Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 2, March 2024, pp. 379∼388
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i2.3499

Educational Data Mining: Multiple Choice Question Classification in Vocational School

Sucipto[1], Didik Dwi Prasetya[2], Triyanna Widiyaningtyas[2]
[1]Universitas Nusantara PGRI, Kediri, Indonesia
[2]Universitas Negeri Malang, Malang, Indonesia

Article Info

Article history:

Received October 24, 2024
Revised November 19, 2024
Accepted March 11, 2024

Keywords:

Education data mining
Difficult index
Classification
Multiple choice questions
Vocational School

ABSTRACT

Data mining on student learning outcomes in the education sector can overcome this problem. This research aimed to provide a solution for selecting quality multiple choice questions (MCQ) using the results of students' mid-semester exams in vocational high schools using a Data Mining approach. The research method used was the Cross-Industry Standard Process for Machine Learning (CRISP-ML) model. Steps to assess the accuracy of analyzing the difficulty level of questions based on student profile data and midterm test results. The data used in this research were the findings of basic computer tests on mid-term exams in mathematics disciplines at vocational high schools. This research used several classification algorithms, including SVM, Naive Bayes, Random Forest, Decision Three, Linear Regression, and KNN. The results of evaluating the classification algorithm using the difficulty index approach model increased evaluation accuracy by 5% -10%.

Corresponding Author:

Didik Dwi Prasetya,
Department of Electrical Engineering and Informatics,
Universitas Negeri Malang, Malang, Indonesia,
Email: didikdwi@um.ac.id.

How to Cite:
S. Sucipto, D. Dwi Prasetya, and T. Widiyaningtyas, "Educational Data Mining: Multiple Choice Question Classification in Vocational School", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 2, pp. 379-388, Mar, 2024.

1. INTRODUCTION

It is undeniable that advances in science and technology, especially information and communication technology in education, continue to grow. ICT has made learning activities more manageable than a few decades ago. ICT is increasingly proving successful in many aspects of life, one of which is the field of education [1, 2]. With the advent of information and communication technology (ICT), educational institutions have been encouraged to use artificial intelligence to improve learning efficiency and flexibility, resulting in better academic outcomes. One of the implementations is the existence of Computer-based Exams [3]. In vocational secondary education, computer-based tests (CBT) are used to evaluate exams. Exams at school are essential for students who determine the value of student learning outcomes. Exams improve students' abilities [4]. Evaluating exams using CBT is used to overcome the capacity of many test takers above 200 participants. CBT tests also solve the logistical problems of running traditional pencil and paper exams. This includes things like requesting places, scoring exams, monitoring, grading on time, and handling conflicts [5–7].

Today, data mining has attracted many researchers from various fields, including education [8**?** , 9]. Especially when it comes to education, obtaining student data makes it easier to make decisions about student academic performance. In education, predicting student performance is very important because it can determine who performs well and poorly in the future after college. There are currently few studies analyzing students' experiences in computer-based examinations and related educational outcomes in an ever-evolving educational landscape where computer-based assessments are growing in popularity [10, 11]. Research on student learning outcomes focusing on multiple-choice questions has been extensively researched, including by Julaeha, who discussed building web-based three-tier multiple-choice. This study assesses students' understanding of evolutionary characters, phylogenetic relationships, clade concepts, MRCA, sister groups, tree topology, and evolutionary sequence. The study concluded that web-based tests effectively measure students' thinking trees [12]. Another study was conducted by Olen Kim, discussing multiple-choice questions using machine learning. This study proposes an algorithm for the readjustment of difficulty in LMS-based online evaluation. This algorithm uses logistic regression classification algorithms and reference thresholds. Group evaluation showed that the average score improved in most groups compared to the difficulty of questions based on correct answers [13]. Another study conducted by Yu discussed the potential of education using questions written by students using online multiple-choice questions. The results showed that students who received explanations from students had a better attitude towards the subject matter compared to students who received explanations from teachers [14].

