Science in Information Technology Letters
Vol. 4., No. 2, November 2023, pp. 90-96
ISSN 2722-4139
http://pubs2.ascee.org/index.php/sitech

# Web log augmented analytics and extraction for e-learning environment

Nur Azizah Mohammad Mokhtar [a,1], Sarina Sulaiman [a,2], Andri Pranolo [b,3,*]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
[b] Informatics Department, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[1] nnurazizahmokhtar@gmail.com; [2] sarina@utm.my; [3] andri.pranolo@tif.uad.ac.id
* Corresponding Author

ARTICLE INFO

**Article history**
Received October 7, 2023
Revised October 22, 2023
Accepted November 16, 2023

**Keywords**
Augmented analytics
E-learning
Classification
Artificial neural network
Support vector machine

ABSTRACT

E-Learning is a commonly used platform by most institutions, especially during the pandemic Covid-19. E-learning services include viewing, submitting, and uploading files, attempting quizzes, viewing forums, and downloading files. The data store in the servers grow on par with the increment of users in e-Learning@UTM every semester. As a result, the data have become extremely huge. These web log data can be used in augmented analytics to find meaningful insights. The web log data extracted are the log files of the history engagement of users and students' grades. Data obtained are used in augmented analytics to study the pattern of the data and insights into meaningful information. This research focuses on classification of data through predictive analytics. Hence, predictive models are required. To prove a better outcome, building the model consists of three types of algorithms; Decision Tree, Artificial Neural Networks and Support Vector Machine which are used and compared. After extracting data from e-learning, the first step in building a predictive model is to do data collection, data pre-processing, and data transformation. These three classifiers use the pre-processed data and split the data into training and test sets afterwards. Each classifiers techniques are built and a confusion matrix is applied as a performance measurement to summarise the performance of a classification algorithm respectively.



## 1. Introduction

E-Learning has become a common platform used by all institutes and universities, especially after the pandemic Covid-19 starts, and has been fully utilized since the establishment of Online Distance Learning (ODL). E-learning@UTM provides excellent teaching and learning services to all UTM lecturers, instructors, and students [1]. Different modules such as assignments, forums, quizzes, upload, and others are used simultaneously by each user. The module's retrieval can be recognized from the log data stored in servers. The web log data extracted from e-learning can be used in finding insights. For example, classify the students by predicting their performance. Machine Learning (ML) techniques in augmented analytics can give meaning to these data and provide valuable information for learning improvement [2]–[4].

Augmented analytics is a way of studying patterns of data and finding meaningful information from it [5]. It gives a method to predict the outcome of the data. It assists in detecting patterns of data and making a prediction of it about future events [6]. To predict students' performance based on the data

extracted from e-learning@UTM, predictive models are built. Hence, classification techniques are used. Classification technique is a technique categorised under a supervised learning algorithm using label data type.

Students' performance is classified into three categories, "high", "middle" or "low" according to UTM standard marks. There are three main classification algorithms used in this research, Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM). Each algorithm is to build the predictive model using training and testing data sets. Results obtained are compared and prove the best algorithm to use in predicting students' performance. Therefore, this research objectives are to determine the best method for applying augmented analytics; to identify the classification of web log data extracted from e-learning using augmented analytics; and to validate the accuracy of the augmented analytics applied to extracted web log data. This paper is structured as follows: Section 2 represents the proposed algorithm, the dataset used, research design and implementation. Meanwhile, Section 3 describes the framework for predictive models. The result analysis is in Section 4. Conclusion is made on the last section, Section 5.

## 2. Method

### 2.1. Supervised Learning Classification Algorithm

To do analytics, ML and Educational Data Mining (EDM) are useful methods. Classification is a method used by many researchers in extracting hidden patterns or relationships between attributes and elements in a dataset [7]. This paper focuses on three classification methods DT, ANN, SVM.

