A STUDY ON EFFECIENCY OF TEACHING PROBABILITY AND STATISTICS COMBINING WITH PROGRAMMING LANGUAGE R FOR STUDENTS IN UNIVERSITY OF INFORMATION AND COMMUNICATION TECHNOLOGY, THAI NGUYEN UNIVERSITY

Quach Thi Mai Lien
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 02/5/2024
Revised: 08/8/2024
Published: 08/8/2024

KEYWORDS

Teaching methodology
Teaching probability and statistics
Programming language
Probabilistic software
Integrative teaching

ABSTRACT

This study investigates the impact of teaching methodologies with probabilistic programming language R on student performance across various topics in probability theory and statistics. Integrative teaching strategies has shown impressive efficiency in enhancing student's critical thinking skills, creativity, and ability to solve complex problems. Here, two distinct groups of students, one (group B) have been taught these topics with traditional methodology and the other (group A) have been taught with R, were exposed to different instructional approaches. The performance of 1000 students from each group in University of Information and Communication Technology – Thai Nguyen University, was evaluated across Bloom's Taxonomy 6 levels, encompassing fundamental to advanced cognitive skills. Employing statistical analyses such as statistical descriptions, MANOVA and machine learning classification, we compared the performance of the two groups and conducted post-hoc analyses to identify specific factors contributing to performance disparities. Results indicate the superior performance of group A across multiple cognitive levels, underscoring the efficacy of methodology A. This research contributes to the ongoing discourse on optimizing teaching methodologies to enhance student learning outcomes in probability and statistics. Especially, it encourages to combine the teaching methodology of these subjects with a probabilistic programming language, such as R.

MỘT NGHIÊN CỨU VỀ SỰ HIỆU QUẢ CỦA VIỆC DẠY HỌC XÁC SUẤT VÀ THỐNG KÊ KẾT HỢP VỚI NGÔN NGỮ LẬP TRÌNH R CHO SINH VIÊN TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG, ĐẠI HỌC THÁI NGUYÊN

Quách Thị Mai Liên
Trường Đại học Công nghệ thông tin và Truyền thông - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 02/5/2024
Ngày hoàn thiện: 08/8/2024
Ngày đăng: 08/8/2024

TỪ KHÓA

Phương pháp giảng dạy
Giảng dạy Xác suất và Thống kê
Ngôn ngữ lập trình
Phần mềm xác suất
Dạy học tích hợp

TÓM TẮT

Nghiên cứu xem xét hiệu quả của việc giảng dạy tích hợp môn học xác suất thống kê với ngôn ngữ lập trình R. Dạy học tích hợp đang cho thấy hiệu quả ấn tượng trong việc củng cố kỹ năng tư duy phê bình, khả năng sáng tạo, kỹ năng giải quyết vấn đề của người học. Hai nhóm sinh viên được xem xét, một nhóm (nhóm B) được dạy các chủ đề này bằng phương pháp truyền thống và nhóm còn lại (nhóm A) được dạy các chủ đề này kết hợp với ngôn ngữ R. Kết quả học tập của 1000 sinh viên từ mỗi nhóm tại trường ĐH Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên được đánh giá theo thang đo 6 bậc Bloom, gồm các kỹ năng nhận thức cơ bản đến nâng cao. Sử dụng các phân tích thống kê mô tả, cũng như MANOVA và các phương pháp phân loại học máy, nghiên cứu đưa ra so sánh sự hiệu quả của các phương pháp dạy học trên hai nhóm sinh viên và tiến hành phân tích hậu kiểm để xác định các yếu tố cụ thể góp phần tạo ra sự chênh lệch về sự hiệu quả của mỗi phương pháp. Kết quả cho thấy hiệu quả vượt trội của nhóm A trên nhiều cấp độ nhận thức. Nghiên cứu này giúp khuyến khích giảng dạy tích hợp môn học xác suất thống kê với ngôn ngữ lập trình như R.

