ISSN: 2289-7844 | e-ISSN: 0127-9750
DOI: https://doi.org/10.37134/jictie.vol10.1.10.2023

Journal of ICT in Education (JICTIE)
Volume 10, Issue 1, Pages 118-134, 2023

Research Article

A Digital Literacy Predictive Model in The Context of Distance Education

Maslinda Mohd Nadzir*, Juhaida Abu Bakar

[1]*School of Computing, UUM College of Arts and Sciences, Universiti Utara Malaysia, Sintok, Kedah, Malaysia; maslinda@uum.edu.my, juhaida.ab@uum.edu.my*

Received: 20 March 2023; Revised: 29 April 2023; Accepted: 28 May 2023; Published: 10 June 2023

**corresponding author*

Abstract

Digital technologies are essential in the distance education environment. Several learners who enrol in distance education programs have difficulty using the technologies during the learning sessions. This study focuses on factors influencing digital literacy in distance education programs. A predictive model was developed based on four phases of the study; preliminary study, dataset preparation, experimental design, and analysis. For this study, 232 survey data were successfully collected. The experimental design has two phases, including data pre-processing and feature selection. - The target class is defined based on the device used and the generation. Five machine learning algorithms have been selected for the classifier model in the data analysis phase. They are logistic regression as a baseline model, k-nearest neighbour, random forest, support vector machine (SVM), and multilayer perceptron. In addition, five cross-validation settings were chosen, which are 2, 3, 5, 10, and 20-fold cross-validation. The comparison result for two-target data will be based on accuracy, precision, recall, and F-measure. Using the SVM algorithm, the best model shows the maximum accuracy with 82.9% for model learning based on generation and 41.4% for model learning based on the device. This study identifies interesting patterns based on learners' generation. It shows that different generations have various ways of using digital literacy. Thus, the proposed predictive model proves that generation is a useful indicator to recognize the level of digital literacy among higher education learners, especially those for distance learning.

Keywords: digital literacy, distance education, online learning, supervised learning approach, predictive model

INTRODUCTION

Learning is a process that has changed from face-to-face to online learning because of the rapid development of digital technologies. Digital technology involves learning via various technology platforms, such as online learning, web conferencing, WhatsApp, Telegram, Facebook, or a

combination of the technologies. During this process, digital literacy appears as practical literacy. Digital literacy can improve lifelong learning and is related to predictions of academic life (Knox & Bayne, 2013).

Digital literacy can be described as finding and evaluating information, generating content, and communicating with people via technology. In this study, digital literacy involves determining what technologies to use based on individual needs and disciplinary practices (Newland & Handley, 2016). The term digital literacy is widely applied in curriculums. However, the study shows that universities do not understand the types of digital literacy competencies that would prepare graduates, for the digital world (Press et al., 2019).

Currently, most educators presume that students are digitally literate. Although students use digital technologies for social purposes, many have not engaged in an online learning environment using digital technologies efficiently (Pangrazio, 2019). Digital technologies provide an opportunity to meet the needs of the learning environment and overcome learning challenges. However, several studies have argued that technology has not significantly affected the learning environment (Sarker et al., 2019). As a result, recent studies show that many undergraduates exhibit poor employment of digital literacy in online learning (Sulianta, 2019).

To be an effective learner in learning environments, especially in online learning, it is highly beneficial to be digitally literate. With digital skills, students can learn, participate, communicate, and prepare themselves to access information. Digital skills are disseminated remotely using various platforms such as online learning portals, online discussion boards, video conferencing, and social media platforms due to the COVID-19 pandemic that places limitations, on physical presence in the same lecture room.

In response to the COVID-19 pandemic, physical distancing is enforced to minimize the spread of the virus. This paradigm change has a significant impact on basic and higher education, making digital technology an essential requirement. Thus, educators should include information technology and use digital technology in the teaching and learning process. In line with that, understanding digital literacy is vital for distance education students during the distance learning process. In the context of this study, distance education students are undergraduate students who have enrolled in higher education via distance learning at a University in the Northern Region of Malaysia.

The distance education system allows students with their own study speed, time, and place flexibility. It also creates a physical separation between students and lecturers most time. face-to-face lectures are held once a month to assist the students in understanding the course topics. However, the COVID-19 pandemic has temporarily transformed face-to-face learning sessions into blended learning through e-learning and web conferencing platforms. Thus, several students struggle during the blended learning sessions. It is consistent with Lowe et al. (2016), who recommended that virtual lectures should be used to enhance physical lectures, rather than a replacement for face-to-face lectures. Although face-to-face lectures will be permitted after the pandemic, the blended learning sessions will complement

the lectures. As a result, distance education students should attain a decent level of digital literacy to use digital technology at the time of the blended learning process.

On the other hand, supervised learning entails learning a mapping between a set of input variables X and an output variable Y and applying this mapping to predict the outputs for unseen data (Cunningham et al., 2008). Supervised learning is used in the majority of machine learning methodologies. Machine learning refers to a computer program, that calculates and deduces the information related to the task and obtains the characteristics of the corresponding pattern (Alloghani et al., 2019). This technology can achieve an accurate and economical classification of digital literacy users; hence, it might be a promising method for predicting digital literacy amongst learners. Hence, this study aims to understand digital literacy amongst distance education students, who are generally adult learners seeking higher education to improve their education level using a supervised learning approach.

