Performance Evaluation of College Students' Google Classroom Engagement Using Data Mining Techniques

Serafin C. Palmares [1], April Joy Abara-Palmares [1], Jan Carlo T. Arroyo [2,3], Allemar Jhone P. Delima [2]

[1]*Northern Negros State College of Science and Technology, Sagay City, Philippines*
[2]*College of Information and Computing Studies, Northern Iloilo State University, Estancia, Iloilo, Philippines*
[3]*College of Computing Education, University of Mindanao, Davao City, Davao del Sur, Philippines*

Abstract **– The purpose of this study is to assess how well college students use Google Classroom as a useful and informative teaching and learning tool. The survey method was utilized in the study to measure student involvement in Google Classroom. This study's sample population included 292 college students from Northern Negros State College of Science and Technology. Algorithms such as Random Forest (RF), C4.5, and Naive Bayes (NB) were utilized with three of the most crucial techniques, such as 60% split, training set, and 8-fold cross-validation, for performing analysis on the student data. After analyzing different metrics for performance (Correctly Classified Instances, FP Rate, ROC Area, F-Measure, TP Rate, Recall, Precision, Time taken to build model, Mean Absolute Error, Root Mean Squared Error, Root Relative Squared Error, Relative Absolute Error) by various algorithms for data mining, the researchers determined which algorithm performs better than others on the student dataset gathered, allowing the researchers to make a recommendation for future improvement in students' Google Classroom engagement.**

DOI: 10.18421/TEM122-48
https://doi.org/10.18421/TEM122-48

Corresponding author: Serafin C. Palmares,
Northern Negros State College of Science and Technology, Sagay City, Philippines.
Email: spalmares@nonescost.edu.ph

Received: 18 February 2023.
Revised: 09 May 2023.
Accepted: 11 May 2023.
Published: 29 May 2023.

Keywords **– C4.5, data mining, decision tree, Google Classroom, Naive Bayes, random forest, Weka.**

1. Introduction

The COVID-19 pandemic has impacted every country in the world, causing widespread health, social, and economic disruptions across the globe. By implementing constraints on community interaction that are implemented physically and improving community strategy, COVID-19 can be stopped from spreading in its entirety [1]. The physical separation strategy, however, can impede the rate of development in several areas, including social, educational, and economic ones. The government's policy in education is to implement study-at-home activities to relocate learning and instruction from classrooms to homes [2]. The role of the instructor has also been altered, with online education transforming the dynamics of teamwork and personalization in the classroom. Therefore, the success of the e-learning process is dependent on the efficacy of the engagement and communication that occurs throughout the class [3]. In an effort to curb the spread of COVID-19, traditional classrooms had to be replaced by online alternatives, but many educators and learners were unprepared for the change [4].

In this study, the researchers analyzed how well students at Northern Negros State Colleges use Google Classroom to supplement their education. In this study, the researchers addressed the following issues:

1. Does using Google Classroom make students more interested in and dedicated to their studies?
2. How helpful are the skills and information acquired in a Google Classroom setting?
3. Does Google Classroom allow students to receive feedback that is valuable to them?

4. Does Google Classroom facilitate the ability to talk to students about their work?
5. What do students think are the biggest problems or limitations when it comes to using Google Classroom?

In this study, the researchers compared C4.5, Naive Bayes (NB), and Random Forest (RF), three different data mining methodologies suitable for classification algorithms. One or more decision trees, each of which has been trained using training data samples, make up the random forest algorithm. The accuracy of the results can be boosted since each child node used to trigger a node is chosen at random using Random Forest methods. This technique can be used to generate a tree structure with such a root node, leaf nodes, and internal nodes. Naive Bayes, on the other hand, is a classification method that is based on Bayes' Theorem and assumes that all features that predict the target value are unrelated to one another. It computes the probabilities for each class before selecting the one with the highest likelihood [5]. According to the Naive Bayes classifier, the features that are utilized to predict the target are unrelated to one another and do not affect one another [6]. The C4.5 algorithm is one of the data mining algorithms included in the classification groups. C4.5 methods are used to generate a decision tree. The decision tree produced by the algorithm C4.5 can describe and reflect the findings of important data investigations, making it easier to extract information or knowledge from the data [7]. This work is based on a survey done on students at NONESCOST during the 1st semester of the academic year 2022–2023, in which, in addition to demographic data, survey responses were collected. This analysis was carried out following the testing and training of the algorithms, allowing conclusions to be drawn on possible determinants of student performance evaluation.

