Journal of Engineering Science and Technology
Special Issue on ICIT2022 Vol. 19, No. 1 (2024) 131 - 139


# AN EARLY WARNING SYSTEM FOR STUDENTS AT RISK USING SUPERVISED MACHINE LEARNING

YAM ZHENG HONG[1], MOHD NORSHAHRIEL ABD RANI[1], NABILAH FILZAH MOHD RADZUAN[2], LIM HUAY YEN[1], SARASVATHI NAGALINGAM[1]

[1]Faculty of Data Science and Information Technology (FDSIT), INTI International University, Persiaran Perdana BBN, Putra Nilai, 71800 Nilai, Negeri Sembilan
[2]Faculty of Computing, Universiti Malaysia Pahang, Pekan 26600 Pahang
Corresponding Author: mnorshahriel.arani@newinti.edu.my

## Abstract

The Covid-19 pandemic has brought numerous social issues to the fore, including poverty alleviation and education. As a result, significant changes have occurred in education, in which teaching is done remotely. To keep up with the course's pace, students must be self-motivated, well-organized, and have time management skills. Without these behaviours, online education is inappropriate for students. According to the research, 52% of students who sign up for a course would never read the course materials. Furthermore, throughout the course of five years, the dropout rate reached a stunning 96%. The main objective of this study is to create an early warning system for educators to use in educational institutions. For that, this study will perform a study of the current issues, factors, and solutions in education student's data, determine the supervised machine learning algorithms, compare which model is the best predict the students' performances, develop an early warning system for the educators to make an early decision in order to assist and consult the at-risk students, and finally conduct testing and evaluation of the system. Besides, this study also focuses on the Decision Tree, Random Forest, and XGBoost models in the system. The system will detect or forecast symptoms of dropout or potential dangers ahead of time, allowing educational institutions to anticipate problems and provide adequate educational services through appropriate intervention and response. A dashboard system will be created with a descriptive data mining model that can examine and make early judgments on at-risk students before they become high-risk. The web-based platform enables educational institutions to explore data patterns and analyse current crucial key performance metrics conditions. The technology will display the analysed results as well as visualized data to help educators gain insight and make better decisions.

Keywords: Education quality, Educational environment, Students risk, SVM, Warning system.

## 1. Introduction

Approximately 50,000 children drop out of school each year [1] Over a quarter of all high school students drop out before graduating. In some major cities, the rate is as high as 40%. Millions of minorities and poor students who are "at-risk" have been impacted by higher standards in public schools [2, 3] The rising number of at-risk middle and high school students on the verge of dropping out due to academic failure or other issues is a big worry in today's education [4]. Despite the high rate of at-risk students is high, researchers, practitioners, and policymakers have paid less attention to at-risk students. This is because at-risk students are usually considered as an unavoidable consequence of family relocation or residence mobility over which schools have little control. Gender, race and ethnicity, handicap status, and geographic location can all affect at-risk students. In a study, lack of attention from educational institutions is one of the biggest factors for at-risk students. Some students are introverted and shy. They are generally embarrassed to seek assistance when they face any problems. They struggle in quiet, hoping that a teacher will notice [5]. As a result, it is important to have a dashboard to help educators analyse and intervene early in when students are struggling. The dashboard can reduce the student dropout rate and reduce the student at-risk.

Many operations throughout the world have been impacted by the coronavirus illness 2019, often known as Covid-19. Most economic operations were put on hold until the pandemic was controlled as a result of the entire or partial lockdown enforced by countries all around the world. Authorities in each impacted country make judgments based on a variety of variables, including the country's financial viability. Following an upsurge in the number of positive COVID-19 cases, Malaysia's prime minister issued a movement control order (MCO) on March 16, 2020. The MCO had a significant influence on the economy, education, politics, and the lifestyle of residents [6-8]. As the COVID-19 pandemic continues to spread, parents and children are realizing that traditional classroom instruction is no longer necessary. Some could even go as far as to argue that the typical classroom is more of a hindrance than a help when it comes to learning. Their lifestyles have been perfectly linked with technology.