In other studies, analyzing exam results in algorithmic, statistical, and data mining approaches, among others, model compression reduces the size of trained language models (PLM). Still, it often causes performance degradation, especially for low-resource tasks such as answering multiple-choice questions. The end-to-end reptile meta-learning (ETER) approach addresses this problem by integrating target adjustments into meta-training. ETER significantly improves compressed PLM performance and outperforms baseline performance on a wide range of data sets [15]. Research conducted by Archana discusses DItractor GENeration (DIGEN), which aims to produce distractors for Multiple Choice Questions (MCQ) in the technical domain. Distractors were evaluated using Item Response Theory, which showed promising results in reducing errors and cumbersome effort in manual MCQ construction [16]. Wasim conducted another study regarding question labels for classification models. This paper introduces a data transformation approach called Label Power Set with logistic regression (LPLR) for multi-label biomedical question classification. It compares its performance with Structured SVM, Boltzmann Machine Limited, and copy transformation-based logistic regression [17]. Another study on classification models was conducted by Yu H. This model uses a knowledge distillation-based two-way Transformer encoder and a convolutional neural network model (TinyBERT-CNN) to classify the intent of question sentences in text. This model can effectively classify the intent of question sentences and provide technical support for future question-and-answer systems [18].

Statistical analysis and data mining approaches do not concentrate much on the difficulty index of computer-based testing multiple choice questions. Xue took several approaches to investigate the effectiveness of transfer learning in predicting parameters of difficulty and response time. Results show that transfer learning improves item difficulty prediction when response time is used as an additional task, with entire item sections being important for predicting response time [19]. Another approach Qiiu took proposes a Document-enhanced Mindfulness-based Neural Network (DAN) framework to predict the difficulty of multiple-choice questions in medical exams. The experimental results showed the effectiveness of the proposed framework [20]. Kumar's difficult index approach was conducted by analyzing 90 multiple-choice questions (MCQs) for 150 physiology students. The study recommends item analysis for all MCQ-based assessments to determine validity and reliability [21]. In previous research, various methods such as difficulty indices, statistics, and data mining were used to analyze the difficulty of multiple-choice questions. However, in this research, a combined analysis was carried out by combining the three. The difference between this research and previous research is the use of various data mining models with the addition of a difficulty index variable.

This project created a predictive data mining model using the SVM, Naive Bayes, Random Forest, Decision Three, Linear Regression, and KNN algorithms. Using a data mining model with a difficulty index variable aims to analyze multiple choice questions based on student profiles and subject midterm exam results. Mathematics to be able to select good-quality questions.

2. RESEARCH METHOD

The researchers use an action research methodology based on the developed research methodology. A technique called action research places a strong emphasis on social activity. It addresses social problems often seen in institutions such as hospitals, factories, schools, etc. The data mining process uses classification techniques with fish diseases as the target. The study expands on previous research that established CRISP-ML, a methodology that helps achieve the level of interpretation stakeholders need to achieve successful real-world solutions. Based on the advantages of CRISP-DM, CRISP-ML overcomes the shortcomings in handling interpretability [22–24]. Figure 1 represents the procedure of this study.

Figure 1. CRISP-ML Diagram of research procedure

The sequence of research starts from the level of observation. At this level, several cases or problems are found, then the problems and cases are formulated, and the purpose of the research is indicated. This degree also stands out in planning during the study. In this study, researchers wanted to analyze midterm exam data through multiple-choice questions to predict student learning outcomes. The data comes from the CBT data extract of midterm exams in mathematics subjects in vocational high schools; the data is useful for predicting student learning outcomes with an educational data mining approach in the future.

The second step of the research procedure is to ”reflect.” In this step, the data used in the study is obtained. Descriptions of feature selection, data selection, class balancing, cleaning data (noise reduction, data imputation), Feature engineering (data construction), Data augmentation, and Data standardization are also discussed. From this step, the data amounted to 8520 data recorded student answers from multiple choice questions.