This study is presented as follows: these sections are a literature review of the overview of digital literacy and the online learning environment, distance education, predictive analytics, and related studies. Then, the following section outlines the methods used in this study, followed by results and discussion. The final section then presented the conclusion and future study.

Digital Literacy and Online Learning Environment

Higher education is transitioning the distance education system into an online learning environment. Several factors must be considered for the effective implementation of online learning including the importance of interactive interaction, full access to relevant digital technology, and the support of higher education institutions to distance education students, who participate in distance learning via specific platforms. To effectively use digital technology in online learning, students should have a sufficient degree of digital literacy, which includes skills such as utilizing and managing digital technology.

Digital literacy is a necessary skill for using digital technologies, and processing and retrieving relevant information (Karpati, 2011). According to Greene, Seung, and Copeland (2014), digital literacy is more than just being able to access digital sources; it is also concerning searching, evaluating, and combining credible information into a meaning-making activity throughout online learning. Ng (2012) distinguished three-intersecting dimensions: digital literacy technical, cognitive, and social-emotional dimensions. Chetty et al. (2018) on the other hand identify digital literacy as a multi-disciplinary concept comprising five disciplines: information literacy, computer literacy, media literacy, communication literacy, and technology literacy. Overall, digital literacy is presented as an approach to perform naturally in online environments and efficiently access the wide range of knowledge embedded in the environments (Martin, 2008).

A study by Alfia, Sumardi, and Kristina (2020) describes digital literacy as a set of skills to improve student's thinking skills and facilitate access to accurate and relevant information. Students are thought to be competent to critically evaluate information gathered because not all information retrieved from

the internet is credible. As a result, a digitally literate student understands how to select and use digital technologies as needed (Ozdamar-Keskin et al., 2015). According to Ozdamar-Keskin et al. (2015), digitally literate students in this study are described as distance education students who can use digital technology for distance learning.

Distance Education

Distance education offers flexible learning opportunities to access various delivery systems and learning materials in order to acquire knowledge. Distance education is essential in providing an alternative for students who are unable to further their studies in higher education due to financial limitations. Distance education drives easy access and convenience of learning opportunities (Purnama et al., 2021). In addition, the distance education program allows opportunities for Malaysians, particularly those who work full-time, to improve their quality of life via higher education. Distance education takes a student-centred approach to learning supported by face-to-face lectures once or twice a month. This approach allows students to determine the time, place, and learning style that suits their everyday lives and needs. As a result, Information and Communication Technology (ICT) should be used to enhance the synchronous and asynchronous interaction between students and lecturers (Maphosa & Bhebhe, 2019).

Furthermore, Hassan and Mirza (2020) support integrating technology, particularly ICT into the distance education system, to meet student's requirements in distance modes, such as content delivery, communication, and feedback. However, there are distance education students with minimal digital literacy levels, making it difficult to study in such situations. As a result, there must be adequate assistance for such students in any distance learning environment.

Predictive Analytics

Knowledge Discovery in Databases (KDD) and data mining approaches have been effectively utilized in several academic areas to extract new and relevant knowledge from historical data (Ghazal and Hammad 2020). KDD is the process of obtaining usable information and contributing to the cause of knowledge discovery using large datasets. The first step in the KDD process is selection. This stage requires building a target data set or focusing on the selection of variables or data samples for which discovery will be made. The second step consists of Data preprocessing; this stage comprises cleaning and preparing the target data to produce consistent data. The third step is transformation; this level involves data manipulation using dimensionality reduction or transformation methods. The fourth step is Data mining; depending on the purpose of the data mining, this stage entails looking for patterns of interest in a certain representational form. The final step is Pattern evaluation; this stage entails interpreting and evaluating the mined patterns (Azevedo, 2019).

The phrase "data mining" was coined in 1990, and it incorporates several fields such as database management systems (Saeed, 2017), statistics, artificial intelligence (Dick, 2019), and machine learning (Jordan and Mitchell, 2015). The goal of its orientation and application is the same since it

entails producing and aligning meaningful patterns using different methods. Prior to the advent of Statistics and Machine Learning approaches, algorithms (Dey, 2016) produced the greatest results for numerical data, but these are currently being created for non-numerical or qualitative data as well. The primary output remains the same: to produce relevant results, advance research methodologies, or create knowledge (Dangeti, 2017). Data mining methods rely solely on statistical approaches, but with the integration of Artificial Intelligence, Machine Learning techniques, and Pattern Recognition (Dangeti, 2017), heterogeneous ideas and anomalies can now be found inside data warehouses.

Williams's (2011) perspective on descriptive and predictive analytics is that descriptive analytics is the problem of representing newly acquired knowledge without necessarily modelling a certain conclusion. This category includes activities like cluster analysis, association and correlation analysis, and pattern finding. From the standpoint of machine learning, we may relate these techniques to unsupervised learning. The goal of unsupervised learning is to find patterns in data that will help us learn more about the world that the data represents. In most cases, there is no unique target variable that we are aiming to represent. Rather, these techniques shed light on the patterns shown by descriptive analytics. Meanwhile, predictive analytics is known as supervised learning.