2. Related Works

Review of Techniques for Data Mining to Predict Student Performance

Predicting students' outcomes becomes more challenging as educational databases expand in size. In Malaysia, there is not yet a framework in place to evaluate and track student outcomes. In general, two factors account for this phenomenon. When it comes to predicting academic achievement in Malaysian schools, not nearly enough study has been done on existing prediction approaches. It is advised that researchers conduct a thorough literature analysis on employing data mining technology to forecast student performance to enhance education. The major purpose is to teach the techniques of data mining that are used to predict student progress. This paper focused on how to develop a prediction algorithm to find the most valuable data features of a student. Data mining with an educational focus has the potential to increase student growth and performance dramatically. It has the potential to benefit schools, classrooms, and educators [8].

A Comprehensive Analysis of the Research on Forecasting Student Achievement Using Data Mining and Learning Analytics Approaches

Predicting student academic progress has piqued the curiosity of many educators. While learning is expected to improve both learning and teaching, predicting student outcomes has yet to be investigated. This study examined studies from 2010 to November 2020 to provide a fundamental grasp of the intelligence strategies utilized for forecasting student accomplishment in a setting where academic achievement is rigorously measured by student learning outcomes. Several electronic bibliographic databases were searched, including Scopus, Science Direct, ACM, Google Scholar, and others. Additionally, the researchers combined and assessed 62 relevant research studies from three perspectives: (1) how learning outcomes are anticipated; (2) predictive and prescriptive models developed to anticipate learning outcomes; and (3) dominating elements impacting student results. To consolidate and report the major findings, the best approaches for conducting a comprehensive review of literature, such as PRISMA and PICO were used. Students' places in their classrooms and averaged test scores were used to evaluate learning results. To identify pupils' outcomes, instructors typically employ regressed and regulated machine learning models. Finally, the most visible markers of academic results were students' interactive learning activities, term assessment grades, and academic feelings [9].

Students' Views on the Usefulness of Google Classroom as a Technological Learning and Teaching Instrument

The study aimed to evaluate how effective Google Classroom is as an LMS, specifically by assessing the perceptions of college students who use it. To achieve this, a survey research approach was utilized to gather data about students' perspectives on the platform's effectiveness in supporting their learning. Descriptive statistics were used to analyze the information. The research showed that when students used Google Classroom, they were more motivated to take an active role in their own education, had better access to educational resources, and were more focused on their studies.

Students are unable to fully utilize Google Classroom due to poor network conditions, which in turn delays the submission of student work. Instructors can help their students succeed academically by using Google Classroom in their existing repertoire of tried-and-true classroom techniques. Google Classroom's online quizzes and assignments can be used to get more students involved in their instructors' lessons on educational technology. If the school keeps its network stable, pupils will have no trouble handing in their work on time using Google Classroom [10].

Methods for Forecasting Student Achievement in an E-Learning Environment Using Data Mining and Machine Learning

The usage of internet technology in education transformed from a traditional to a virtual environment recognized as the ELE or referred to as E-Learning Environment. This is a brand-new area of study for researchers. Using the ELE platform in conjunction with Face to face training improved students' comprehension and performance. All academic institutions shifted to online learning during the COVID-19 epidemic, which increased the value of the online learning environment. Realistic and reliable assessments of students' performance on ELE present the biggest problem for educational institutions. Forecasts of actual students' progress will be valuable to instructors or course coordinators early in the course when students demand attention and support. Throughout the last decade, education-related data mining has proven to be an effective approach for identifying useful knowledge and structures from massive educational datasets. It comprises using data mining (DM) techniques for big datasets. In contrast, machine learning (ML), classification, and regression algorithms are more successful and accurate in anticipating student performance today. The type of characteristics used, the size of a sample, and the variety of the dataset all have a significant impact on prediction accuracy and efficiency. ML approaches such as AdaBoost, SVC, ANN, Random Forest (RF), Decision Tree (DT), and k-NN are employed as important methods for forecasting students' success on the ELE datasets based on regression and classification analyses [11].