DOI: https://doi.org/10.34238/tnu-jst.10256

Email: qtmlien@ictu.edu.vn

1. Introduction

1.1. Teaching subjects of probability and statistics with R for students in University of Information and Communication Technology, Thai Nguyen University

Teaching probability and statistics using the programming language R offers a multitude of benefits that enrich students' learning experiences [1]-[3] and prepare them for success [4], [5] in data-driven fields. By leveraging R's powerful visualization capabilities [6], students gain a deeper understanding of abstract statistical concepts and probability distributions [3]-[5], as they can explore and interact with data through graphical representations [7]. Hands-on experience with real-world datasets in R fosters practical skills in data manipulation, analysis, and interpretation, empowering students to tackle complex problems and make informed decisions based on data-driven insights. R's emphasis on reproducibility and transparency instills good research practices, ensuring that students are well-prepared for academic and professional environments where rigorous documentation and validation of results are paramount.

Teaching probability with a probabilistic programming language, such as R, in University of Information and Communication Technology, (ICTU), Thai Nguyen University, provides students, who are mainly in Information technology majors, with a comprehensive and integrated approach to understanding probabilistic concepts and their applications in Information Technology (IT) - related domains [5], [7], [8]. The integration of probability and programming cultivates critical thinking, analytical reasoning, and problem-solving skills essential for success in Information technology - related professions and advanced studies. Since the academic year 2023-2024, ICTU has deployed the integrated teaching the course probability and statistics with programming language R, an open source software for statistical computing and graphic, in the undergraduate programs for IT engineering students. This plan is one of ICTU's endeavour to help IT engineering students improve their fundamental science knowledge with practical skills which connect to their major. It even benefits their professional skills; enables them attained the objectives and learning outcome of their training program as well when connecting probability and statistics, data sciences to programming skills; provides them chances to develop career skills needed to enter the high quality, modern labor market; helps them gain a competitive edge and are better prepared for advanced coursework and careers that require statistical programming and data analysis skills.

1.2. Problem and Dataset

The research aims to prove the efficiency of integrated teaching strategy applied on subjects of probability with a statistical programming language R by comparing two teaching strategies: integrated one and classical one (teaching subjects without a programming language). Using powerful techniques from data analysis, this point will be clear [7]-[9]. The data is collected for this research by author. The quizzes are designed for the purpose of the research with respect to each of 7 subjects and a Bloom's Taxonomy 6 level. These 7 subjects are constructed in curriculum of IT engineering program. The grades are collected for each student shown in a row, in scale 10, for each subject and the Boom's Taxonomy 6 level. These quizzes are designed to guarantee the Learning Outcome of the probability and statistics course in the training program for the IT engineering major of ICTU. Here, the rubrics of quizzes are implemented for each of 7 subjects of the course: Basic on probability measurement; Random variables and Probabilistic distribution, Expectation; Sampling theory and Inference Statistics; Parameter estimation; Hypothesis testing; Simple and Multiple linear regression; Logistic regression. Six levels of Bloom's Taxonomy as mentioned in subsection 2.1 for each subject are set up in each quiz. They include: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating. These quizzes are constructed on the basis of bank of problems for the course which has been edited by author with 14 years in reality of teaching experience for students in IT engineering major of ICTU. The dataset includes scores of 1000 students in ICTU who have been taught these 7

subjects integrated with R, labeled this group A, and scores of other 1000 students who have been taught these subjects without any programming language for probability integrated, but following the traditional strategy, labeled as group B. Dataset is available in Github at https://raw.githubusercontent.com/tiep1144/ ictu/main/ictu.csv'.

2. Materials and Methods

2.1. Six levels of Bloom's Taxonomy in teaching

Bloom's Taxonomy, developed by educational psychologist Benjamin Bloom in the 1950s, is a hierarchical framework that categorizes educational objectives and cognitive skills into six levels. This taxonomy provides educators with a structured approach to designing learning experiences, assessments, and instructional strategies that promote critical thinking and intellectual development in students. The six levels of Bloom's Taxonomy progress from lower-order thinking skills, such as remembering and understanding, to higher-order thinking skills, such as analyzing, evaluating, and creating. Each level represents a distinct cognitive process that students can engage in to demonstrate their understanding and mastery of learning material. By incorporating Bloom's Taxonomy into curriculum design and instructional practices, educators can facilitate deeper learning, foster intellectual growth, and prepare students for success in academic and real-world contexts [7], [9], [10]. These levels are termed by numbers from level_0 to level_5 in the dataset, and denoted from N1 to N6, including:

Remembering: This level involves recalling or recognizing information, concepts, or principles. It includes tasks such as recalling facts, defining terms, or identifying basic concepts.