Gen Alpha is a group of people born between 2010 and 2025. They were born in a century when advanced technologies were available 24/7 around the world. To them, technology is everything. Their lives revolve around technology, from entertainment to gaming, networking with friends, and even schooling in the aftermath of the COVID-19 epidemic. [9, 10]. Social media is a way of life for them, and technology is merely an extension of their own consciousness and personality. According to a survey published by Dell Technologies in 2017, 85 percent of the professions that Generations Z and Alpha will begin in 2030 have yet to be developed. 65 percent of today's elementary school students will work in jobs that do not yet exist [11, 12] This generation of young pre-schoolers also has the most non-traditional family arrangements. While many Alpha students may be unaware of COVID-19's influence on their education, they will undoubtedly experience it for years to come. Given the negative impact of the COVID-19 pandemic on pedagogical methods and student learning, reform is critical to guarantee that educators present the curriculum in ways that are relevant to students in Generations Z, Alpha, and beyond [13-14].

The majority of students in today's educational institutions come from Generation Z, which has grown up in a fully world community and is closely intertwined with technology. This generation, the oldest of whom is currently 25 years old, must consider their education in the context of a genuinely worldwide epidemic, with many students facing missed examinations, athletic events, and even graduation ceremonies [15]. Generation Z students are used to receiving immediate feedback and communication via programmed such as Facebook Messenger, WhatsApp, and WeChat. But they also recognize the value of working together to address the world's most pressing issues: their agenda includes not just COVID-19, but also climate change and mental health problems (World Economic Forum 2020). In the COVID-19 epidemic, the characteristics of generation Z learners will undoubtedly have an impact on education. In reality, many of generation Z's inherent talents and preferences, as outlined above, may contribute to beneficial developments in this otherwise challenging period [16] . Students have been transferred to nearly all online study, which corresponds to their desire for technology integration in learning environments.

Education is the controlled transmission of socially meaningful experience from past generations to future generations. The process of transmitting and receiving was defined as education. The most common way to obtain an education is to enrol in an educational institution's training programme [17, 18]. Education for growth, education as direction, and education as preparation for adult responsibilities are the three purposes of education towards individuals [19, 20]. As a result of the COVID-19 outbreak, the majority of nations have undertaken lockdown and social separation measures, resulting in the closure of schools, training institutes, and institutions of higher education. Educators are providing quality instruction through a variety of online media, which represents a paradigm shift [21]. Despite the difficulties educators and learners confront, online education, distance learning, and continuing education have shown to be effective in combating this unprecedented global epidemic.

Throughout the epidemic, e-learning platforms played a key role in supporting schools and universities in promoting student learning while colleges and schools were closed [22]. While adapting to the new alterations, staff and student preparation must be monitored and supported. As a result of increased and unstructured time spent online, students are at greater risk of exposure to potentially harmful and violent content and cyberbullying. As a result of school closures and strict containment measures, more families are relying on technology and digital solutions to keep their children engaged in learning, entertained, and connected to the outside world, but not all students have the necessary knowledge, skills, and resources to stay safe online [23]

An at-risk student refers to students who have a high probability of flunking a class or dropping out of their school. A large proportion of dropout students are still roaming the streets, unable to receive ideal education or re-education. The society and government should carry out the responsibilities by reducing the number of students from dropping out and providing suitable educational services to at-risk students. At-risk students face a variety of challenges compare to other students. Students with low socioeconomic level, particularly boys, experience emotions of isolation and alienation in their schools [24, 25].

Every year, over 50,000 youngsters drop out of school [26, 27]. Over a quarter of high school students do not complete their education. The percentage might be as high as 40% in some urban cities. Higher standards in public schools have impacted millions of minorities and poor students who are "at-risk" [27, 28].

For the early identification of at-risk students, a well-established system is needed, as well as diagnostic instruments that can detect or forecast symptoms of dropout or possible dangers. Implementing an early warning system to identify at-risk students from dropping out or who may be at risk will allow education institutions to anticipate signals of difficulty and provide appropriate assistance through appropriate intervention and response, ensuring systemic management of at-risk students [29, 30].