A clustering analysis for descriptive analysis to determine the relationship between ICT use and literacy and the academic level of a group of teachers who develop their academic activities at a university in the city of Sincelejo, Department of Sucre, Colombia, is shown in Vitola & Sierra (2021). The closest neighbours were used in a cluster analysis using the KNN or K-techniques. The KNN or K- techniques, or closest neighbours, were used to do cluster analysis, employing the Metric Distance or Euclidean Distance as a measure of similarity. Therefore, it was possible to determine that the variables related to the dimension "Use of ICT and Technology Literacy" and the Level of Training of University Teachers had no evident pattern of association. Teachers' scores varied depending on their professional training in all of the variables that make up the analyzed dimension, as evidenced by the fact that several Teachers with Doctoral or Master's Training received a Low or Nil rating in both variables, while others with Basic Professional Training or Specialization received Very High qualifications in the same variables. A study by Ozdamar-Keskin (2015) was set to investigate the digital literacy skills and learning patterns of students enrolled in Anadolu University's open and remote education system in Turkey. Descriptive data analysis for demographic information and The Principal Component Factor Analysis was used to organize and categorize the learners' attitudes and assertions about their personal learning preferences, problem-solving capabilities, project-work skills, and ability to use digital resources for learning.

The OU Analyze (OUA) System, which falls under Predictive Learning Analytics (PLA), was utilized in research by Herodotou et al. (2020). The purpose of this research is to consider the macro-level of adoption by describing use, obstacles, and characteristics that facilitate acceptance at an organizational level, as well as the micro-level of adoption, which is instructors' perceptions of OUA. Every week, OUA predicts whether (or not) a particular student will submit their next teacher-marked assignment. The OUA dashboard displays predictive data for individual students, such as who is at risk of missing their next assignment due date, as well as cohort-level engagement and assignment submission rates

in Virtual Learning Environments (VLEs). Herodotou et al. (2020) offer predictive modelling, which is the Naive Bayes classifier (NB), the classification and regression tree (CART), and the k-Nearest Neighbours technique, three machine learning algorithms used by OUA (k-NN). Faculty involvement with OUA, instructors as "champions," evidence collection and sharing, digital literacy, and attitudes toward online education were all cited as critical aspects for scaling up PLA implementation. Confusion matrix data (True Positive, True Negative, False Positive, and False Negative) for all courses was averaged each week to balance the relative size of various courses to evaluate OUA's predictions. This matrix was used to determine Precision, Recall, F-measure, and Accuracy for the courses under consideration.

In the study by Urbancikova et al. (2017), research on the social, economic, demographic, and geographical elements that influence digital literacy as a foundation for digital success was conducted. The most significant characteristics of digital literacy in Slovakia are age, education, income, and home type, which contribute to the social digital gap. The size of the city and the sector of the economy are less relevant, while the impact of region, gender, and nationality is relatively marginal. Pooling, Random effects, and Fixed-effects regression models, as well as the Item Cluster Analysis approach, are used to analyze cross-sectional data. The coefficients' values indicate the strength of the relationship between socio-demographic characteristics and digital literacy components.

Research by Reddy et al. (2020) was to examine the digital capabilities of first-year university students at a higher education institute. To derive the scale, the Explanatory Factor Analysis (EFA) and the Cronbach alpha were used to validate *digilitfj*. According to the findings, 86.15 per cent of pupils are digitally literate in the range of average to very literate, while 14.71 per cent are digitally literate in the area of very low to low. The most relevant and impactful characteristics in determining an individual's digital literacy proficiency were evaluated using RapidMiner's predictive modelling machine learning approach. Deep Learning, Decision Tree, Random Forest, and Support Vector Machine (SVM) techniques were used to create predictive models. The four algorithms were chosen by the writers of this work because they were the most regularly utilized algorithms for educational data mining. When compared to other algorithms, Deep Learning, SVM, Random Forest, and Decision Tree algorithms were shown to be the best suited for prediction in this study. To assess the predictive modelling, the authors employ Error Rate, Accuracy, and Precision.

In Gómez-Galán et al., (2021) study, the study examines how information and communication technologies (ICTs) are integrated in the academic context. This study looks at how university students in Hispanic nations utilize ICTs, particularly the Internet. The descriptive analysis employed in this study was based on the text's replies to the COBADI® instrument's selected questions, which were of the 'free answer' kind. This research was carried out using the R programming language. Text mining is a technique for extracting as much information as possible from a large number of answers, recognizing trends, and comprehending current information.

RELATED STUDIES

Kim (2019) examined the relationship between South Korean college student's digital literacy, learning strategies, and core competencies. The study's findings revealed that digital literacy had a direct impact on essential abilities such as online information search strategies; moreover, learning strategies mediated the relationship. In addition, this study suggests that future researchers should contribute to the relationship between digital literacy and basic learner characteristics such as learning style via various digital technologies such as social media and online learning platforms.

The level of digital literacy skills within a group of university students in Papua New Guinea, as well as their capability to participate in the digital environment, was investigated. The finding indicated that many students at the university have a limited understanding of digital literacy. Further efforts are needed to promote digital literacy awareness of digital technologies within the online learning environment (Kolodziejczyk et al., 2020). Clark (2020) on the other hand, conducted an autoethnographic study of developing digital literacy, which perceives digital literacy as an interaction between skills, practices, and identity. The findings suggest that identity is important in enhancing an individual's digital literacy. Further study on identity change in digital literacy is suggested.