The next step is a ”plan,” where in this step, the selection of variables and the research applied several classification algorithms, such as SVM, Nave Bayes, Random Forest, Decision Three, Linear Regression, and KNN [25], [26], to determine the accuracy of the difficulty level analysis of questions based on mid-term test results. The final stage is the implementation of an algorithm to determine the quality of the elucidation data. This stage evaluates the model by validating model performance with confusion matrix testing. Accuracy, precision, gain (also called sensitivity), and F Measure range from 0 to 1, and the calculation Equations are shown at (1), (2), (3) and (4) respectively [27, 28].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Data mining is analyzing data from multiple points of view and turning it into meaningful information. Technically, data mining can be defined as the process of finding patterns or correlations of thousands of fields drawn from large relational databases. The education data mining model is presented in Figure 2 in the flowchart.

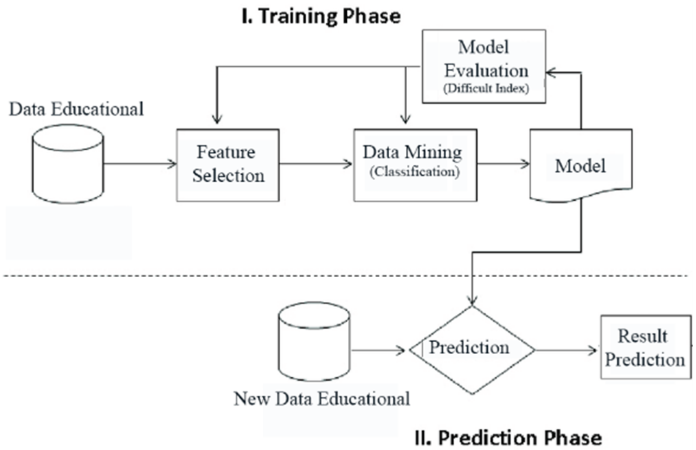

Figure 2. Model education data mining

The model stage in Figure 2 starts from educational data obtained in the CBT database that has been carried out. Data Preprocessing with steps Raw data is often imperfect, including missing, duplicate, or anomalous data. Data cleansing includes filling in missing data, deleting invalid data, and performing additional processing. Variable coding, data normalization, or the creation of additional features are examples of data transformation, including converting raw data to a format that can be used for analysis. Afterward, proceed to the feature selection stage, where the most relevant subset of features is selected for this process analysis. Selecting this feature helps reduce data dimensions and increase prediction rates.

Data mining involves several algorithms, including SVM, Nave Bayes, Random Forest, Decision Three, Linear Regression, and KNN. The data mining model was evaluated, and a difficulty index was taken from the student score variable of each question. Once the model is appropriate, the data mining model produces the expected predictions. The final model is tested with new data and generates new classes. The old model and optimization results were compared to confusion matrix testing.

3. RESULT AND ANALYSIS

Analyzing student learning outcomes data from multiple choice questions starts from collecting data taken from midterm exam results in vocational school mathematics subjects in one of the Kediri City, Indonesia schools. The variables used consist of 12, which are presented in Table 1. The variables are retrieved from the CBT database. The number of records is 8520 records. The question model uses multiple choice questions of 20 data. The total number of examinees was 426 students.

Table 1. Variable of Research

Variables	Describe
user_id	Student ID
th_akademic	Year Academic
grup_class	Student Class
grup_name	Student Departmen
grup_rombel	Student Goup
start_test_time	Start Test Time
difference_time	difference time
tessoal_soal_id	Question ID
tessoal_order	Question Order
dif_index	Difficult Index
bloom	Teacher Definition Taxnomoni
tessoal_nilai	Test Result per question

The variable ”user_id” is the student’s test id. Each student does math problems in CBT with a processing time of 70 minutes and 20 questions. The questions shown to each student were randomized. Random sequence variables are taken on the variable ”tessoal_order.” The exam’s start time is stored in the variable ”start_test_time,” and the difference in the time of doing each question in the form of a second time is stored in the variable ”difference_time.” The teacher performs a taxonomy definition of the role to map the difficulty level of the question stored in the ”bloom” variable. The student’s answer to each question is stored on ”tessoal_nilai.” After defining the variables used at the ”reflect” stage. The next stage is cleaning data. The result data is presented in Figure 3. The result is cleaning data on blank data and standardizing data from nominal to integer data. Data captured.

