10th International Scientific Conference
Technics, Informatics, and Education – TIE 2024
20-22 September 2024

Session: *Educational Technology*
Professional paper
DOI: 10.46793/TIE24.213M

Predictive Model for Early Detection of Students with Difficulties in Online Learning

Katarina Mitrović[1*] [0000-0002-4035-5987], Dijana Stojić[1] [0000-0003-0582-2911] and Mladen Janjić[1] [0000-0001-5324-0874]
[1]University of Kragujevac, Faculty of Technical Sciences Čačak, Serbia
[*]katarina.mitrovic@ftn.kg.ac.rs

Abstract: *Online learning has become increasingly prevalent in all education levels during recent years. While in highly developed regions transition from traditional to online learning happens without significant difficulties, in underdeveloped and developing countries introducing students to online learning is typically followed by complications and frustration. Many researchers conducted studies to solve the issue of conforming to online learning and provide equal opportunities to all students regardless of their demographical characteristics and environmental factors. Introducing artificial intelligence tools to this problem can provide valuable insight into patterns and predictors in online education. This study proposes a machine learning model for predicting the low-level student adaptability to online learning. This model can indicate students who might have difficulties adapting to online learning with 94% accuracy based on their demographical and environmental characteristics. The model is developed using locally weighted learning with a C4.5 decision tree classifier. This paper contributes to understanding the problems underlying online learning adaptability and offers an accurate tool for detecting students prone to online learning issues, which can help persons of authority provide dependable and rapid aid.*

Keywords: *adaptability; artificial intelligence; education; machine learning; online learning*

1. INTRODUCTION

Distance education has a long history [1], but online education emerged in the 20th century during the 80's and 90's with the emergence of the Internet and Web [2]. Besides some benefits, this also introduced difficulties and challenges [3, 4], but even traditional education faced major challenges of online learning due to the pandemic of COVID-19 [5, 6]. Besides many other issues, the transition to online learning accentuated the problem of predicting student failure [7]. This paper presents one approach to early detection of students with difficulties in online learning.

In recent years, artificial intelligence (AI) has become increasingly prevalent in diverse areas of educational research. Machine learning (ML), a subset of AI within the realm of computer science, typically employs statistical methods to enable computers to learn from data autonomously, without explicit programming [8]. This discipline has a broad spectrum of applications, excelling in pattern recognition and adaptive learning across diverse fields. A plethora of research has utilized ML algorithms to predict outcomes from new inputs and to uncover underlying data structures and relationships. Significant discoveries have been made in education-related studies through the application of ML techniques. Online learning has gained significant traction across all educational levels in recent years. In highly developed areas, the shift from traditional classroom settings to online learning generally occurs smoothly [5]. However, in underdeveloped and developing nations, the introduction of online learning often brings challenges and frustration [9]. Numerous researchers have conducted studies to address the difficulties of adapting to online education and to ensure equal opportunities for all students, irrespective of their demographic and environmental backgrounds, concluding that the following ML techniques outperformed any other: Deep Neural Networks [10], Decision Tree Algorithm [11], Weighted Voting Classifier [12], and Locally Weighted Learning model [9]. For instance, in [13] ML techniques were used to estimate students' performance in Blended Learning and Complete Virtual Courses, while in [14] authors developed eight ML algorithms for predicting students' performance in STEM courses with recommendations for using ML models in education. In [15], authors analyzed the feedback in online courses to improve the quality of learning.

Papers [10, 11, 12, 9] investigated the prediction of students' pass rates as one metric of learning success. Thus, it can be assumed that the integration of AI tools into this field can offer valuable insights into the patterns and predictors of successful online education.

This research introduces an ML model specifically designed to predict low student adaptability to online learning environments. The model demonstrates a 94% accuracy in identifying students who may struggle with adapting to online learning, based on an analysis of their demographic and environmental characteristics. The development of this model utilizes locally weighted learning (LWL) in conjunction with a C4.5 decision tree classifier. By offering deeper insights into the underlying issues of online learning adaptability, this paper contributes significantly to the field. Furthermore, it provides an accurate and practical tool for detecting students who are likely to encounter difficulties, thereby enabling authorities to deliver dependable and swift assistance. The model's implementation can also guide the development of targeted interventions and support systems, ultimately fostering a more inclusive and effective online learning experience for all students.

The remainder of the paper is organized as follows. The second section describes the methods used, with emphasis on participants and data acquisition, ML models, and software and hardware requirements. Results are specified in the third section, while the discussion is described in the fourth section. The final section brings concluding remarks.