3. Methodology

In this section, the researchers go over the techniques they employed to accomplish the study's goals. First, they go over the data collection to come up with the dataset that was used for this study. Second, they tackle files and attributes that have been imported into the data mining tool.

After that, they go over the Data Mining and classification algorithms they utilized with the WEKA tool to get the findings.

Data Collection and Dataset

The researchers gathered and examined the survey responses from 292 undergraduate students using online Google forms. Also, the researchers conducted a survey to collect data from students in different NONESCOST colleges and programs. According to the study, it is evident that women make up most of the sample population, as indicated in Figure 1. This is plainly evident from the survey's huge portion (62%) of female respondents vs. its low number (38%) of male respondents.

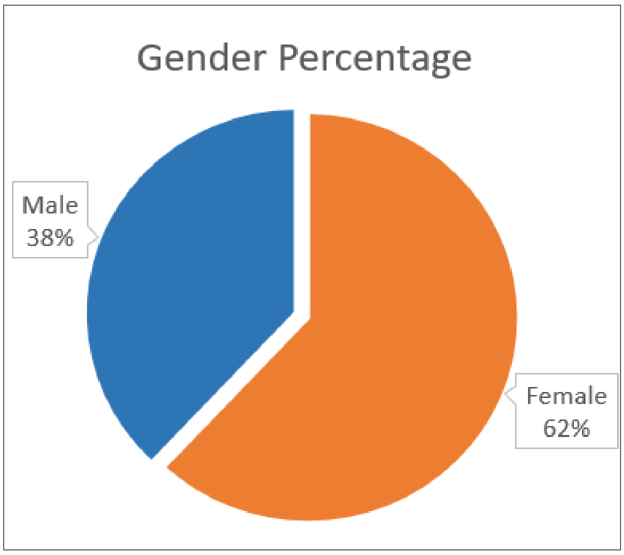

Figure 1. Gender Percentage

File and Attributes

The dataset is stored in the Microsoft Excel CSV file format and was collected using the reliable data mining program WEKA. For the goal of analyzing college students' performance evaluation in their Google Classroom engagement, the researchers constructed the questionnaire with 3 demographic profile attributes and 20 survey questions for the functions and usefulness of Google Classroom grouped into 5 sets of questions and 1 target attribute value, then derived and selected 9 attributes for the datasets. The survey questionnaire's statements served as the foundation for all the qualities. Table 1 provides a detailed overview of all the attributes that have been utilized. Figure 1 presents the findings of the survey conducted. That sums up how many pupils there were and how they responded to each attribute. The researchers analyzed and anticipated the student's performance based on the survey results.

Table 1. Attributes Definition

ATTRIBUTE	DESCRIPTION
Course	Student’s Course
Year Level	Student’s Year level
Gender	Student’s Gender
Q1	Does using Google classroom make students more interested in and dedicated to their studies?
Q2	How helpful are the skills and information acquired in a Google Classroom setting?
Q3	Does Google Classroom allow students to receive feedback that is valuable to them?
Q4	Does Google Classroom facilitate the ability to talk to students about their work?
Q5	What do students think are the biggest problems or limitations when it comes to using Google Classroom?
Effective Tool?	Is Google Classroom an effective tool?

Modeling

Modeling stages depend on the model algorithm utilized in the study. This time, investigate modeling using three models: the Naive Bayes method, the random forest, and C4.5. WEKA, a data mining tool, is used to process the algorithm models [12]. For this purpose, the researchers explained each method as well as some of the algorithms. Using data from 292 students, they examined the performance evaluations in their Google Classroom engagement.