Early detection of potential dropouts has been discovered by researchers [31, 32]. At-risk students are the students who live in a school district that is in flux, have a low-income family member, have low academic skills, have parents who did not complete high school, have negative self-perceptions, and poor self-esteem are the characteristics of at-risk students [33, 34].

## 2. Methodology

In order to make educated judgments and take action in the event of a disaster, early warning systems provide individuals with relevant and timely information in a systematic manner prior to the event [35, 36]. It can be seen as the next step in the process of predicting how well students do in school [37]. This research focuses on supervised machine learning and employs the best accurate algorithm to provide a web-based dashboard [38]. The Decision Tree, Random Forest, and XGBoost algorithms will be compared [39, 40]. The process of this research will be implemented through the CRISP-DM methodology [41, 42]. The CRISP-DM methodology consists of six stages which are business understanding, data understanding, data preparation, modelling, evaluation, and deployment. It teaches you how to plan a data mining project in a systematic way [43, 44].

## 3. Results and Discussion

Results show that by analysing Figs. 1 and 2, the research can compare and conclude that the random forest has the highest accuracy compared to the decision tree and gradient boosting [45, 46]. Four different methods have been used which are accuracy score, cross-validation, confusion matrix, and classification report [47]. Those methods show that random forest has the highest accuracy percentage. Thus, the research will deploy random forest as the ML model into the student at-risk prediction system [48, 49].

Besides, in the dashboard interface, the researcher can analyse the student's information such as name, gender, average score, GPA, attendance, and risk. Besides, there are some graphs such as pie charts, and bar charts to visualize student performance. The performance includes exam results, count by gender and grade, average subject score, etc. The dashboard is only one page this is because too much info will make it messy and not user-friendly as shown in Fig. 3.

```
print('----------------------Models Score---------------------------\n')
print('Decision Tree Score :     ', Dtree.score(x_test, y_test.values.ravel()))
print('Gradient Boosting Score : ', GBC.score(x_test, y_test.values.ravel()))
print('Random Forest Score :     ', RF.score(x_test, y_test.values.ravel()))
print('\n')
print('----------------------Cross Validation Mean Score-----------------------------\n')
print('CV mean score: ', Dtree_cv_score.mean())
print('CV mean score: ', GBC_cv_score.mean())
print('CV mean score: ', RF_cv_score.mean())
print('\n')
print('----------------------Confusion_matrix-----------------------------\n')
print('Decision Tree confusion_matrix : \n', confusion_matrix(y_test, Dtree_ypred, labels=[0,1]))
print('Gradeint Boosting confusion_matrix : \n', confusion_matrix(y_test, GB_ypred, labels=[0,1]))
print('Random Forest confusion_matrix : \n', confusion_matrix(y_test, RF_ypred, labels=[0,1]))
print('\n')
print('----------------------Classification Report-----------------------------\n')
print('Decision Tree Report \n', classification_report(y_test, Dtree_ypred, target_names=['fail','pass']))
print('Gradeint Boosting Report \n', classification_report(y_test, GB_ypred, target_names=['fail','pass']))
print('Random Forest Report \n', classification_report(y_test, RF_ypred, target_names=['fail','pass']))
```

```
----------------------Models Score---------------------------

Decision Tree Score :      0.8652389413635089
Gradient Boosting Score :  0.8315720564855513
Random Forest Score :      0.9454471772810873


----------------------Cross Validation Mean Score-----------------------------

CV mean score:  0.8618127497861549
CV mean score:  0.826317740166235
CV mean score:  0.9509497800912637


----------------------Confusion_matrix-----------------------------

Decision Tree confusion_matrix :
 [[ 5009   645]
 [ 3678 22747]]
Gradeint Boosting confusion_matrix :
 [[ 4655   999]
 [ 4404 22021]]
Random Forest confusion_matrix :
 [[ 5529   125]
 [ 1625 24800]]
```