A DT is similar to a flowchart. It consists of two entities, nodes, and leaves. The nodes split the data into leaves. Leaves in the decision tree are its decision-making or the final outcomes. The top node is called a root node. Classification and regression trees are the two different forms of decision trees. Classifier generate a decision based on a certain sample of data either uni-variate or multivariate predictors [8]. This algorithm uses information gain ratio, entropy, and pruning. In an information gain, information that counts as useful and necessary information is when the outcome is not the expected outcome [9], [10]. Otherwise, it is not new information. Entropy determines the level of impurity of a specifically labeled dataset [11].

ANN is an interconnection structure between nodes and different layers of neurons [12]. Weighted links connect the neurons [13]. ANN is a technique that implies the studies of the brain and the nervous system of three layers; input layer, hidden layer and output layer [14], [15]. Data is first passed to the input layer. Each input is an independent variable for one attribute. Hidden layer performed an operation named weighted sum and activation function that converts the neuron's weighted sum. Output layer produces output value either continuous, binary or categorical form.

SVM is a supervised learning algorithm that assists in analysing and recognizing the patterns within data as a classification and regression prediction tool [16]–[18]. The algorithm of SVM allows the selection of the lines or known as a separator in a manner that accurately mimics the underlying function in the target space [19]. SVM mapped the data to a high-dimensional feature space for categorization. [20] by separating data between categories until the line is shaped into hyperplanes.

### 2.2. Data Collection and Preparation

Dataset used for this research is from one of the sections consists of students from different section TIS subject. The dataset is obtained from 3 different classes where each class consists about 30-35 students. The data consist of 95 students from 3 different sections are combined into one log files. There are two web log data that are retrieved from the e-learning@UTM, log files of history of e-learning engagement and log files of student's grades. In the log files of student's engagement, a total of 114624 of student's engagement for all modules which highest and lowest individual engagement are 1039 and 298 respectively. Meanwhile, grade log files recorded raw data of history of activity in e-learning student's

grade respectively. In simpler terms, log files of student's grades contains the grades student's obtained in assessment made in the e-learning. For example, quizzes and assignments submitted in the e-learning. Each log files has different and/or same attributes and need to be cleaned and filtered later on in pre-processing. The data obtained need to be filtered and cleaned beforehand by removing noise data, unrelated data. The attributes of merged data sets are as presented in Table 1. There are two types of attributes, features and label. Features are used as predictors and label is the predicted output. From now onwards, "Result" attribute is known as label and the rest are features.

**Table 1.** Attributes in Web Log Files

| Attributes | Description |
|---|---|
| Course view | No of times course viewed |
| Resource view | No of times resource viewed |
| Assignment view | No of assignments viewed |
| Assignment submit | No of assignments submit |
| Forum view and submit | No of times forum viewed and submit |
| File and folder | No of file and folder accessed |
| Quiz view | No of times quiz viewed |
| Quiz submit and review | No of quizzes submitted and reviewed |
| Total | Total marks of course |
| Result | Student's category (High/Average) |

### 2.3. Research Design and Implementation

Referring to Table 1, there are 10 attributes used in building the predictive model. The model is to predict whether the students fall under the category of "High", "Average" or "Low" based on their engagement in e-learning. However, in the dataset used, no students fall under the category "Low" subjected to remove this class from the classification. Fig 1 shows the process diagram of a predictive model.

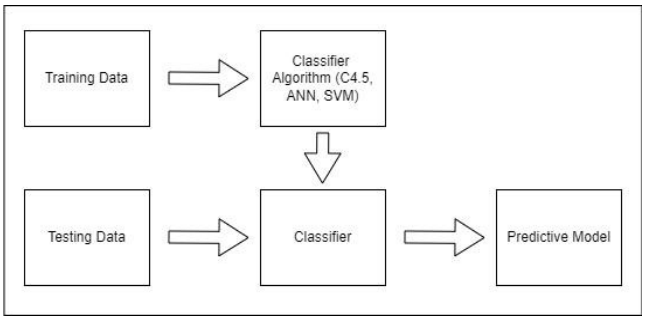


**Fig. 1.** Process Diagram of Predictive Model

Prediction model consists of two steps, to predict a class labelled whether it is nominal or discreet and to classify data to construct a model using Decision Tree, ANN, and SVM. The data is discreet as this research focus on classification. There are two types of data, training data and testing data. Training data is where the data is applied to the training set by applying the training data on the three classification algorithm. Meanwhile testing data, also known as validation data is the data that is used to evaluate the accuracy of the model. To build predictive model, each algorithm split the pre-processed data into train-test set with a ratio of 80:20.