Recently, systematic literature reviews have been conducted on digital literacy themes and predictive analytics. The study trends show that digital literacy is a fundamental competency that learners must improve in various fields and stages. Accordingly, Park et al. (2020) suggested increasing the originality of digital literacy-related studies and future research should investigate a broader range of digital literacy environments. Meanwhile, Sanchez-Caballe et al. (2020) proposed that it is essential to emphasize digital literacy development in educators and learners to adapt to the pace of digital literacy technology evolution. Furthermore, most of the documents selected in the systematic literature review do not show that the students have acquired an adequate level of digital literacy. On the other hand, most related studies that work on predictive analytics show that statistical (i.e., logistic regression) and machine learning such as KNN, random forest, SVM, and multilayer perceptron (MLP) have been utilized. Thus, there is a need to study the broader scope and techniques of digital literacy environments with distance education students who are generally adult learners pursuing higher education and working simultaneously.

METHODOLOGY

In this chapter, methodology-related studies will be discussed and the numerous ideas and techniques needed will be discussed. As discussed in the previous section, many literature issues are motivating the study. There are four phases in this study, which are a preliminary study, dataset preparation, experiment, and analysis. In the first phase, the researchers identify the problem statement, literature review, gap, and objective of the study.

The second phase focuses on the method for obtaining datasets for this study. A standard questionnaire taken from the study by Ozdamar-Keskin et al. (2015) was used in this study. The study works on

digital literacy competencies and learning habits of learners enrolled in the open and distance education system of Anadolu University in Turkey. The learner's attitudes and statements about their learning preferences, problem-solving capabilities, project work skills, and ability to use digital technologies for learning were grouped and classified using Principal Component Factor Analysis. Their personal learning preferences formed five components, according to the findings: visual, auditory, dependent, collaborative, and reading-writing learning styles.

The third phase is the experimental phase, which examines the dataset to obtain interesting patterns. This study constructs a prediction model that can be developed based on digital literacy patterns in an online learning environment. This study focuses on students enrolled in open and distance education managed by *Professional and Continuing Education, Universiti Utara Malaysia* (PACE). Students were divided into four generations for this purpose: Baby Boomers, X, Y, and Z. There were four generational groups identified among the respondents to the questionnaire. The attribute 'Device', which shows the learners' abilities to use digital technologies, has been tested as target data. A Comparison study between these two target data has been evaluated.

The last phase, Data Analysis, discusses the results of experiments performed with specific settings such as determining the ranking of attributes for this study, the relationship between attributes and classes produced in the previous phase, and the performance and evaluation models based on standard statistical methods such as Accuracy and F1-measure. Figure 1 shows the research methodology.

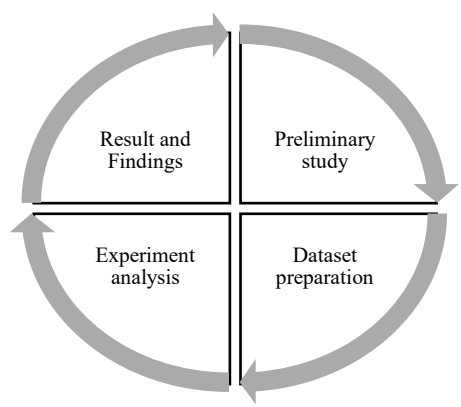

Figure 1: Research methodology

RESULTS AND DISCUSSION

Dataset and Specifications

The questionnaire from the study has been adopted from the study of Ozdamar-Keskin et al. (2015). PACE Universiti Utara Malaysia (UUM) has been organised to spread the questionnaire among open

and distance students and 232 surveys were successfully collected from 11 May 2021 to 26 October 2021. The specifications of the dataset are shown in Table 1.

Table 1: Dataset and Specifications

Subject	Details
1	Demographic information
	A. Gender
	B. Age
	C. Faculty (School)
	D. Year
2	Learners' Abilities to Use Digital Technologies
	A. Ownership of Information and Communication Technologies
	B. The Frequency of Using Technology and the Purposes
	C. Learning Habits
3	Personal learning preferences
	A. 1st Factor: Visual Learning Style
	B. 2nd Factor: Auditory Learning Style
	C. 3rd Factor: Dependent Learning Style
	D. 4th Factor: Collaborative Learning Style
	E. 5th Factor: Reading-writing Learner Style
4	Problem-solving skills
	A. 1st Factor: Working on a problem
	B. 2nd Factor: Evaluating the solution methods
	C. 3rd Factor: To avoid solving a problem
5	Project work skills
	A. 1st Factor: Planning
	B. 2nd Factor: Project Management
	C. 3rd Factor: Evaluation of the Results
6	Ability to use digital tools for learning purposes
	A. 1st Factor: Ability to use digital learning tools
	B. 2nd Factor: Managing digital learning platforms
	C. 3rd Factor: Ability to use advanced-level digital tools
	D. 4th Factor: Security and Ethics

Experimental Design

The experiment phase design has been set into two phases, which are Data pre-processing and feature selection. A supervised learning approach has been applied. Therefore, our aim in this experiment is to identify target data and to perform the evaluation based on the experiment settings. In the next phase, the performance of the model will be evaluated based on the standard statistical measurements, i.e., accuracy, F-1 measure, precision, and recall.

a) Data pre-processing

Statistical analysis has been done in this phase. First, duplicate usernames are traced to make sure only one entry is answered by the specific students. duplicate usernames are removed. After this process, 19 duplicate values were found and removed. 210 unique values remained. Then, missing values have also been identified. All features do not have any missing values. This is because the Google form created earlier has been set to require all answers when the respondents fill in the questionnaire.