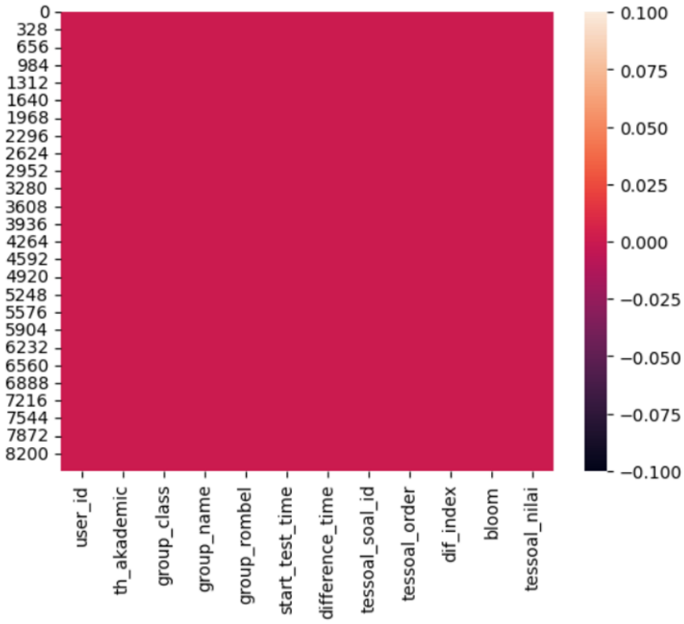

Figure 3. Dataset IsNull

In the next stage, carry out the ”plan” stage by determining the data analysis process using 2 model scenarios as in the groove presented in Fig 2. The first scenario is analyzed with several algorithms and evaluated using a confusion matrix. The second scenario is to analyze the data more carefully. The analysis is based on the fundamental statistical analysis in Table 2 and Table 3. It is known to be more specific that the data has the proximity of the overall correct and false answer results data such as Table 2.

Table 2. Statistics Global Questions Answer

Total Record	8520
Global Answer	
Correct	4510
Wrong	4010

The number of right and wrong answers has a difference of 500 records, so it will tend to have difficulty predicting the classification of student learning outcomes, [29]. Based on the statistical analysis of more specific answers presented in Table 3, the second step obtained better results regarding the level of difference in student answer results. The approximate results in Table 3 are included in the difficult index analysis with the Formula (5).

$$DI = \frac{NCR}{\Sigma NA} \times 100\% \tag{5}$$

DI : Difficult Index
NCR : Number of Correct Responses
NA : Number of Attempts

Formula (5) calculates how difficult a question or problem is based on the percentage of people who answer it correctly. A low-difficulty index indicates a more manageable problem, while a high-difficulty index indicates a difficult problem [19, 20, 30]. In education, the difficulty index of a question or test is usually calculated as the ratio of the number of people who answered the question correctly to the total number of people who tried to solve it. This index is usually indicated in percentage form. Implementing both scenarios in the classification algorithm uses the configuration in Table 4. The configuration compares 80:20 testing data to obtain training data of 6816 and testing data of 1704.