Understanding: At this level, students demonstrate comprehension by explaining ideas or concepts in their own words, interpreting data, or summarizing information. They grasp the meaning of information and can explain it in their own terms.

Applying: Students apply acquired knowledge and understanding to new situations or contexts. They can solve problems, use procedures, or apply concepts in different scenarios.

Analyzing: This level involves breaking down information into its component parts, identifying patterns or relationships, and drawing inferences or conclusions. Students analyze data, dissect arguments, or differentiate between components.

Evaluating: At this level, students assess or judge the value or quality of ideas, methods, or materials based on specific criteria. They make judgments, critique arguments, or assess the validity of claims.

Creating: The highest level of Bloom's Taxonomy involves synthesizing information, generating new ideas or products, or designing solutions to problems. Students demonstrate creativity, innovation, and originality in their work.

2.2. Topics in probability and statistics in the research

This research focuses on 7 subjects of Probability and Statistics teaching for students in ICTU, including: Basic on probability measurement; Random variables and Probabilistic distribution, Expectation; Sampling theory and Inference Statistics; Parameter estimation; Hypothesis testing; Simple and Multiple linear regression; Logistic regression. These subjects are crucial in Probability and Statistics course taught for Engineering. They are termed in this research by number from 0 to 6, respectively, in the dataset.

In integrated teaching these subjects with R, R can be used for visualizations and simulations of random experiments, helping students grasp abstract concepts through practical, interactive methods; plotting distributions, computing expectations, and performing Monte Carlo simulations; demonstrating sampling methods, calculating confidence intervals, and performing bootstrap methods, providing hands-on experience with real data sets; assisting in point estimation, maximum likelihood estimation, and least squares estimation, making complex calculations more accessible; computing test statistics, p-values, and conducting power analysis,

ensuring students understand the practical applications of theoretical concepts; fitting simple and multiple linear regression models, creating diagnostic plots, interpreting coefficients, and modeling binary outcomes with logistic regression.

2.3. Methods

This paper presents some approaches to analyze the data using the tools of statistical visualization, MANOVA, logistic regression and machine learning models to make comparison between two groups A and B. These analyses, presented in Section 3, show the positive effect of the integrated teaching subjects of probability with the programming language R.

3. Results and Discussions

3.1. Data visualization

3.1.1. Bar plot

Bar plot (Figure 1) is a method to compare the mean grades between each group. From this plot, we observe that for each pair of a level and subject, the performance in group A is much higher than that in group B. It shows the superior of integrated teaching probability and statistics with R to the traditional teaching strategy which does not use integrated teaching the course with any programing language.

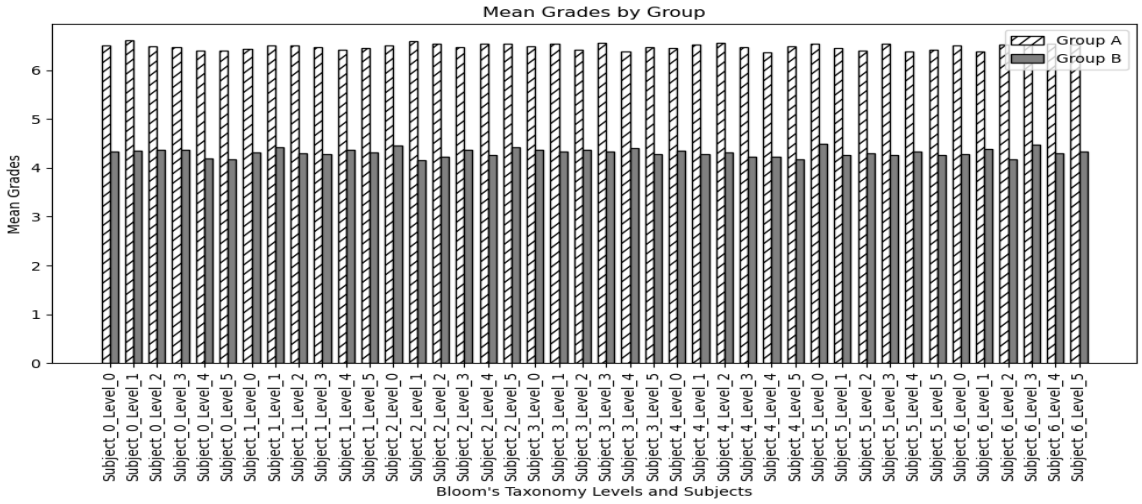

Figure 1. *Bar plot to compare the mean grades of each subject and each of Bloom's Taxonomy 6 levels between groups A and B*