1. *Naïve Bayes Algorithm*

The Bayes method, which uses the likelihood function as a prerequisite, is an effective machine learning technique based on data training. According to Gata et al., [13], a statistic on the classification that can be used to foretell the possibility that a group will have members is the Naive Bayes Classifier. The Bayesian classification, which is based on the Bayes theorem, is named after Thomas Bayes, a mathematician, and minister of the Presbyterian Church in the United Kingdom [14]. A rule in the Naive Bayes technique can be used to determine how likely a class is. The Naive Bayes algorithm offers a way to integrate the chance or opportunity advance with the phrases likely to be a calculation that can be used to determine the probabilities of that happening in any case. Bayes' rule applies to the Standard form of the theorem as follows:

$$GainAverage = H(T) - Hsaving(T)$$

2. *Random Forest algorithm*

A single decision tree derivative is developed using the process known as Random Forest (RF). RF techniques consist of one or more trees or decision trees, where each tree has finished training on sample data. Because the evocation of each child node by the other nodes is random, the Random Forest (RF) techniques can improve the accuracy of the outcomes [15]. This method develops a tree structure with leaf nodes, internal nodes, and parent nodes using random data attributes and applicable laws. The root node is an informal name for the top node of a decision tree, often known as the root. An internal node is a branching node that has at least two outputs, as opposed to the terminal node or leaf node, which is the final node and has no output but has only one input. The following equations can be used to determine how valuable information gain and entropy are:

$$Entropy(Y) = -\sum_i p(c \mid Y) \log 2\, p(c \mid Y)$$

Where Y is the set of case and p (c | Y the Y value) is the proportion of class c.

Information Gain

$$(Y, a) = Entropy(Y) \sum_{v \in Value(a) \frac{|Yv|}{|Ya|}} Entropy(Yv)$$

Where values (a) represent all possible values in a set of cases. YV is a sub class of the Y class v relating to class a. Yes, are all the values that correspond to a.

The selection of attributes as nodes, whether root or internal nodes, is based on the attribute with the highest information gain.

$$SplitInformation(S, A) = \sum_{\frac{c}{i}=1} \left(\frac{|Si|}{|S|}\right) \log 2 \left(\frac{|Si|}{|S|}\right)$$

Where the split information (S, A) is the value of the input variable entropy estimation of S that have class c and/Si///S/ is the probability of class I attribute.

$$GainRatio(S, A) = \frac{InformationGain(S, A)}{SplitInformation(S, A)}$$

3. *C4.5 Algorithm*

The decision tree algorithm, also known as C4.5, is an algorithm with a concept on the strategy of divide-and-conquer for a classification procedure that is a problem.

A decision tree could be created by using rules that have been acquired in C4.5 [16] essentially in three steps: establishing the data of the node; deciding on training (the selected node is the one with the least value of the results of an entropy search); and finally, creating the decision tree itself. The decision tree is created using the algorithm C4.5, which includes numerous steps like preparing the training data, finding, and calculating the entropy before searching each entropy class, and calculating the amount of the gain as well as the averaged gain.

Calculate Entropy

$$H(X) = \sum_{j} -pj \log_2(pj)$$

Calculate the value of the gain and the average gain.

$$GainAverage = H(T) - Hsaving(T)$$

Evaluation

Weka collected the findings from the accuracy value after running the NB, RF, and C4.5 algorithms in a data mining tool so that it could determine the confusion matrix of each approach. Additionally, the three most crucial procedures, including training set, percentage split (70%), and 8-fold cross-validation, are being used. After analyzing various key performance indicators (Mean Absolute Error, Time to build Classifier, Root Relative Squared Error, Precision, Recall, Relative Absolute Error, F-Measure, ROC Area, and Root Mean Squared Error) through various data mining algorithms, the researchers were able to determine which data mining algorithm is performing better than the other. Researchers can now develop guidelines for future improvements to student participation in Google Classroom as a result.

4. Results and Discussion

Three classification algorithms namely NB, RF, and J48 or C4.5 are tested on the dataset used in this study. These algorithms are used in conjunction with the three most important procedures, such as the training set, 70% split, and 8-fold cross-validation. Tables 2-4 provide the complete statistical findings. The accuracy of all classifiers has been compared, and it has been determined that Multi-Layer Perception technique performs best, with accuracy in the Training Set of classifiers using instances that have been Correctly Classified being 100% using Random Forest Algorithm, and for the Percentage Split (70%) and 8-folds Cross-validation of classifiers using instances that have been Correctly Classified, the C4.5 algorithm outperformed the others with 85%. In Tables 5-7, the accuracy rating for each algorithm is provided.