2. METHODS

2.1. Participants and data acquisition

The dataset used in this paper was acquired by a group of researchers from Daffodil International University (Dhaka, Bangladesh) [16] and was made publicly available on the Kaggle dataset repository [17]. The data was collected from 1205 students using online and paper surveys and preprocessed into a form suitable for ML algorithm training. A detailed explanation of the preprocessing phase can be found in [8].

The dataset consists of 13 input variables that represent mainly the demographical and environmental characteristics of participants:

- Gender – Gender type of student,
- Age – Age range of the student,
- Educational level – Education institution level,
- Institution type – Education institution type,
- IT student – Studying as an information technologies student or not,
- Town – Is student located in town,
- Load-shedding – Level of reduction of electricity supply,
- Financial condition – Financial condition of the student's family,
- Internet type – Mostly used Internet type,
- Network type – Network connectivity type,
- Class duration – Daily class duration,
- Self LMS – Institution's Learning Management System (LMS) availability, and
- Device – Mostly used device for online learning.

Output variable shows whether a student has issues with online learning or not using two values:

- Yes – in the case when a student has problems adapting to online learning, and
- No – in the case when a student can adapt to online learning with the minor to no difficulties.

All variables are of nominal type and have no missing values. Generalization of the output variable is performed comparing to the original dataset values in order to focus on indicating students with difficulties in online learning rather than recognizing levels of students' adaptability.

2.2. Machine learning

In this paper, the implemented model is based on lazy learning. Lazy learning is a type of ML where the model delays the generalization process until a query for prediction is requested from the model [18]. These methods do not involve an explicit training phase where a model is built. Instead, they perform computation during the prediction phase. This is in contrast to eager learning methods, where the model generalizes from the training data before receiving queries. This approach typically entails storing the training data in memory and retrieving relevant data from the database to respond to specific queries, which can lead to high memory usage and computational cost during prediction. Known also as memory-based learning, this method measures relevance using a distance function, where points closer to the query are deemed more relevant. Predictions are made based on local approximations of the target function around the query point, rather than a global approximation over the entire input space.

A variant of lazy learning used in this research, called LWL, utilizes locally weighted training to average, interpolate, extrapolate, or combine training data [18]. Instead of building a global model that captures the entire data space, LWL focuses on fitting simpler models to localized subsets of the data. This approach allows for flexible modeling of complex, nonlinear relationships in the data. LWL focuses on a small neighborhood around the query point, giving more weight to data points that are closer to the query point. The hypothesis is that points near the query are more relevant for making predictions. Each training example is assigned a weight based on its distance from the query point using the selected weighting function.

The main tasks of the LWL implementation process are presented in Fig. 1. First step is determining the neighborhood of training points given a query point.

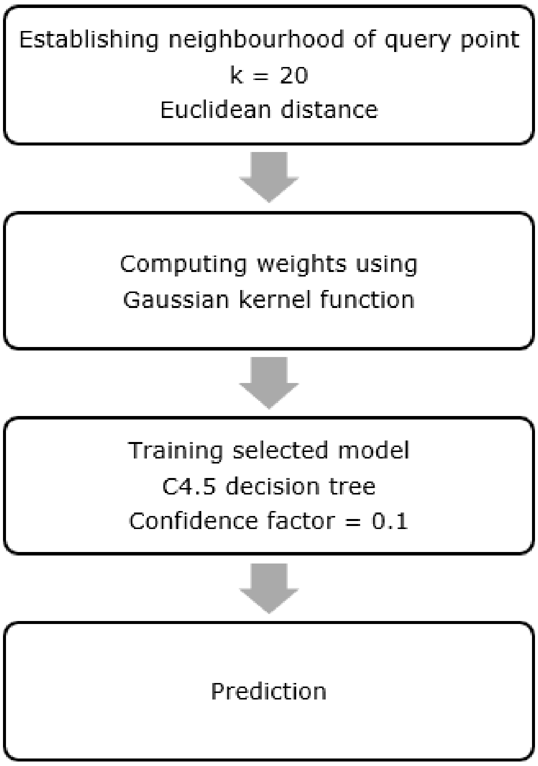

Figure 1. *Locally weighted learning based model for early detection of students with difficulties in online learning*

In this research, the number of neighbors was set to 20, whereas the search was performed using the Euclidean distance function. Further, the calculation of weights for each training point is performed using a chosen kernel function. The weight represents the distance of each point from the query point. In this paper, the Epanechnikov function is used for calculating the weights, which is mathematically represented as follows [19, 20, 21]:

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2), & |x| < 1; \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

After determining the query point space, an ML algorithm can be applied for training. The model proposed in this paper implements C4.5 decision tree as a base model for LWL with a confidence factor of 0.1. C4.5 is a prominent algorithm developed by Ross Quinlan in 1993 for generating decision trees, widely recognized for its effectiveness in ML tasks [22]. The model is presented in Fig. 1.

