palmer, E. Personalising learning: Exploring student and teacher perceptions about flexible learning and assessment in a flipped university course. *Computers & Education* **2015**, *88*, 354–369.
34. Likas, A.; Vlassis, N.; Verbeek, J.J. The global k-means clustering algorithm. *Pattern recognition* **2003**, *36*, 451–461.