3.1.2. Boxplot

This boxplot (Figure 2) compares the performance of students in group A and B corresponding to each subject and each of Bloom's Taxonomy 6 levels. This is shown by taking the mean grades between two groups. We see that, on each pair, the mean grades of group A always takes higher performance.

3.1.3. Heat map

This heat map (Figure 3) shows the effectiveness of integrated teaching strategy for each pair of subject and Bloom's Taxonomy level. It shows the most effectiveness on group A for the pairs subject0_N2, subject6_N4, subject2_N2, subject4_N3, and subject3_N4, while as this on group B for subject5_N1, subject6_N4, subject2_N1, subject2_N6, and subject1_N2. The integrated teaching strategy shows impressive effect in helping students understanding the basic concepts of probability, analyzing logistic regression, understanding concepts of random variables and

probability distributions. Moreover, the heat map shows the more concentration in high levels of Bloom's Taxonomy on group A than that on group B.

All of three data visualization methods mentioned above give a clear proof of the fact that the efficiency of integrated teaching strategy with R is superior than that of traditional teaching strategy, especially on subjects of random variable, probability distribution, basic concepts of probability, logistic regression modeling.

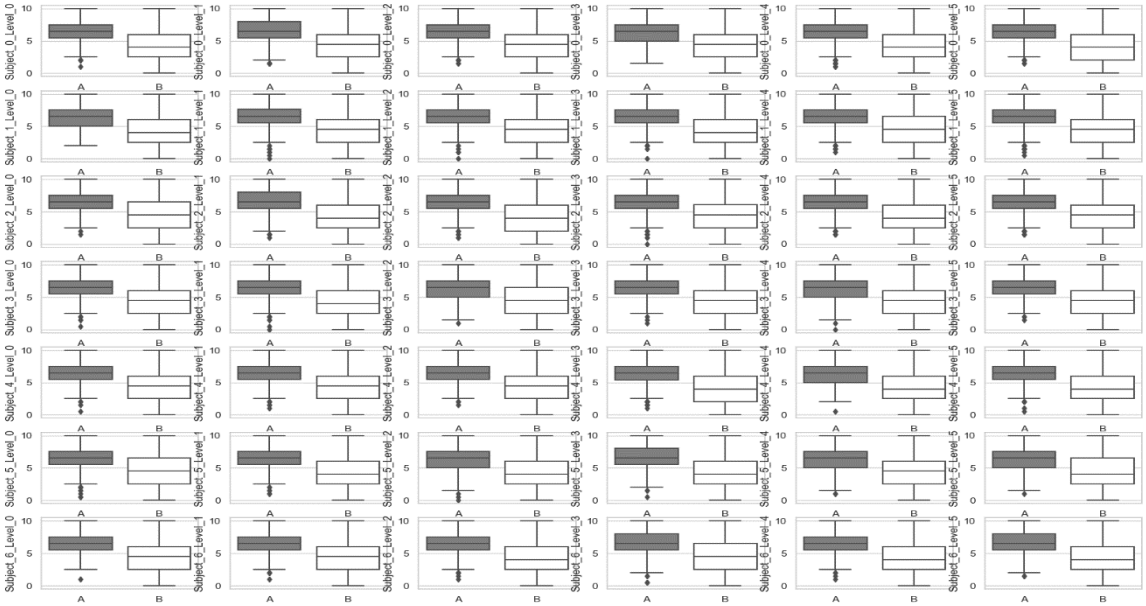

Figure 2. *Boxplot to compare the mean grades of each subject and each of Bloom's Taxonomy 6 levels between groups A and B*