Building on its predecessor ID3, C4.5 uses the information gain ratio as its splitting criterion, which normalizes the information gain to mitigate bias towards attributes with many distinct values. The algorithm constructs the decision tree by recursively selecting the attribute with the highest gain ratio, splitting the dataset, and continuing the process until the stopping criteria are met. Once the full tree is generated, C4.5 prunes it to remove branches with a low contribution to the model's predictive power, ensuring better generalization. The main hyperparameter of the C4.5 algorithm is the confidence factor, which is used during the pruning phase of the decision tree construction. If the error rate with pruning is estimated to be within the confidence interval of the error rate without pruning, the node or branch can be pruned. A higher confidence factor results in less pruning, leading to a more complex tree that closely fits the training data. Conversely, a lower confidence factor results in more aggressive pruning, producing a simpler tree that may generalize better to unseen data.

2.3. Software and hardware requirements

Model development was performed in the Waikato Environment for Knowledge Analysis version 3.8.5. The hardware used for model implementation consists of NVIDIA GeForce GTX 1650 Ti GPU, AMD Ryzen 54600H 3.00GHz CPU, and 8GB of RAM.

3. RESULTS

This study is based on data that contains demographical and environmental characteristics of 1205 participants. In Table 1, the values of each input variable are presented, as well as their occurrence in relation to the output variable values. The occurrence is shown as the number of students and in a percentage format. Additionally, the total number of instances for each input value is calculated.

Within this research, multiple ML algorithms were tested in order to develop a model that can detect students with difficulties in online learning with high accuracy. The LWL-based model with the C4.5 decision tree achieved the best results with over 94% accuracy. As shown in Table 2, the classification model demonstrates a strong performance, with 94.02% of correctly classified and only 5.98% of incorrectly classified instances. This high accuracy is complemented by a Kappa statistic of 0.8753, indicating a strong agreement between the predicted and actual classifications. The model's errors are relatively low, as evidenced by a mean absolute error (MAE) of 0.0741 and a root mean squared error (RMSE) of 0.2071.

These metrics suggest that the predictions are close to the actual values. Additionally, the model exhibits a relative absolute error (RAE) of 15.46% and a root relative squared error (RRSE) of 42.30%, reflecting its robustness compared to a simple predictor. Overall, these results highlight the model's effectiveness and reliability in making accurate predictions with minimal error.

Table 1. *Participant characteristics*

Attribute	Value	Output NoP (%) Yes	Output NoP (%) No	Total NoP (%)
Gender	Female	235 (43.36)	307 (56.64)	542 (44.98)
	Male	245 (36.95)	418 (63.05)	663 (55.02)
Age	1-5	17 (20.99)	64 (79.01)	81 (6.72)
	6-10	24 (47.06)	27 (52.94)	51 (4.23)
	11-15	120 (33.99)	233 (66.01)	353 (29.29)
	16-20	144 (51.80)	134 (48.2)	278 (23.07)
	21-25	139 (37.17)	235 (62.83)	374 (31.04)
	26-30	36 (52.94)	32 (47.06)	68 (5.64)
Educational level	University	178 (39.04)	278 (60.96)	456 (37.84)
	College	120 (54.79)	99 (45.21)	219 (18.17)
	School	182 (34.34)	348 (65.66)	530 (43.98)
Institution type	Gov.	234 (61.26)	148 (38.74)	382 (31.70)
	Non-gov.	246 (29.89)	577 (70.11)	823 (68.30)
IT student	Yes	89 (29.28)	215 (70.72)	304 (25.23)
	No	391 (43.40)	510 (56.60)	901 (74.77)
Town	Yes	309 (33.05)	626 (66.95)	935 (77.59)
	No	171 (63.33)	99 (36.67)	270 (22.41)
Load-shedding	Low	380 (37.85)	624 (62.15)	1004 (83.32)
	High	100 (49.75)	101 (50.25)	201 (16.68)
Financial condition	Poor	129 (53.31)	113 (46.69)	242 (20.08)
	Middle	341 (38.84)	537 (61.16)	878 (72.86)
	Rich	10 (11.76)	75 (88.24)	85 (7.05)
Internet type	Wi-Fi	192 (37.65)	318 (62.35)	510 (42.32)
	Mobile data	288 (41.44)	407 (58.56)	695 (57.68)
Network type	4G	278 (35.87)	497 (64.13)	775 (64.32)
	3G	186 (45.26)	225 (54.74)	411 (34.11)
	2G	16 (84.21)	3 (15.79)	19 (1.58)
Class duration	0	144 (93.51)	10 (6.49)	154 (12.78)
	1-3	290 (34.52)	550 (65.48)	840 (69.71)
	4-6	46 (21.80)	165 (78.20)	211 (17.51)
Self LMS	Yes	52 (24.76)	158 (75.24)	210 (17.43)
	No	428 (43.02)	567 (56.98)	995 (82.57)
Device	Tab	2 (6.67)	28 (93.33)	30 (2.49)
	Mobile	438 (43.24)	575 (56.76)	1013 (84.07)
	Computer	40 (24.69)	122 (75.31)	162 (13.44)