Table 2. Evaluation using the student's dataset in training set mode

Metrics for Performance	Random Forest	Naïve Bayes	C4.5
TP Rate	1.00	0.87	0.89
FP Rate	1.00	0.19	0.20
F-Measure	1.00	0.87	0.89
Recall	1.00	0.87	0.89
Precision	1.00	0.87	0.89
ROC Area	1.00	0.90	0.90
Time taken to test model on training data	0.02 sec	0.01 sec	0 sec
Relative absolute error	21.02%	38.32%	44.37%
Mean absolute error	0.0857	0.1562	0.1808
Root relative squared error	29.48%	70.65%	66.66%
Root mean squared error	0.133	0.3187	0.3007

Table 3. Evaluation using the student's dataset in Cross-validation (8-folds) mode

Metrics for Performance	Random Forest	Naïve Bayes	C4.5
TP Rate	0.82	0.85	0.86
FP Rate	0.32	0.24	0.26
F-Measure	0.82	0.85	0.85
Recall	0.82	0.85	0.86
Precision	0.82	0.84	0.85
ROC Area	0.85	0.88	0.81
Time taken to build model	0.02 sec	0 sec	0 sec
Relative absolute error	57.69%	40.27%	53.47%
Mean absolute error	0.2352	0.1641	0.2180
Root relative squared error	79.45%	72.85%	79.25%
Root mean squared error	0.3584	0.3286	0.3575

Table 4. Evaluation using the student's dataset in Percentage Split (70%) mode

Metrics for Performance	Random Forest	Naïve Bayes	C4.5
TP Rate	0.80	0.81	0.85
FP Rate	0.32	0.26	0.26
F-Measure	0.79	0.81	0.85
Recall	0.80	0.81	0.85
Precision	0.79	0.81	0.85
ROC Area	0.83	0.82	0.80
Time taken to test model on test split	0.01 sec	0 sec	0 sec
Relative absolute error	62.94%	50.29%	51.33%
Mean absolute error	0.2625	0.2097	0.2141
Root relative squared error	84.03%	82.22%	77.42%
Root mean squared error	0.3933	0.3848	0.3624

Table 5. Training Set of classifiers using instances that have been Correctly Classified

Mining Technique	Accuracy
Random Forest	**100%**
Naïve Bayes	86.99%
C4.5	89.04%

Table 6. Percentage Split (70%) of classifiers using instances that have been Correctly Classified

Mining Technique	Accuracy
Random Forest	79.55%
Naive Bayes	80.68%
C4.5	**85.23%**

Table 7. Cross-validation (8-folds) of classifiers using instances that have been Correctly Classified

Mining Technique	Accuracy
Random Forest	82.19%
Naive Bayes	84.59%
C4.5	**85.62%**

5. Conclusion

The findings of this study demonstrate how Google Classroom improves the process of teaching and learning. It is also reliable, useful, and efficient at captivating students' attention and creating learning opportunities. Google Classroom activities assist students in moving from inactive to active learning. Google Classroom offers online assessments that let students simply track their progress. Parents can easily and also, at their convenience, check on and track the academic development of their kids. This study supports the findings made by Nizal et al. [17] that Google Classroom improves teaching and learning. On the other hand, inadequate network conditions restrict students from utilizing Google Classroom properly, which leads them to complete their assignments after the due date. Therefore, it can be concluded that using Google Classroom will help both educators and students interact, collaborate, create coursework, grade students, and publish lecture notes considering the current circumstances of the worldwide COVID-19 pandemic, restricted movement, and social isolation.

The areas where students do not comprehend the material can also be the focus of investigation. Additionally, Google Classroom makes it simple to manage student data due to the increase in undergraduate enrollment. Records of students' online exams can always be recovered and are simple to access. The student's account can be used to resolve any issues with missing grades. With online tests and homework, pupils are not constrained by what they have been taught; students can use internet tools to seek additional information on a specific subject, giving them a deeper knowledge of what was covered in class. Instructors should consider utilizing Google Classroom as another channel of communication with their students to support in-person teaching and learning by quickly tracking students who skip assessments or submit them late. Academic institutions may consider upgrading the connection to accommodate all students on the campus to address the issue of the poor network impeding students' involvement.