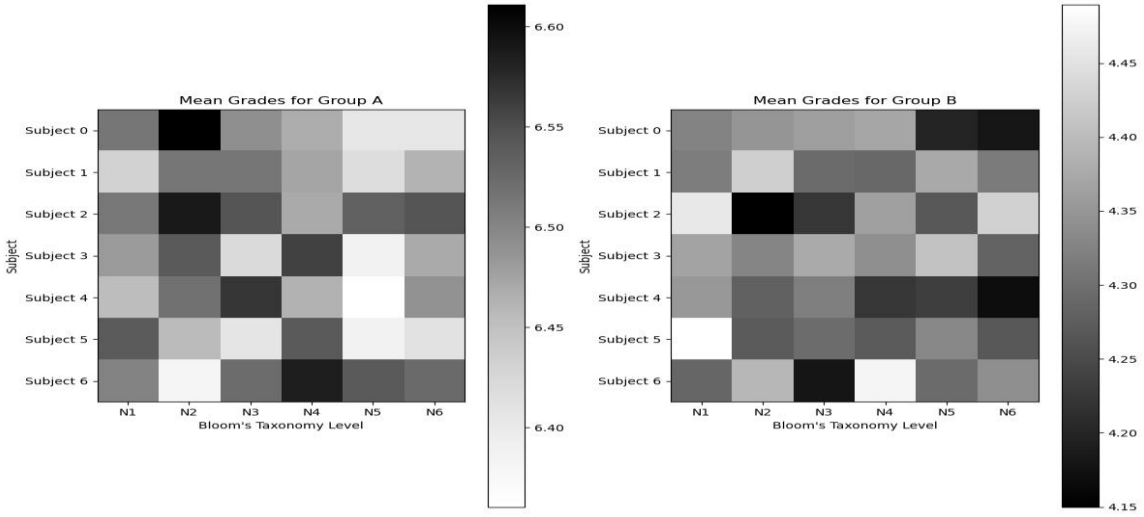

Figure 3. *The mean grades of each subject and each of Bloom's Taxonomy 6 levels of groups A and B*

3.2. Multivariate Analisis of Variation MANOVA

This section uses Multivariate Analysis of Variation (MANOVA) to analysis the significant difference shown in the performance of two groups A and B. A multivariate linear model are constructed with the package *statsmodels.multivariate.manova* in Python. In this model, the dependent variable is the label, taking on values A and B. The independent variables are pairs of

subject and level of Bloom's Taxonomy, including 42 variables, Subject_i_Level_j, for $0 \leq i \leq 5, 0 \leq j \leq 6$. MANOVA allows us to test whether there are statistically significant differences in the means of multiple dependent variables simultaneously. In the output shown in Figure 4, Wilks' lambda, Pillai's trace, Hotelling-Lawley trace, Roy's greatest root are indices, which shows a significant result, indicating that there are differences between the groups across the dependent variables, the performance of two groups A and B (P-values shown in the last column are significant) [11]. Tukey's Honestly Significant Difference (HSD) test, a post-hoc test, is also used to compare group means across multiple dependent variables, implemented from the package statsmodels.stats.multicomp with module pairwise_tukeyhsd. The result could be found in source code mentioned in Conclusion. In HSD test, each pair of subject and Bloom's Taxonomy level shows the superior difference in mean grade of group A comparing to that of group B.

```
                 Multivariate linear model
================================================================
----------------------------------------------------------------
       Intercept         Value     Num DF    Den DF   F Value  Pr > F
----------------------------------------------------------------
          Wilks' lambda    0.0053 42.0000 1957.0000 8700.4324 0.0000
         Pillai's trace    0.9947 42.0000 1957.0000 8700.4324 0.0000
 Hotelling-Lawley trace 186.7236 42.0000 1957.0000 8700.4324 0.0000
    Roy's greatest root 186.7236 42.0000 1957.0000 8700.4324 0.0000
----------------------------------------------------------------

----------------------------------------------------------------
            Group           Value   Num DF    Den DF  F Value  Pr > F
----------------------------------------------------------------
            Wilks' lambda  0.0871 42.0000 1957.0000 488.2624 0.0000
           Pillai's trace  0.9129 42.0000 1957.0000 488.2624 0.0000
   Hotelling-Lawley trace 10.4788 42.0000 1957.0000 488.2624 0.0000
      Roy's greatest root 10.4788 42.0000 1957.0000 488.2624 0.0000
================================================================
```