* NoP – Number of participants; Gov. – Government; Non-gov. – Non-government; IT – Information technologies; LMS – Learning management system

Table 2. *Performance of the classification model*

Correctly Classified Instances	94.02%
Incorrectly Classified Instances	5.98%
Kappa statistic	0.8753
Mean absolute error	0.0741
Root mean squared error	0.2071
Relative absolute error	15.46%
Root relative squared error	42.30%

The confusion matrix which provides a detailed breakdown of the classification model's performance is presented in Table 3. Out of the actual positive instances, the model correctly predicted 444 and incorrectly predicted 36 as negative. Conversely, for the actual negative instances, the model correctly identified 689 and mistakenly classified 36 as positive. This demonstrates a high level of accuracy in both identifying true positives (444) and true negatives (689), with relatively low false negatives (36) and false positives (36). The balanced distribution of errors indicates that the model performs consistently well across both classes, maintaining a strong capability to correctly distinguish between positive ("Yes") and negative ("No") instances.

Table 3. *Confusion matrix of the classification model*

Real \ Predicted	Yes	No
Yes	444	36
No	36	689

The performance metrics that evaluate the model's effectiveness in distinguishing between two classes are presented in Table 4.

Table 4. *Performance metrics of the classification model*

Class	Yes	No	Weighted average
TP rate	0.925	0.950	0.940
FP rate	0.050	0.075	0.065
Precision	0.925	0.95	0.940
Recall	0.925	0.950	0.940
F-measure	0.925	0.950	0.940
MCC	0.875	0.875	0.875
ROC area	0.984	0.984	0.984
PRC area	0.972	0.979	0.976

* TP – true positive; FP – false positive; MCC - Matthews correlation coefficient; ROC - Receiver operating characteristic curve; PRC - Precision-recall curve

4. DISCUSSION

This study focuses on the development and validation of an ML model aimed at identifying students who are likely to face difficulties in adapting to online learning environments. This model is particularly relevant in the context of increased reliance on online education, amplified by the COVID-19 pandemic. The COVID-19 pandemic has profoundly impacted education systems worldwide, catalyzing an unprecedented shift from traditional in-person instruction to online learning [5]. This transition has highlighted both the potential and the challenges of online education, influencing students, educators, and institutions in various ways. The abrupt shift has forced educational stakeholders to adapt quickly, often with limited resources and preparation, leading to a range of outcomes that underscore the importance of effective online learning strategies and support systems [6]. This rapid transition necessitated a significant digital transformation, including the adoption of LMS, video conferencing tools, and digital resources. Institutions that were previously resistant or slow to adopt these technologies had to overcome logistical, technical, and pedagogical challenges swiftly. Despite the challenges, the pandemic-induced shift to online learning has also revealed several benefits and opportunities. Online learning can provide greater flexibility, increased accessibility, and expanded opportunities for students, allowing them to learn at their own pace and on their own schedule, breaking down the financial and locational barriers [4]. The model developed in this work leverages demographic and environmental data to achieve a highly accurate detection of students prone to difficulties in online learning, which is crucial for timely intervention and support.