Decision tree algorithm uses entropy to determine the level of impurity of each features split. The nodes will stop splitting when the leaves are pure or equals to 0. It is important to fit the model beforehand to avoid over fitting. The last steps in decision tree model is to predict its outcome. According to [9], the formula of entropy and gain ratio equation are as shown in (1) and (2) respectively.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (2)$$

ANN model has three layers; input, hidden, and output layer. Input layers are the features define after load the data set into the program and label is used for output layer. The input layer consists of 9 features and the output layer has 1 label. There are two hidden layers, 10 and 6 neurons at the first and second hidden layer respectively The output layer has only one neuron as it is a binary classification (0,1).

SVM model applies mathematical approach to draw a straight line (hyperplanes) between categories. The mathematical representation of its formula is in (3). This algorithm classify data by mapping data points and find the hyper-plane to divide data into classes. There are three types of kernel, linear, RBF, and polynomial. Linear is the best option because the dataset can be separated using a single line. The data has been classified into categories with labels of 0 and 1.

$$K(X_1, X_2) = exp\left(-\frac{\|X_1 - X_2\|^2}{2\sigma^2}\right) \quad (3)$$

### 2.4. Paremeters testing

This section explains the steps of parameter testing method. The parameters used in this research is the web log data extracted from the e-learning environment. The web log data contains student's activity in e-learning and applied into augmented analytics to find insights. The data is classify into two, "High" and "Average" by splitting into training and testing with ratio of 80:20. All three predictive models are built based on this parameter to predict students' performance. Accuracy of each model is calculated and validate using confusion matrix at the end of the model. The results obtained will be taken and compared to decide which algorithm gives better outcome

The performance of all three algorithms are evaluated based on confusion matrix. There are four aspects to validate the performance of the models using confusion matrix; precision, recall, f1-score, and accuracy. Table 2 shows the formulation of each metric and the description. Further elaboration is discussed on section 4.

**Table 2.** Performance Model Evaluation

| Attributes | Formula | Description |
|---|---|---|
| Accuracy | $\frac{T_P + T_N}{T_P + T_N + F_P + F_N}$ | Ratio of the total number of correct classifications to the total number of all classifications |
| Precision | $\frac{T_P}{T_P + F_P}$ | The proportion of positive test cases that is correctly classified |
| Recall | $\frac{T_P}{T_P + F_N}$ | The fraction of actual positive test cases that were properly classified |
| F1-Score | $\frac{2T_P}{2T_P + F_P + F_N}$ | The mean of precision and recalls |

## 3. Results and Discussion

In the predictive model, the attributes are assigned as features and labels. All attributes are features except for the result. There are two categories of in-label that students fall under, "High" and "Average". In the classification model, 80% of the data set are randomly chosen as training data and another 20% for testing data. The effectiveness of the model is evaluated using confusion matrix. Table 3 shows the result of performance measurement for each model.

**Table 3.** Requirements of Evaluation

| Algorithm | Category | Precision (%) | Recall (%) | F1-Score (%) | Accuracy (%) |
|---|---|---|---|---|---|
| DT | Average | 60 | 75 | 67 | 84 |
| | High | 93 | 87 | 90 | |
| ANN | Average | 100 | 40 | 57 | 84 |
| | High | 81 | 100 | 90 | |
| SVM | Average | 100 | 50 | 67 | 89 |
| | High | 88 | 100 | 94 | |

The model with the highest accuracy is the SVM model with a percentage of 89%. Meanwhile, the other two models, C4.5 and ANN have the same accuracy percentage, 84%. Fig. 1 shows the visualization of the performance.