Table 3. Statistics Questions Answer

Question ID	**Σ students**	**Answer**	
		Correct	**Wrong**
24787	426	192	234
24788	426	307	119
24789	426	269	157
24790	426	297	129
24791	426	241	185
24792	426	247	179
24793	426	181	245
24794	426	313	113
24795	426	60	366
24796	426	205	221
24797	426	201	225
24798	426	147	279
24799	426	143	283
24800	426	279	147
24801	426	165	261
24802	426	266	160
24803	426	284	142
24804	426	228	198
24805	426	199	227
24806	426	286	140

Table 4. Standard Preprocessing Data

test_size	0.20
random_state	42
sklearn.preprocessing	StandardScaler

After determining the division of the dataset [27], the data was randomized with parameters random_state 42. The random seeds that divide the data into training and testing data are assigned with these parameters. Data normalization function using "StandardScaler." To perform normalization, StandardScaler subtracts the average of each feature and divides it by the standard deviation. As a result, each feature will have an average of 0 and a variance of 1. The results of the analysis of student learning outcomes are based on MCQ in Table 5, Table 6, and Table 7. In Table 5, the initial scenario with some algorithms shows an average accuracy of 59%. The difference between Precision, Recall, F1-score, and Accuracy is 1-2%. The random forest algorithm shows the best accuracy with 65% accuracy and the lowest algorithm on the SVM algorithm with 55% accuracy.

Table 5. Experimental results of MCQ in situation1

Algorithm	**Precision**	**Recall**	**F1-score**	**Accuracy**
SVM	0.55	0.53	0.49	0.55
KNN	0.58	0.58	0.58	0.58
RF	0.64	0.64	0.64	0.65
NB	0.59	0.57	0.54	0.56
LR	0.56	0.55	0.54	0.56
DT	0.64	0.63	0.63	0.64
Average	0.59	0.58	0.57	0.59

Table 6. Experimental results of MCQ in situation2

Algorithm	Precision	Recall	F1-score	Accuracy
SVM	0.65	0.64	0.63	0.64
KNN	0.62	0.62	0.62	0.62
RF	0.65	0.64	0.64	0.65
NB	0.63	0.63	0.63	0.63
LR	0.63	0.63	0.63	0.63
DT	0.67	0.66	0.65	0.66
Average	0.64	0.64	0.63	0.64

Table 7. Difference Experimental results of MCQ

Algorithm	Precision	Recall	F1-score	Accuracy
SVM	0.1	0.11	0.14	0.09
KNN	0.04	0.04	0.04	0.04
RF	0.01	0	0	0
NB	0.04	0.06	0.09	0.07
LR	0.07	0.08	0.09	0.07
DT	0.03	0.03	0.02	0.02
Average	0.05	0.05	0.06	0.05

The findings in the results of this research are based on the results of increasing accuracy obtained in Table 6 on the addition of variables resulting from the difficulty index model with an average accuracy of 64% with the best accuracy using the decision tree algorithm and the lowest accuracy using the KNN algorithm with an accuracy of 62%. The difference between Precision, Recall, F1 score, and Accuracy is 1%. Table 7 represents the average increase of each algorithm, around 5%. On the Random Forest algorithm, there is no improvement [31]. The most significant improvement in the SVM algorithm was %. At the highest result on the DT algorithm, the increase was only 2%. The increase in accuracy is good because the data used has a reasonably close distribution value calculation and a multi-class number of questions with a total of 20 [32].

4. CONCLUSION

In this study, we explore methods of analyzing student learning outcomes to determine the quality of questions in mathematics subjects in vocational school students based on predictions of student answers to CBT data. First, we performed a detailed statistical analysis of the supporting variables in the CBT results. The results showed that students had almost harmonious results between questions that were successfully answered right and wrong on all questions tested. We propose a more specific analysis model by adding variables from the results of difficult index analysis from the results of student answers to each question to determine the quality of student learning outcomes. In the future, we will continue to study accuracy improvement models with specific classification algorithms to improve the accuracy of student learning outcome predictions based on MCQ data.

5. ACKNOWLEDGEMENTS

This research is part of a doctoral dissertation research. The authors thank the State Universitas Negeri Malang for the support that made this important research feasible and effective.