In this phase, the researchers also re-evaluate answers given for the features below the subject 'Learners' Abilities to Use Digital Technologies', that is the question 'Which digital tool do you use for learning purposes?' Data reduction has been applied as data discretization. Data Discretization is used to obtain a reduced representation of the data while minimizing the loss of information content. The original dataset had 31 categorical values, which had duplicate information, so the researchers changed the form to numeric value by calculating how many devices/digital tools were used for learning purposes. Therefore, in the dataset, a new attribute is created named 'Device'. This attribute has five data values which are numbered 1–5 based on the number of devices used for learning purposes. In this phase, several attributes have been identified as meta-data, such as attribute 64: 'Timestamp', attribute 65: 'Username', attribute 2: 'Age', and attribute 5: 'Which digital tool do you use for learning purposes?'. Figure 2 shows the pre-processing steps.

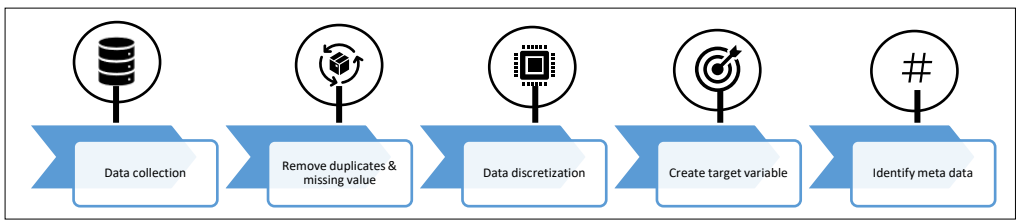

Figure 2: Data pre-processing steps

b) Feature selection

Feature selection is important to determine the input feature suitable for the training. Several statistical approaches can be used in determining the importance of one attribute such as finding the correlation pattern and ranking. In this part, the target class will be determined. The correlation analysis in Table 2 shows 20 attributes that correlate with the other attributes.

i. Correlation analysis

Based on the analysis, there is a strong correlation below one subject. All 20 most correlated information above is from the subject 'Ability to Use Digital Tools for the Purpose of Learning'

Table 2: Twenty correlation analysis

Correlation	Feature	Feature
+0.872	Digital platforms files	Using several applications
+0.863	Digital platforms	Digital platforms files
+0.836	Using several applications	Digital objects
+0.829	Digital objects control	Digital objects
+0.814	Digital platforms	Using several applications
+0.813	Digital ownership	Online communication
+0.802	Digital objects control	Using several applications
+0.799	Digital objects control	Digital platforms
+0.798	Digital platforms	Digital objects
+0.797	Online ethics	Digital ownership
+0.796	Social media	Digital platforms
+0.795	Google list	Online campaigns
+0.788	Social media	Using several applications
+0.782	Online ethics	Online communication
+0.774	Social media	Digital platforms files
+0.773	Digital platforms files	Digital objects
+0.765	Digital objects control	Digital platforms files
+0.756	Online collaboration	Social media
+0.749	Cloud applications	Online collaboration
+0.748	Digital content	Using several applications

ii. Ranking

Table 3 shows the ranking attributes based on their importance. The 10 based rank scores are based on Information Gain. Information Gain is a statistical property that measures how well a given attribute separates the training examples according to their target classification (Mitchell 1999, Pandey, 2008).

Table 3: Ten attributes based on ranking

Attributes	Subject	Information Gain
Online moderator	Abilities to use digital tools for learning purposes	0.050
Drawing – summarise	Personal learning preferences	0.049
Telling than reading	Personal learning preferences	0.049
Programme	Demographic Information	0.049
Study group	Personal learning preferences	0.040
Trusting peers	Problem-solving skills	0.039
QR code	Abilities to use digital tools for learning purposes	0.039
Social media	Abilities to use digital tools for learning purposes	0.038
Seeing others	Personal learning preferences	0.038
Devices	Learners' Abilities to Use Digital Technologies	0.035

Based on the findings above, the researchers see that most of the important attributes are below the subject personal learning preferences with four important attributes, the subject 'Abilities to use digital tools for learning purposes' with three important attributes and one important feature for each subject

below 'Demographic Information', 'Problem-solving skills' and 'Learners' Abilities to Use Digital Technologies'.

iii. Target class

To use this dataset, the researchers have created new attributes named Generation. In this attribute, it will act as a Target Data. The purpose of creating this new attribute is to see the significance between features and generation. Is there any interesting pattern between digital literacy and generation differences? Based on the Age feature, we constructed new attributes with four types of categorical values: Baby boomers, X, Y, and Z. Tables 3 and 4 show the total instances falling below each of these categorical values.

Table 4: Total instances for Generation

Generation	Total instances
Baby Boomers	1
X	6
Y	173
Z	30

Other than that, researchers also try using 'Device' to act as Target Data. The data type has been set to categorical data type. Later in the analysis phase, we will compare the results between these two target data.