Based on Table 4, the precision of "Average" class for ANN and SVM are equal, which means, all students that are in an "Average" class are correctly predicted as average. ANN and SVM model once again obtain the same result in recall with 100% for "High" class. Recall indicates these two models have the best classifier ability in finding all the positive instances be they predicted correctly or incorrectly. Moreover, all three models obtain an f1-score above 85% for "High" class. F1-score balances precision and recall to provide a balance measure of model performance. Hence, these models can be conclude as models with balanced measure for "High" class. Even though, the difference in f1-scores for "High" and "Average" classes are quite a gap, they are still above 50% which can be said as pretty decent measurement for "Average" class.

All three models achieve an overall accuracy of 84% to 89%. The difference between models are their performances. C4.5 performs exceptionally well for both classes while SVM perform better for the "High" class. The ANN model obtains an extraordinary precision for "Average" class but very low recall for that same class. It is a struggle to choose which model with the best algorithm as it is not too distinguishable. Thus, the choice depends on the specific requirements and the importance of prediction such as balance performance, correct amount of instances, and precision measurement as shown in Table 4.

**Table 4.** Requirements of Evaluation

| Attributes | Formula | Description |
|---|---|---|
| Balanced Performance | SVM | Has better precision, recall, f1-score, and accuracy |
| Correct Instances | SVM, ANN | Achieved perfect recall percentage for "High" class |
| Precision Measurement | SVM | Has better overall precision percentage compare to other models |

If the priority is balanced performance for both classes, SVM is the most suitable model as it consists an overall high percentage of precision, recall, f1-score, and accuracy. If the number of correct instances are the priority, SVM and ANN are the better choice as both model have high percentage of recall for "High" class. However, if the precision is the top priority, SVM is the most suitable model because for both "Average" and "High" class, the precision are comparatively high. Overall, from the requirements above, SVM model meets all the conditions. To conclude, the best model cannot be chosen based on the accuracy alone and results of all three models are relatively similar. That being said, the most suitable model to choose based on various perspective is SVM.

## 4. Conclusion

This research focuses on augmented analytics and extraction of web log data from the e-learning environment. Therefore, Section I emphasizes the introduction, objectives, scopes, problem statement and significance of this research. There are three objectives of this research. The first objective is achieved in Section 2. E-learning services such as assignments, quizzes and forums present different modules. The number of retrievals for each module can be recognized from log data stores in servers. The log data stored in e-learning can be extracted and used in augmented analytics to find insights into the data. There is much proof from the work of other researchers that augmented analytics has a wide spectrum. Thus, three classification algorithms; DT, ANN, and SVM are introduced to build a predictive model. The best method to apply augmented analytics based on the web log data is to classify data using a predictive model. The second objective is achieved in Section III and IV. It describes how the raw data can be cleaned and pre-processed before applying it to the predictive model. It also explains the steps of building the model including the type of data used. It also describes the implementation of the proposed solution in Chapter 3. The data set is classified into two groups, "Average" and "High" which represent the student's final results based on their engagement in e-Learning@UTM. Three predictive models, DT, ANN, and SVM are built and presented in the classification report. Each model is required to validate its performance to ensure a low error percentage. Validation of each model is part of the third objective. The last objective is achieved in Section V. Each model provides the confusion matrix to calculate the performance measurement; recall, precision, accuracy, and f1-score. By comparing all three models from different perspectives, the SVM model has the highest accuracy with a percentage of 89%. Furthermore, the model has the highest percentage for recall, precision, and f1-score in both the average and high groups. Hence, the SVM model is the most suitable model for augmented analytics.

### Acknowledgment

The authors wish to thank all experts, sources, and individuals involved for contribution of this research. Massive gratitude to the Universiti Teknologi Malaysia for the experience given in writing this paper.

### References

[1] Z. Hussaini, "Prediction of students' performance in e-learning environment of UTMSPACE program," pp. 9-15, 2017. [Online]. Available at: http://eprints.utm.my/79092/.

[2] S. J. Qin and L. H. Chiang, "Advances and opportunities in machine learning for process data analytics," *Comput. Chem. Eng.*, vol. 126, pp. 465–473, Jul. 2019, doi: 10.1016/j.compchemeng.2019.04.003.