**Fig. 1. Model comparison 1.**

```
----------------------Classification Report-----------------------------

Decision Tree Report
              precision    recall  f1-score   support

        fail       0.58      0.89      0.70      5654
        pass       0.97      0.86      0.91     26425

    accuracy                           0.87     32079
   macro avg       0.77      0.87      0.81     32079
weighted avg       0.90      0.87      0.88     32079

Gradeint Boosting Report
              precision    recall  f1-score   support

        fail       0.51      0.82      0.63      5654
        pass       0.96      0.83      0.89     26425

    accuracy                           0.83     32079
   macro avg       0.74      0.83      0.76     32079
weighted avg       0.88      0.83      0.85     32079

Random Forest Report
              precision    recall  f1-score   support

        fail       0.77      0.98      0.86      5654
        pass       0.99      0.94      0.97     26425

    accuracy                           0.95     32079
   macro avg       0.88      0.96      0.91     32079
weighted avg       0.96      0.95      0.95     32079
```

**Fig. 2. Model comparison 2.**

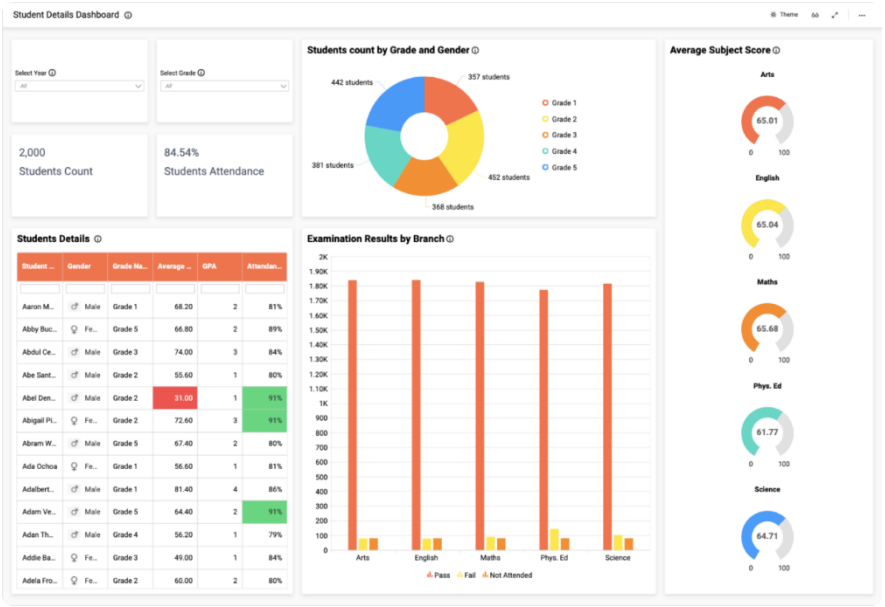


**Fig. 3. Dashboard interface.**

## 4. Conclusions

The random forest has the highest accuracy compared to the decision tree and gradient boosting. The dashboard and model can be deployed into a mobile application or website. The dataset can be uploaded into the cloud server and use the API to pass the data into the Random Forest model. By deploying the dashboard into a mobile application or website, the user can access the dashboard without using the Power BI Desktop software using the local computer. Besides, uploading the dataset into cloud storage can prevent the student's dataset lost. Cloud storage also allows the user to extract and load the dataset more safely and conveniently. Besides, the system can record the system usage and provide feedback for further enhancement.

## Acknowledgments

Financial support for this research was provided by INTI International University. Thanks, are also due to all authors on preparing the manuscript.

## References

1. Akçapınar, G.; Altun, A.; and Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. *International Journal of Educational Technology in Higher Education*, 16, 40.
2. Akçapınar, G.; Coşgun, E.; and Altun, A. (2013). Mining Wiki usage data for predicting final grades of students. *Proceedings of the International Academic Conference on Education, Teaching and E-learning in Prague* 2013 (*IAC-ETeL* 2013), Prague, Czech Republic, 1-6.
3. Akçapınar, G.; Hasnine, M.N.; Majumdar, R.; Flanagan, B. and Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using eBook interaction logs. *Smart Learning Environments*, 6(1), 4.