Data Analysis

In this phase, the two processes have been done. One is about model selection and another is learning.

a) Model selection

Five machine-learning algorithms have been selected for the classifier model. There is logistic regression as a baseline model, k-nearest neighbour (KNN), random forest, support vector machine (SVM), and multilayer perceptron (MLP). KNN is a lazy method classifier when its prediction is required for a new data instance, the algorithm will search via the training dataset for the k-most similar instances (Phyu, 2009). Meanwhile, Random Forest, SVM, and MLP are eager method classifiers, which the system tries to generalize the training data before receiving queries. The target function will be approximated globally during model development. The random forest is a tree classifier, SVM is a function classifier and MLP is based on the neural network (NN) technique. The NN analytic technique is modelled after the hypothesized processes of learning in the neurological functions of the brain (Tso & Yau, 2007).

b) Learning

The learning of this model is based on the different settings of partitions. The model performance has been performed based on k-fold cross-validation. Five settings of cross-validation were chosen, which

are 2, 3, 5, 10 and 20-fold cross-validation have been used in this study. The comparison result for two-target data will be based on accuracy (Acc) and F-measure (F1).

Based on the data analysis above, it shows that support vector machines get the highest accuracy with 82.9% for model learning based on generation. Other machine learning classifiers, such as Random Forest and K-nearest neighbours, get the second and third maximum accuracy with 82.4%, Multilayer Perceptron with 80% and the baseline classifier logistic regression with 77.1%. Cross-validation between 5, 10, and 20-fold shows no difference in the analysis.

However, data analysis based on device use for learning purposes shows low accuracy for all experiments. This means that the experiment does not find any significant difference between using 1–5 devices for learning purposes. In this experiment, it shows that support vector machines also got the highest accuracy with 41.4% for model learning based on the device. Other machine learning classifiers such as K-Nearest neighbours get the second maximum accuracy at 38.6%, Random Forest at 37.1%, Logistic Regression at 34.8%, and Multilayer Perceptron with the lowest accuracy of 31.4%. The 10-fold cross-validation shows the best performance for the model.

c) Testing on unseen data

The data collected after 30 June 2021 has been set to become testing data. Three new instances were tested for the validation of the training model. The prediction on the dataset shows that SVM predicts the instance class for all testing data as Generation Y (Gen Y). Other machine learning techniques, such as Random Forest, K-Nearest Neighbours, MLP, and Logistic Regression, predict new unseen data to class Gen Y, respectively. The actual target data is Gen Y for all testing data. This good prediction happened because of the high accuracy achieved in the learning model.

On one hand, when we run prediction to predict the device, all machine learning classifiers predict various numbers for one instance, such as SVM, Neural Network and Logistic predict 2, Random Forest and k-NN predict 1 and Naive Bayes predict 5. Actual target data is 2. The difference in prediction based on the device happens due to the low accuracy achieved in the learning model. Table 7 shows the overall result of the testing results. Figure 3 shows the illustration of the proposed experiment.

Table 5: Evaluation of prediction model performance based on Generation

Techniques	**Logistic Regression**		**k-Nearest Neighbor**		**Support Vector Machine**		**Random Forest**		**Multilayer Perceptron**	
Tr-Te	***Acc***	***F1***	***Acc***	***F1***	***Acc***	***F1***	***Acc***	***F1***	***Acc***	***F1***
2-fold cross-validation	**77.1**	74.9	81.9	74.2	82.4	74.4	81.9	74.2	**80.0**	75.2
3-fold cross-validation	74.8	72.8	**82.4**	74.4	82.4	74.4	80.5	73.5	80.0	74.9
5-fold cross-validation	77.1	74.7	81.0	73.7	**82.9**	75.5	81.9	74.2	79.5	75.8
10-fold cross-validation	75.7	72.8	81.0	73.7	82.9	75.5	**82.4**	74.4	79.5	75.0
20-fold cross-validation	76.7	74.2	81.9	75.0	82.9	75.5	81.9	75.0	80.0	75.6

Table 6: Evaluation of prediction model performance based on Device

Techniques	Logistic Regression		k-Nearest Neighbor		Support Vector Machine		Random Forest		Multilayer Perceptron	
Tr-Te	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*	*Acc*	*F1*
2-fold cross-validation	25.7	24.4	32.9	26.7	36.7	26.2	36.2	33.6	27.6	26.1
3-fold cross-validation	33.8	32.9	36.2	29.0	31.4	22.4	31.4	29.4	29.0	27.5
5-fold cross-validation	**34.8**	33.7	**38.6**	31.5	38.6	29.2	**37.1**	33.0	28.6	27.5
10-fold cross-validation	31.0	30.3	35.7	28.3	**41.4**	32.9	35.7	33.6	27.1	26.6
20-fold cross-validation	30.5	29.8	34.8	27.7	39.5	32.1	33.8	31.3	**31.4**	31.0

Table 7: Testing data result based on two classifiers; SVM and K-NN

Prediction	Actual			Predicted (SVM)			Predicted (kNN)			Accuracy (SVM)	Accuracy (kNN)
Based on Generation	Y	Y	Y	Y	Y	Y	Y	Y	Y	100	100
Based on Device	2	1	1	2	1	1	1	1	1	100	66.7%