The shift to online learning has been more challenging in underdeveloped and developing regions, where infrastructural and socio-economic factors play a significant role [3]. Creating inclusive educational tools that account for these disparities is of great importance, providing equal opportunities and quality education for all students around the world [4]. The predictive model proposed in this paper is a step towards ensuring that students at risk of falling behind are identified early and provided with the necessary support.

The dataset used in this study includes demographic and environmental characteristics of 1205 students. The dataset was preprocessed and used to train an LWL algorithm combined with a C4.5 decision tree classifier. The choice of LWL allows the model to adapt to local variations in the data, providing a more nuanced prediction compared to global algorithms. The model achieved a 94.02% accuracy rate, demonstrating its reliability in predicting students' adaptability to online learning.

The implementation of this predictive model holds a significant potential for educational institutions, especially in resource-limited regions. By pinpointing students who are struggling, educators and administrators can customize their support strategies to address individual needs, which could lead to improved educational outcomes overall. This model has the potential to enhance educational equity and inclusivity by ensuring that at-risk students in online learning environments are identified early. Such a proactive approach allows educators to provide tailored support and resources for all students.

In areas where disparities in access to technology and stable internet connections are common, this model can help bridge the gap for students from various socio-economic backgrounds. Educational institutions often face limitations in resources and manpower, and by employing this predictive model, schools and universities can allocate their limited resources more efficiently to the students who need them most. This targeted intervention can optimize the use of educational resources, such as tutoring, mentoring programs, and technical support, ensuring they have the greatest possible impact.

The insights gained from implementing such a predictive model can also inform policy decisions at both institutional and governmental levels. Policymakers can utilize the data to develop strategies that improve online learning infrastructures, address digital divides, and invest in areas where students are most vulnerable. This data-driven approach can lead to more effective educational policies that promote long-term improvements in online learning. Additionally, the model's ability to predict and identify students' difficulties in online learning can foster greater involvement from parents and the community. By providing timely information to parents about their children's learning challenges, educational institutions can encourage a collaborative effort between educators and families to support students' educational journeys. Community programs can also be developed to support students outside the school environment.

To enhance the model's robustness and generalizability, future research should focus on expanding the dataset to include a more diverse population. Incorporating students from different geographical regions, socio-economic backgrounds, and educational systems can help refine the model and ensure its applicability across various contexts. Future iterations of the model could also include additional variables that may influence online learning success, such as psychological factors like motivation and self-regulation, environmental factors like home learning conditions, and pedagogical factors like teaching methods and curriculum design. A more comprehensive set of predictors can improve the model's accuracy and provide deeper insights into the factors affecting online learning.

Integrating the predictive model into existing LMS can facilitate real-time monitoring and alerts. Such integration would allow educators to receive immediate feedback on students' performance and adaptability, enabling prompt interventions. Additionally, LMS integration can streamline the process by automatically collecting relevant data, ensuring the model has access to up-to-date information.

Given the constant evolution in technology and education, conducting longitudinal studies to track the long-term impact of early interventions identified by the predictive model can provide valuable insights. These studies can help understand how early support influences academic trajectories and overall student well-being, informing further refinements to the model and intervention strategies.

5. CONCLUSION

This study provides a robust framework for early detection of students with difficulties in online learning through an innovative application of ML. The high accuracy and reliability of the model suggest its potential for broad application, offering a valuable tool for educators aiming to provide equitable and effective online education. This research underscores the importance of leveraging technology to address educational challenges and highlights the need for continued innovation in this field. By ensuring that all students have the support they need to succeed, this model contributes to the broader goal of inclusive and accessible education for all.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, and these results are parts of the Grant No. 451-03-66 / 2024-03 / 200132 with University of Kragujevac – Faculty of Technical Sciences Čačak.

REFERENCES

[1] Kentnor, H. E. (2015). Distance education and the evolution of online learning in the United States. *Curriculum and teaching dialogue, 17*(1-2), 21-34.

[2] Miller, G., Benke, M., Chaloux, B., Ragan, L. C., Schroeder, R., et al. (2014). *Leading the e-learning transformation of higher education: Meeting the challenges of technology and distance education.* Stylus publishing.

[3] Aung, T. N., & Khaing, S. S. (2016). Challenges of implementing e-learning in developing countries: A review. *Genetic and evolutionary computing: Proceedings of the Ninth international conference on genetic and evolutionary computing, 2*(9), 405-411. doi: 10.1007/978-3-319-23207-2_41

[4] Gilbert, B. (2015). *Online learning revealing the benefits and challenges.* Fisher digital publications.