[3] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Comput. Oper. Res.*, vol. 119, p. 104926, Jul. 2020, doi: 10.1016/j.cor.2020.104926.

[4] A. Christopoulos, S. Mystakidis, N. Pellas, and M.-J. Laakso, "ARLEAN: An Augmented Reality Learning Analytics Ethical Framework," *Computers*, vol. 10, no. 8, p. 92, Jul. 2021, doi: 10.3390/computers10080092.

[5] J. Richardson, R. Sallam, K. Schlegel, A. Kronz, and J. Sun, "Magic Quadrant for Analytics and Business Intelligence Platforms," *Gart. Magic Quadr.*, no. February, pp. 1–60, 2020, [Online]. Available at: https://bpmtraining.net/wp-content/uploads/2020/10/gartner-magic-quadrant-for-analytics-and-business-intelligence-platforms-feb-2020.pdf.

[6] E. M. Onyema *et al.*, "Prospects and Challenges of Using Machine Learning for Academic Forecasting," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–7, Jun. 2022, doi: 10.1155/2022/5624475.

[7] A. Fattah, M. M., P. S., and T. F., "A Decision Tree Classification Model for University Admission System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 10, 2012, doi: 10.14569/IJACSA.2012.031003.

[8] W. Lin, Y. Lu, and C. Tsai, "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Syst.*, vol. 36, no. 1, p. e12335, Feb. 2019, doi: 10.1111/exsy.12335.

[9] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[10] N. A. Priyanka and D. Kumar, "Decision tree classifier: a detailed survey," *Int. J. Inf. Decis. Sci.*, vol. 12, no. 3, p. 246, 2020, doi: 10.1504/IJIDS.2020.108141.

[11] A. Malekian and N. Chitsaz, "Concepts, procedures, and applications of artificial neural network models in streamflow forecasting," in *Advances in Streamflow Forecasting*, Elsevier, 2021, pp. 115–147, doi: 10.1016/B978-0-12-820673-7.00003-2.

[12] M. O. Okwu and L. K. Tartibu, "Artificial Neural Network," in *Studies in Computational Intelligence*, vol. 927, Springer Science and Business Media Deutschland GmbH, 2021, pp. 133–145, doi: 10.1007/978-3-030-61111-8_14.

[13] S. Walczak and N. Cerpa, "Artificial Neural Networks," *Encycl. Phys. Sci. Technol.*, pp. 631–645, Jan. 2003, doi: 10.1016/B0-12-227410-5/00837-1.

[14] N. A. Al-Sammarraie, Y. M. H. Al-Mayali, and Y. A. Baker El-Ebiary, "Classification and diagnosis using back propagation Artificial Neural Networks (ANN)," in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, Jul. 2018, pp. 1–5, doi: 10.1109/ICSCEE.2018.8538383.

[15] I. M. Nasser and S. S. Abu-Naser, "Predicting Tumor Category Using Artificial Neural Networks," *Int. J. Acad. Heal. Med. Res.*, vol. 3, no. 2, pp. 1–7, 2019, [Online]. Available at: https://philpapers.org/rec/NASPTC.

[16] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha, "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models," *J. Pet. Sci. Eng.*, vol. 200, p. 108182, May 2021, doi: 10.1016/j.petrol.2020.108182.

[17] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," in *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, Feb. 2019, pp. 24–28, doi: 10.1109/ISS1.2019.8908018.

[18] I. Ibrahim and A. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 10–19, Mar. 2021, doi: 10.38094/jastt20179.

[19] "How SVM Works," *IBM Documentation*, 2021. [Online]. Available at: https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works.

[20] W. Wiguna and D. Riana, "Diagnosis Of Coronavirus Disease 2019 (Covid-19) Surveillance Using C4.5 Algorithm," *J. Pilar Nusa Mandiri*, vol. 16, no. 1, pp. 71–80, Mar. 2020, doi: 10.33480/pilar.v16i1.1293.