6. DECLARATIONS

AUTHOR CONTIBUTION

Sucipto: Conceptualization, Methodology, Writing - Original Draft, Writing. Didik Dwi Prasetya: Writing- Review & Editing, Supervision. Triyanna Widiyaningtyas: Investigation, Data Curation, Data Validation. All authors have read and agreed to the published version of the manuscript.

FUNDING STATEMENT

This research received no specific grant from any funding agency.

COMPETING INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] P. Garca-Alcaraz, V. Martnez-Loya, J. L. Garca-Alcaraz, and C. Snchez-Ramrez, "The Role of ICT in Educational Innovation," 2019, pp. 143–165, https://doi.org/10.1007/978-3-319-93716-8_7. [Online]. Available: http://link.springer.com/10.1007/978-3-319-93716-8{_}7

[2] D. D. Prasetya, A. P. Wibawa, T. Hirashima, and Y. Hayashi, "Digital Content Model for E-Learning System in Higher Education," in *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*. IEEE, Oct. 2019, pp. 192–196, https://doi.org/10.1109/ICEEIE47180.2019.8981461. [Online]. Available: https://ieeexplore.ieee.org/document/8981461/

[3] D. D. Prasetya, H. W. Herwanto, and W. S. G.I, "Design of Web-Based Interactive Whiteboard Application to Facilitate Online Learning," *Proceedings of Vocational Engineering International Conference*, vol. 5, pp. 118–122, 2023. [Online]. Available: https://proceeding.unnes.ac.id/veic/article/view/2819

[4] J. Bergner, J. J. Filzen, and M. G. Simkin, "Why use multiple choice questions with excess information?" *Journal of Accounting Education*, vol. 34, pp. 1–12, Mar. 2016, https://doi.org/10.1016/j.jaccedu.2015.11.008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S074857511500086X

[5] C. Zilles, M. West, G. Herman, and T. Bretl, "Every University Should Have a Computer-Based Testing Facility," in *Proceedings of the 11th International Conference on Computer Supported Education*. SCITEPRESS - Science and Technology Publications, 2019, pp. 414–420, https://doi.org/10.5220/0007753304140420. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0007753304140420

[6] R. Efendi, L. S. Lesmana, F. Putra, E. Yandani, and R. A. Wulandari, "Design and Implementation of Computer Based Test (CBT) in vocational education," *Journal of Physics: Conference Series*, vol. 1764, no. 1, p. 012068, Feb. 2021, https://doi.org/10.1088/1742-6596/1764/1/012068. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1764/1/012068

[7] R. Carpenter and T. Alloway, "Computer Versus Paper-Based Testing: Are They Equivalent When it Comes to Working Memory?" *Journal of Psychoeducational Assessment*, vol. 37, no. 3, pp. 382–394, Jun. 2019, https://doi.org/10.1177/0734282918761496. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0734282918761496

[8] C. Xu, G. Zhu, J. Ye, and J. Shu, "Educational Data Mining: Dropout Prediction in XuetangX MOOCs," *Neural Processing Letters*, vol. 54, no. 4, pp. 2885–2900, Aug. 2022, https://doi.org/10.1007/s11063-022-10745-5. [Online]. Available: https://link.springer.com/10.1007/s11063-022-10745-5

[9] M. B. Priyantono, M. Ahnan, M. A. Widhianto, and D. D. Prasetyo, "Optimasi Sistem Pelabelan Topik Skripsi menggunakan Algoritma Naive Bayes dengan Pendekatan Design Thinking," *Jurnal Edukasi dan Penelitian Informatika (JEPIN)*, vol. 8, no. 1, p. 168, Apr. 2022, https://doi.org/10.26418/jp.v8i1.50702. [Online]. Available: https://jurnal.untan.ac.id/index.php/jepin/article/view/50702

[10] M. Zheng and D. Bender, "Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance," *Medical Teacher*, vol. 41, no. 1, pp. 75–82, Jan. 2019, https://doi.org/10.1080/0142159X.2018.1441984. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0142159X.2018.1441984