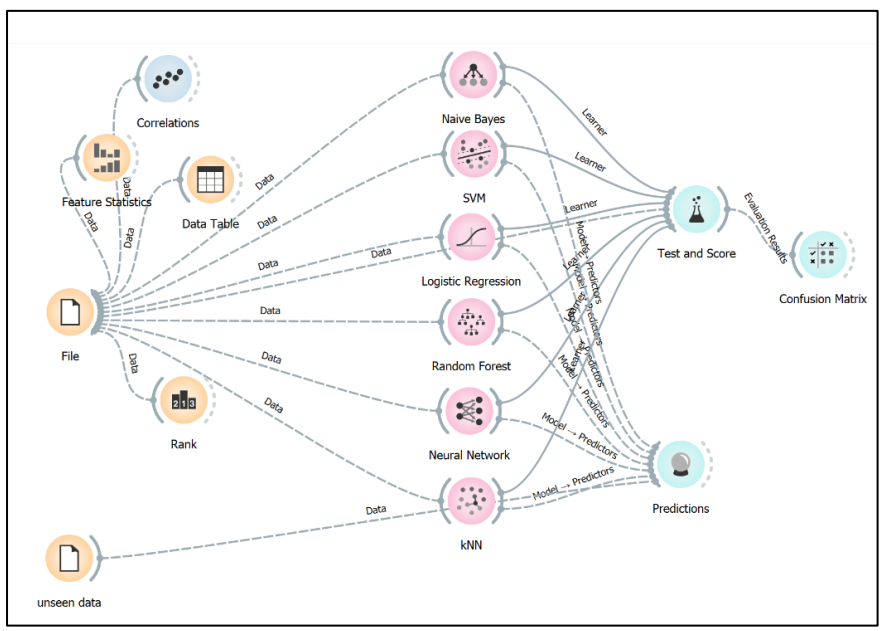

Figure 3: The illustration of the proposed experiment using *Orange*

CONCLUSION

The main aim of this study is to identify interesting patterns based on digital literacy among students in an online learning environment. It shows that there is less interesting or significant pattern dependent on the skills of learners to use digital technologies. Most learners in this study stand out in their ability

to use digital tools for learning purposes. This is due to 20 correlation analyses revealing that they all fall into the same subject. Learners do not have any problems using digital tools. However, the earlier plan of this study was to investigate interesting patterns in specific subjects such as problem-solving skills and project management skills. Based on the analysis, one attribute for problem-solving skills shows an important attribute to be used, which is trusting peers to solve a problem (0.039). It shows that all learners choose similar answers, and it makes the pattern consistent for this question. Generally, the learners in this study believe and trust their friends to solve problems assigned by the lecturer. However, there is no important and interesting pattern detected for the project work skill.

Overall, this study identifies interesting patterns based on learners' generation. It shows that different generations have different ways of using digital literacy. The prediction model based on digital literacy shows 82.9% accurate results in predicting unseen data. New unseen data also show correct prediction between actual and predicted for all ML classifiers. It shows the stability of the predicted model. However, the prediction based on device use does not show a good learning model. Thus, the prediction on unseen data also shows the instability of the result with various predictions. Future studies should consider using different data mining tasks such as association rule mining and sequential pattern mining.

ACKNOWLEDGMENTS

This research has been carried out under Geran Penjanaan (SO Code: 14916) provided by Universiti Utara Malaysia (UUM), Malaysia.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

DECLARATION OF GENERATIVE AI

The authors declare that no generative AI was used in the writing of the manuscript.

DATA AVAILABILITY STATEMENT

Data available within the article or its supplementary materials.

REFERENCES

Alfia, N., Sumardi, S., & Kristina, D. (2020). Survival skills in the digital era: An integration of digital literacy into EFL classroom. *Indonesian Journal of EFL and Linguistics*, *5*(2), 435–451.

Alloghani, M., Al-Jumeily, D., Baker, T., Hussain, A., Mustafina, J., & Aljaaf, A. J. (2019). Applications of machine learning techniques for software engineering learning and early prediction of students' performance. In *Soft Computing in Data Science: 4th International Conference, SCDS 2018, Bangkok, Thailand, August 15-16, 2018, Proceedings 4* (pp. 246-258). Springer Singapore.

Azevedo, A. (2019). Data mining and knowledge discovery in databases. In *Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics* (pp. 502-514). IGI Global.

Chetty, K., Qigui, L., Gcora, N., Josie, J., Wenwei, L., & Fang, C. (2018). Bridging the digital divide: measuring digital

literacy. *Economics*, *12*(1), 2017–2069. https://doi.org/10.5018/economics-ejournal.ja.2018-23

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval* (pp. 21-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

Dey, A. (2016). Machine learning algorithms: A review. *International Journal of Computer Science and Information Technologies* (IJCSIT). *7*(3),1174-1179. https://doi.org/10.21275/ART20203995

Dick, S. (2019). Artificial Intelligence. Harvard Data Science Review, 11, 1.

Ghazal, M. M., & Hammad, A. (2022). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. *International Journal of Construction Management*, *22*(9), 1632-1646. https://doi.org/10.1080/15623599.2020.1738205

Gómez-Galán, J., Martínez-López, J. Á., Lázaro-Pérez, C., & Fernández-Martínez, M. D. M. (2021). Usage of internet by university students of hispanic countries: Analysis aimed at digital literacy processes in higher education. *European. Journal of Contemporary Education*, *10*(1), 53-65.

Greene, J. A., Seung, B. Y., & Copeland, D. Z. (2014). Measuring critical components of digital literacy and their relationships with learning. *Computers & education*, *76*, 55-69. https://doi.org/10.1016/j.compedu.2014.03.008

Hassan, M. M., & Mirza, T. (2020). Information and communication technology (ICT) in the distance education system: An overview. *IOSR Journal of Research & Method in Education (IOSR-JRME)*, *10*(6), 38-42. https://doi.org/10.9790/7388-1006053842

Herodotou, C., Rienties, B., Hlosta, M., Boroowa, A., Mangafa, C., & Zdrahal, Z. (2020). The scalable implementation of predictive learning analytics at a distance learning university: Insights from a longitudinal case study. *The Internet and Higher Education*, *45*, 100725. https://doi.org/10.1016/j.iheduc.2020.100725

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa841

Karpati, A. (2011). Digital literacy in education', IITE *Policy Brief May 2011*. UNESCO Institute for Information Technologies in Education. https://unesdoc.unesco.org/ark:/48223/pf0000214485.