4. Alzubi, J.; Nayyar, A.; and Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series*, 1142, 012012.
5. Analytics Vidhya (2021). Random forest: Introduction to random forest algorithm. Retrieved 28 March 2022 from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
6. Arora., A.; and Jha., A.K. (2020). Understanding pattern of online gaming addiction among Indian teenagers. *Our Heritage*, 68(1), 13190-13100.
7. Baker, R.S.J.D. (2010). Data mining for education. In McGaw, B.; Peterson, P.; and Baker, E. (Eds.) *International encyclopedia of education* (3rd ed.). Oxford, UK: Elsevier.
8. McLeod, S. (2018). Questionnaire: Definition, examples, design and types. Retrieved 13 Apr. 2022 from https://www.simplypsychology.org/questionnaires.html
9. Bienkowski, M.; Feng, M.; and Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Center for Technology in Learning. SRI International, U.S. Department of Education.
10. Busetto, L.; Wick, W.; and Gumbinger, C. (2020). How to use and assess qualitative research methods. *Neurological Research and Practice*, 2, 14.
11. Business Research Methodology (BRM) (2010). Interviews - Research-methodology. Retrieved 14 Apr. 2022 from https://research-methodology.net/research-methods/qualitative-research/interviews/
12. Chen, T.; and He, T. (2022). xgboost: eXtreme gradient boosting. Retrieved 30 Mar. 2022 from https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf
13. Chung, J.Y.; and Lee, S., (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*. 96, 346-353.
14. Dayan, P. (1999). *Unsupervised learning*. In Wilson, R.A.; and Keil, F.C. (Eds.) *The MIT encyclopedia of the cognitive sciences*. The MIT Press, 857-859.
15. Donnelly, M. (1987). At-risk students. Service Learning General, 305.
16. d'Orville, H. (2020). COVID-19 causes unprecedented educational disruption: Is there a road towards a new normal. *Prospects*. 49(1), 11-15.
17. Druian, G.; and Butler, J. (1987). Effective schooling practices and at-risk youth: What the research shows. School Improvement Research Series Retrieved March 23. 2022 from https://educationnorthwest.org/sites/default/files/effective-schooling-practices.pdf.
18. SlideShare (2014). Fact - Finding techniques. Retrieved April 12. 2022, from https://www.slideshare.net/gomzy22/fact-finding-techniques.
19. Hughes, J. (2020). Getting to know generation Alpha: 10 takeaways for higher ed. Retrieved 10 Feb. 2022 from https://www.keg.com/news/getting-to-know-generation-alpha-10-takeaways-for-higher-ed.
20. KEDI (2018). *Statistical yearbook of education*. Department of Education, Korean Educational Development Institute.