References

[1]. Hassan, M. M., Mirza, T., & Hussain, M. W. (2020). A critical review by teachers on the online teaching-learning during the COVID-19. *International Journal of Education and Management Engineering*, *10*(8), 17-27. Doi: 10.5815/ijeme.2020.05.03

[2]. Pangestu, F. A. (2022). Application Of Correlation Data Mining the Covid-19 Pandemic With Student Achievement Level Using A Priori Method. *International Journal of Health Engineering and Technology*, *1*(3).

[3]. Al-Maroof, R. A. S., & Al-Emran, M. (2018). Students acceptance of google classroom: An exploratory study using PLS-SEM approach. *International Journal of Emerging Technologies in Learning*, *13*(6), 112-123. Doi:10.3991/ijet.v13i06.8275

[4]. Maity, B. (2021). Online Education and Learning Loss During the Covid-19 Pandemic. *International Journal of Multidisciplinary Educational Research*, *10*(11), 59–64. Retrieved from: http://s3-ap-southeast-1.amazonaws.com/ijmer/pdf/volume10/volume10-issue11(2)/12.pdf [accessed: 5 March 2023]

[5]. Islam, M. J., Wu, Q. M. J., Ahmadi, M., & Sid-Ahmed, M. A. (2008). Investigating the Performance of Naive-Bayes Classifiers and K-Nearest Neighbor Classifiers. In *Proceedings of 2008 11th IEEE International Conference on Computer and Information Technology* (1541–1546). Doi: 10.1109/iccit.2007.148.

[6]. Dong, L., Li, X., & Xie, G. (2014). Nonlinear methodologies for identifying seismic event and nuclear explosion using random forest, support vector machine, and naive bayes classification. *Abstract and Applied Analysis*, *2014*. Doi: 10.1155/2014/459137.

[7]. Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm. *International Journal on Computer Science and Engineering*, *2*(8), 2526-2534.

[8]. Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422. Doi: 10.1016/j.procs.2015.12.157.

[9]. Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences (Switzerland), 11*(1), 1–28. Doi: 10.3390/app11010237.

[10]. Hussaini, I., Ibrahim, S., Wali, B., Libata, I., & Musa, U. (2020). Effectiveness of Google classroom as a digital tool in teaching and learning: Students' perceptions. *International Journal of Research and Innovation in Social Science (IJRISS)*, *4*(4), 51-54.

[11]. Kumar Verma, B., Srivastava, N., & Kumar Singh, H. (2021). Prediction of Students' Performance in e-Learning Environment using Data Mining/ Machine Learning Techniques. *Journal of University of Shanghai for Science and Technology*, *23*(5), 596–593. Doi:10.51201/JUSST/21/05179

[12]. Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459. Doi: 10.11591/ijeecs.v9.i2.pp447-459

[13]. Gata, W., Priyandoko, D., Zainuddin, R., Aziz, A., & Susilo, B. (2019). Algorithm Implementations Naïve Bayes, Random Forest. C4.5 on Online Gaming for Learning Achievement Predictions. *In 2nd International Conference on Research of Educational Administration and Management (ICREAM 2018)* (1-9). Atlantis Press.

[14]. Leonard, T. H. (2014). A personal history of Bayesian statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(2), 80–115. Doi: 10.1002/wics.1293

[15]. Granitto, P. M., Gasperi, F., Biasioli, F., Trainotti, E., & Furlanello, C. (2007). Modern data mining tools in descriptive sensory analysis: A case study with a Random forest approach. *Food quality and preference*, *18*(4), 681-689. Doi: 10.1016/j.foodqual.2006.11.001

[16]. Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very Fast C4.5 Decision Tree Algorithm. *Applied Artificial Intelligence*, *32*(2), 119-137. Doi: 10.1080/08839514.2018.1447479

[17]. Nizal, I., Shaharanee, M., Jamil, J. M., Syamimi, S., & Rodzi, M. (n.d.). The Application of Google Classroom as a Tool for Teaching and Learning. *Journal of Telecommunication, Electronic and Computer Engineering*, *8*(10), 5-8.