[5] Graham, R. (2022). *Crisis and catalyst: The impact of COVID-19 on global practice in engineering education.* MIT libraries.

[6] Besser, A., Flett, G. L., & Zeigler-Hill, V. (2022). Adaptability to a sudden transition to online learning during the COVID-19 pandemic: Understanding the challenges for students. *Scholarship of Teaching and Learning in Psychology, 8*(2), 85–105. doi: 10.1037/stl0000198

[7] Veloso, B., Barbosa, M.A., Faria, H., Marcondes, F.S., Durães, D., et al. (2023). A systematic review on student failure prediction. *International conference in methodologies and intelligent systems for techhnology enhanced learning, 538,* 43-52. doi: 10.1007/978-3-031-20257-5_5

[8] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: with applications in Python*. Springer Nature.

[9] Biswas, A.A., Majumder, A., Mia, M. J., Nowrin, I., & Ritu, N. A. (2019). Predicting the enrollment and dropout of students in the post-graduation degree using machine learning classifier. *International Journal of Innovative Technology and Exploring Engineering*, *8*(11), 3083–3088. doi: 10.35940/ijitee.K2435.0981119

[10] Ma, X., Yang, Y., & Zhou, Z. (2018). Using machine learning algorithm to predict student pass rates in online education. *Proceedings of the 3rd international conference on multimedia systems and signal processing*, 156–161. doi: 10.1145/3220162.3220188

[11] Tan, M., & Shao, P. (2015). Prediction of student dropout in e-learning program through the use of machine learning method. *International journal of emerging technologies in learning*, *10*(1), 11-17. doi: 10.3991/ijet.v10i1.4189

[12] Zulfiker, M.S., Kabir, N., Biswas, A.A., Chakraborty, P., & Rahman M.M. (2020). Predicting students' performance of the private universities of Bangladesh using machine learning approaches. *International journal of advanced computer science and applications*, *11*(3), 672–679. doi: 10.14569/IJACSA.2020.0110383

[13] Ciolacu, M., Tehrani, A.F., Beer, R., & Popp, H. (2017). Education 4.0–fostering student's performance with machine learning methods. *IEEE 23rd international symposium for design and technology in electronic packaging*, 438–443. doi: 10.1109/SIITME.2017.8259941

[14] Uskov, V.L., Bakken, J.P., Byerly, A., & Shah, A. (2019). Machine learning based predictive analytics of student academic performance in stem education. *Procceedings of the 2019 IEEE global engineering education conference*, 1370–1376. doi: 10.1109/EDUCON.2019.8725237

[15] Pinheiro, Cavalcanti, A., Ferreira Leite de Mello, R., Rolim, V., André, M., Freitas F. et al. (2019). An analysis of the use of good feedback practices in online learning courses.

IEEE 19th international conference on advanced learning technologies, *2161*, 153–157. doi: 10.1109/ICALT.2019.00061

[16] Suzan, M. H., Samrin, N. A., Biswas, A. A., & Pramanik, A. (2021). Students' adaptability level prediction in online education using machine learning approaches. *12th International conference on computing communication and networking technologies*, 1-7. doi: 10.1109/ICCCNT51525.2021.9579741

[17] Kaggle. (2022). Students adaptability level in online education. Retrieved April 28, 2024, from https://www.kaggle.com/datasets/mdmahmudulhasansuzan/students-adaptability-level-in-online-education

[18] Atkeson, C.G., Moore, A.W., & Schaal, S. (1997). Locally weighted learning. *Lazy learning*. Springer, Dordrecht. doi: 10.1007/978-94-017-2053-3_2

[19] Rajagopalan, B., & Lall, U. (1995). A kernel estimator for discrete distributions. *Journal of nonparametric statistics, 4*(4), pp. 409–426. doi: 10.1080/10485259508832629

Y. Soh, Y. Hae, A. Mehmood, R. Hadi Ashraf and I. Kim. (2013). Performance evaluation of various functions for kernel density estimation, *Open Journal of Applied Sciences, 3*(1B), 58–64. doi: 10.4236/ojapps.2013.31B012

[20] Chi-Yang Chu, C.-Y., Henderson, D.J., & Parmeter, C.F. (2017). On discrete Epanechnikov kernel functions. *Computational statistics & data analysis, 116*, 79–105. doi: 10.1016/j.csda.2017.07.003

[21] Quinlan, J. R. (2014). *C4.5: Programs for machine learning*. Elsevier.