Kim, K. T. (2019). The structural relationship among digital literacy, learning strategies, and core competencies among South Korean college students. *Educational Sciences: Theory and Practice*, *19*(2), 3-21.

Knox, J., & Bayne, S. (2013). Multimodal profusion in the literacies of the Massive Open Online Course. *Research in Learning Technology*, *21*. https://doi.org/10.3402/rlt.v21.21422

Kolodziejczyk, I., Gibbs, P., Nembou, C., & Sagrista, M. R. (2020). Digital skills at Divine Word University, Papua New Guinea. *IAFOR Journal of Education*, *8*(2), 107-124.

Lowe, T., Mestel, B., & Williams, G. (2016). Perceptions of online tutorials for distance learning in mathematics and computing. *Research in Learning Technology*, *24*. https://doi.org/10.3402/rlt.v24.30630

Martin, A. (2008) 'Digital literacy and the 'digital society'', In *C. Lankshear, & M. Knobel (Eds.), Digital Literacies: Concepts, Policies & Practices,* New York: Peter Lang, pp. 151–176.

Maphosa, C., & Bhebhe, S. (2019). Digital literacy: A must for open distance and e-learning (ODEL) students. *European Journal of Education Studies*. *5*(10), 186–199. http://dx.doi.org/10.46827/ejes.v0i0.2274

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

Newland, B., & Handley, F. (2016). Developing the digital literacies of academic staff: An institutional approach. *Research in Learning Technology*, *24*.https://doi.org/10.3402/rlt.v24.31501

Ng, W. (2012). Can we teach digital natives' digital literacy? *Computers & Education*, *59*(3), 1065-1078. https://doi.org 10.1016/j.compedu.2012.04.016

Ozdamar-Keskin, N., Ozata, F. Z., Banar, K., & Royle, K. (2015). Examining digital literacy competences and learning habits of open and distance learners. *Contemporary Educational Technology*, *6*(1), 74-90.

Pandey, S. (2008). *Methods For Approximating Forward Selection Of Features In Information Retrieval Problems Using Machine Learning Methods* (Doctoral dissertation). University of Minnesota, Duluth

Pangrazio, L. (2018). *Young people's literacies in the digital age: Continuities, conflicts and contradictions*. Routledge.

Park, H., Kim, H. S., & Park, H. W. (2020). A scientometric study of digital literacy, ICT literacy, information literacy, and media literacy. *Journal of Data and Information Science*, *6*(2), 116-138. https://doi.org/10.2478/jdis-2021-0001

Phyu, T. N. (2009). Survey of classification techniques in data mining. In *Proceedings of The International Multiconference Of Engineers and Computer Scientists* (Vol. 1, No. 5).

Press, N., Arumugam, P. P., & Ashford-Rowe, K. (2019). Defining digital literacy: A case study of Australian universities. In *ASCILITE 2019 Conference Proceedings: 36th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education* (pp. 255-263).

Purnama, S., Ulfah, M., Machali, I., Wibowo, A., & Narmaditya, B. S. (2021). Does digital literacy influence students' online risk? Evidence from Covid-19. *Heliyon*, *7*(6). https://doi.org/10.1016/j.heliyon.2021.e07406

Reddy, P., Chaudhary, K., Sharma, B., & Chand, R. (2020, December). Digital Literacy: A Catalyst for the 21st Century Education. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* (pp. 1-6). IEEE. https://doi.org/10.1109/CSDE50874.2020.9411548

Saeed, A. M. (2017). Role of database management systems (DBMS) in supporting information technology in sector of

education. *International Journal of Science and Research (IJSR)*, *6*(5). https://doi.org/1462-1466. 10.21275/ART20173499

Sanchez-Caballe, A., Gisbert-Cervera, M, & Esteve-Mon, F. (2020).The digital competence of university students: A systematic literature review, *Aloma*, *38*(1). https://doi.org/10.51698/aloma.2020.38.1.63-74

Sarker, M. N. I., Wu, M., Cao, Q., Alam, G. M., & Li, D. (2019). Leveraging digital technology for better learning and education: A systematic literature review. *International Journal of Information and Education Technology*, *9*(7), 453-461. https://doi.org/10.18178/ijiet.2019.9.7.1246

Sulianta, F., & Supriatna, N. (2019). Digital Content Model Framework Based on Social Studies Education. *International Journal of Higher Education*, *8*(5), 214-220.

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761-1768. https://doi.org/10.1016/j.energy.2006.11.010

Urbancikova, N., Manakova, N., & Ganna, B. (2017). Socio-economic and regional factors of digital literacy related to prosperity. *Quality Innovation Prosperity*, *21*(2), 124-141. https://doi.org/10.12776/QIP.V21I2.942

Vitola, L., & Sierra, J. E. (2021). Cluster Analysis to Determine the Relationship Between Use Of ICT and ICT Literacy. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *18*(10), 3455-3471.

Williams, G. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media.