[11] J. M. Harley, N. M. Lou, Y. Liu, M. Cutumisu, L. M. Daniels, J. P. Leighton, and L. Nadon, "University students' negative emotions in a computer-based examination: the roles of trait test-emotion, prior test-taking methods and gender," *Assessment & Evaluation in Higher Education*, vol. 46, no. 6, pp. 956–972, Aug. 2021, https://doi.org/10.1080/02602938.2020.1836123. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/02602938.2020.1836123

[12] S. Julaeha, T. Hidayat, and N. Y. Rustaman, "Development of web-based three tier multiple choice test to measure student's tree thinking; try out," *Journal of Physics: Conference Series*, vol. 1521, no. 4, p. 042024, Apr. 2020, https://doi.org/10.1088/1742-6596/1521/4/042024. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1521/4/042024

[13] E. Kim, "A study on the difficulty adjustment of programming language multiple-choice problems using machine learning," *Journal of Korea Society of Industrial Information Systems*, vol. 27, no. 2, pp. 11–24, 2022, https://doi.org/10.9723/JKSIIS.2022.27.2.011. [Online]. Available: http://dx.

[14] F.-Y. Yu and C.-Y. Chen, “Student- versus teacher-generated explanations for answers to online multiple-choice questions: What are the differences?” *Computers & Education*, vol. 173, p. 104273, Nov. 2021, https://doi.org/10.1016/j.compedu.2021.104273.

[15] M. Yan and Y. Pan, “Meta-learning for compressed language model: A multiple choice question answering study,” *Neurocomputing*, vol. 487, pp. 181–189, May 2022, https://doi.org/10.1016/j.neucom.2021.01.148. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231221016441

[16] A. P. Kumar, A. Nayak, M. S. K., S. Goyal, and Chaitanya, “A novel approach to generate distractors for Multiple Choice Questions,” *Expert Systems with Applications*, vol. 225, p. 120022, Sep. 2023, https://doi.org/10.1016/j.eswa.2023.120022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417423005249

[17] M. Wasim, M. N. Asim, M. U. Ghani Khan, and W. Mahmood, “Multi-label biomedical question classification for lexical answer type prediction,” *Journal of Biomedical Informatics*, vol. 93, p. 103143, May 2019, https://doi.org/10.1016/j.jbi.2019.103143. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046419300619

[18] H. Yu, C. Liu, L. Zhang, C. Wu, G. Liang, J. Escorcia-Gutierrez, and O. A. Ghoneim, “An intent classification method for questions in ”Treatise on Febrile diseases” based on TinyBERT-CNN fusion model,” *Computers in Biology and Medicine*, vol. 162, p. 107075, Aug. 2023, https://doi.org/10.1016/j.compbiomed.2023.107075. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482523005401

[19] K. Xue, V. Yaneva, C. Runyon, and P. Baldwin, “Predicting the Difficulty and Response Time of Multiple Choice Questions Using Transfer Learning,” in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 193–197, https://doi.org/10.18653/v1/2020.bea-1.20. [Online]. Available: https://www.aclweb.org/anthology/2020.bea-1.20

[20] Z. Qiu, X. Wu, and W. Fan, “Question Difficulty Prediction for Multiple Choice Problems in Medical Exams,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, Nov. 2019, pp. 139–148, https://doi.org/10.1145/3357384.3358013. [Online]. Available: https://dl.acm.org/doi/10.1145/3357384.3358013

[21] D. Kumar, R. Jaipurkar, A. Shekhar, G. Sikri, and V. Srinivas, “Item analysis of multiple choice questions: A quality assurance test for an assessment tool,” *Medical Journal Armed Forces India*, vol. 77, pp. S85–S89, Feb. 2021, https://doi.org/10.1016/j.mjafi.2020.11.007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0377123720302379

[22] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Mller, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, Apr. 2021, https://doi.org/10.3390/make3020020. [Online]. Available: https://www.mdpi.com/2504-4990/3/2/20