21. lluminate Education (2016). Create an early warning system and give your teachers x-ray vision! Retrieved March 31. 2022 from https://www.illuminateed.com/blog/2016/01/give-your-teachers-x-ray-vision/
22. Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms*. John Wiley & Sons.
23. Arnold, K. (2010). Signals: Applying academic analytics. *Educause Review*. Retrieved 18 Jan. 2022 from https://er.educause.edu/articles/ 2010/3/signals-applying-academic-analytics
24. Knowles, J.E. (2015). Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.
25. Li, C.; and Lalani, F. (2020). The COVID-19 pandemic has changed education forever. This is how. Retrieved February 8. 2022 from https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/
26. Liu, Y.; Wang, Y.; and Zhang, J., (2012). New machine learning algorithm: Random Forest. In: Liu, B.; Ma, M.; and Chang, J. (Eds.) *Information computing and applications*. *ICICA* 2012. Lecture Notes in Computer Science, vol 7473. Springer, Berlin, Heidelberg, 246-252.
27. Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research* (*IJSR*). 9, 381-386.
28. Maheswari Jaikumar. (2019). Functions Of Education. Retrieved 1 Feb. 2022 from https://www.slideshare.net/mahes warijaikumar/functions-of-education-177832869
29. Marshall, A.L.; and Wolanskyj-Spinner, A., (2020). June. COVID-19: challenges and opportunities for educators and generation Z learners. *Elsevier: Mayo Clinic Proceedings*, 95(6), 1135-1137.
30. McDill, E.L.; Natriello, G.; and Pallas, A.M. (1986). A population at risk: Potential consequences of tougher school standards for student dropouts. *American Journal of Education*, 94(2), 135-181.
31. McMillan, J.H.; and Reed, D.F. (1994). At-risk students and resiliency: factors contributing to academic success. *The Clearing House*, 67(3), 137-140.
32. Menhat, M., Mohd Zaideen, I.M., Yusuf, Y., Salleh, N.H.M., Zamri, M.A. and Jeevan, J. (2021). The impact of Covid-19 pandemic: A review on maritime sectors in Malaysia. *Ocean & Coastal Management*, 209, 105-638.
33. Natriello, G. (2022.). School dropouts - Extent of the problem, factors associated with early school leaving, dropout prevention programs and their effects. Retrieved February 5. 2022 from https://education.stateuniversity.com/pages/1921/Dropouts-School.html #ixzz7LgE1EqIQ
34. Naziev, A. (2017). What is an education? *Proceedings of the International Conference on The Future of Education*. Florence, Italy, 1-5.
35. Payne, D. (2019). How young universities aiming for high rankings are preparing for "Generation Alpha." Nature Index. Retrieved February 9. 2022 from https://www.natureindex.com/news-blog/how-young-universities-aiming-high-for-rankings-are-preparing-for-generation-alpha

36. Pereira, J. (2019). Pros and cons of online education. Retrieved February 9. 2022 from https://www.magzter.com/stories/Business/The-Observer-of-Management-Education/Pros-And-Cons-Of-Online-Education
37. Pokhrel, S.; and Chhetri, R., (2021). A literature review on impact of COVID-19 pandemic on teaching and learning. *Higher Education for the Future*, 8(1), 133-141.
38. QuestionPro .(2018). Quantitative data: Definition, types, analysis and examples. Retrieved February 14. 2022 from https://www.questionpro.com /blog/quantitative-data/
39. Reich, J.; and Ruipérez-Valiente, J.A. (2019). The MOOC pivot. *Science*, 363(6423), 130-131.
40. Rumberger, R.W.; and Larson, K.A., (1998). Student mobility and the increased risk of high school dropout. *American journal of Education*, 107(1), 1-35.
41. Shorten, A.; and Smith, J. (2017). Mixed methods research: expanding the evidence base. *Evidence-Based Nursing*, 20(3), 74-75.
42. Singh, M.; Sharma, S.; and Kaur, A. (2013). Performance analysis of decision trees. *International Journal of Computer Applications*, 71(19), 10-14.
43. Smerdon, B.A. (2002). Students' perceptions of membership in their high schools. *Sociology of Education*, 75(4), 287-305.
44. Smith, R.A. (2021). Pandemic and post-pandemic digital pedagogy in hospitality education for generations Z, alpha, and beyond. *Journal of Hospitality & Tourism Research*, 45(5), 915-919.
45. Sutton, R.S.; and Barto, A.G. (2018). *Reinforcement learning: An introduction*. (2nd ed.). *MIT press.*
46. Sruthi, E.R. (2021). Random forest: Introduction to random forest algorithm. Retrieved 28 Jan. 2022 from https://www.analyticsvidhya.com/blog/2021/06/ understanding-random-forest/
47. Subedi, S.; Nayaju, S.; Subedi, S.; Shah, S.K.; Shah, J.M. (2020). Impact of e-learning during COVID-19 pandemic among nursing students and teachers of Nepal. *International Journal of Science and Healthcare Research*, 5(3), 68-76.
48. UTA Libraries (2019). Subject and course guides: quantitative and qualitative research: Understand what qualitative research is. Retrieved 15 Feb. 2022 from https://libguides.uta.edu/quantitative_and_qualitative_research/qual
49. Morde, V. (2019). XGBoost algorithm: Long may she reign! Retrieved 30 Jan. 2022 from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.