[23] I. Kolyshkina and S. Simoff, “Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach,” *Frontiers in Big Data*, vol. 4, May 2021, https://doi.org/10.3389/fdata.2021.660206. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2021.660206/full

[24] I. N. Switrayana, D. Ashadi, H. Hairani, and A. Aminuddin, “Sentiment Analysis and Topic Modeling of Kitabisa Applications using Support Vector Machine (SVM) and Smote-Tomek Links Methods,” *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 2, no. 2, pp. 81–91, Sep. 2023, https://doi.org/10.30812/ijecsa.v2i2.3406. [Online]. Available: https://journal.universitasbumigora.ac.id/index.php/IJECSA/article/view/3406

[25] D. D. Prasetya, A. Prasetya Wibawa, and T. Hirashima, “The performance of text similarity algorithms,” *International Journal of Advances in Intelligent Informatics*, vol. 4, no. 1, p. 63, Mar. 2018, https://doi.org/10.26555/ijain.v4i1.152. [Online]. Available: http://ijain.org/index.php/IJAIN/article/view/152

[26] H. Hairani and T. Widiyaningtyas, “Augmented Rice Plant Disease Detection with Convolutional Neural Networks,” *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 1, pp. 27–39, Feb. 2024, https://doi.org/10.29407/intensif.v8i1.21168. [Online]. Available: https://ojs.unpkediri.ac.id/index.php/intensif/article/view/21168

[27] I. Saifudin and T. Widiyaningtyas, “Systematic Literature Review on Recommender System: Approach, Problem, Evaluation Techniques, Datasets,” *IEEE Access*, vol. 12, pp. 19 827–19 847, 2024, https://doi.org/10.1109/ACCESS.2024.3359274. [Online]. Available: https://ieeexplore.ieee.org/document/10415424/

[28] T. Widiyaningtyas, I. Hidayah, and T. B. Adji, “Comparing User Rating-Based Similarity to User Behavior-Based Similarity in Movie Recommendation Systems,” in *2022 International Conference on Electrical and Information Technology (IEIT)*. IEEE, Sep. 2022, pp. 52–58, https://doi.org/10.1109/IEIT56384.2022.9967884. [Online]. Available: https://ieeexplore.ieee.org/document/9967884/

[29] A. Parhizkar, G. Tejeddin, and T. Khatibi, “Student performance prediction using datamining classification algorithms: Evaluating generalizability of models from geographical aspect,” *Education and Information Technologies*, vol. 28, no. 11, pp. 14 167–14 185, Nov. 2023, https://doi.org/10.1007/s10639-022-11560-0. [Online]. Available: https://link.springer.com/10.1007/s10639-022-11560-0

[30] J. L. D’Sa and M. L. Visbal-Dionaldo, “Analysis of Multiple Choice Questions: Item Difficulty, Discrimination Index and Distractor Efficiency,” *International Journal of Nursing Education*, vol. 9, no. 3, p. 109, 2017, https://doi.org/10.5958/0974-9357.2017.00079.4. [Online]. Available: http://www.indianjournals.com/ijor.aspx?target=ijor:ijone{&}volume=9{&}issue=3{&}article=024

[31] S. H. Hasanah and E. Julianti, “Analysis of CART and Random Forest on Statistics Student Status at Universitas Terbuka,” *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 6, no. 1, pp. 56–65, Feb. 2022, https://doi.org/10.29407/intensif.v6i1.16156. [Online]. Available: https://ojs.unpkediri.ac.id/index.php/intensif/article/view/16156

[32] Sucipto, Kusrini, and Emha Luthfi Taufiq, “Classification method of multi-class on C4.5 algorithm for fish diseases,” in *2016 2nd International Conference on Science in Information Technology (ICSITech)*. IEEE, Oct. 2016, pp. 5–9, https://doi.org/10.1109/ICSITech.2016.7852598. [Online]. Available: http://ieeexplore.ieee.org/document/7852598/