Figure 4. *MANOVA of this dataset*

3.3. Logistic regression, Random forest classification, Decision tree

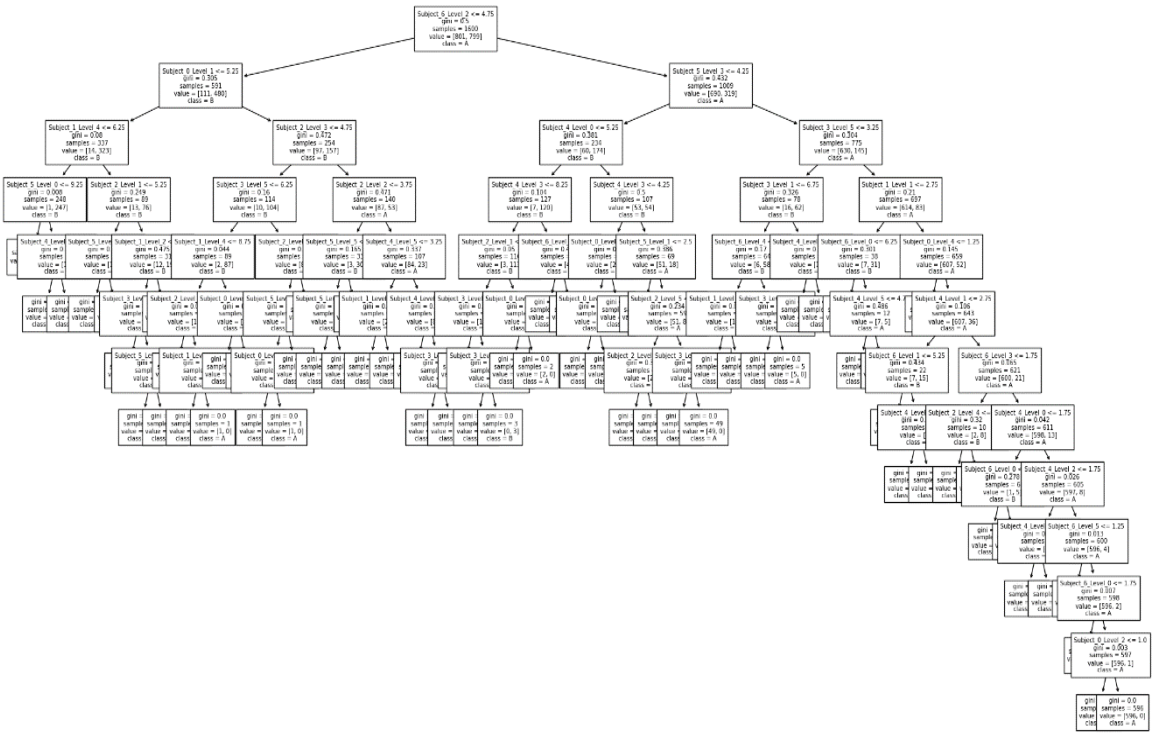

Figure 5. *Decision tree model for the dataset is ploted by using plot_tree from package sklearn.tree*

These models allow to classify the students in groups A and B basing on their scores. Hence, these classifications will tell us how distinct it is between students in two groups. They all present the superior in score of students in group A to ones in group B. The first two models show a high accuracy, precision and recall, f1-score. These models are constructed by using functions LogisticRegression from package sklearn.linear_model and RandomForestClassifier from sklearn.ensemble. Results are shown as below.

Accuracy: 1.0

Classification Report:

	precision	recall	f1-score	support
A	1.00	1.00	1.00	302
B	1.00	1.00	1.00	298
accuracy			1.00	600
macro avg	1.00	1.00	1.00	600
weighted avg	1.00	1.00	1.00	600

Decision tree model (Figure 5) is helpful in understanding how the features are used to make predictions. It indicates the importance of each feature in making predictions. Feature Subject_6_Level_2 has highest importance score of 0.27, followed by Subject_5_Level_3 of 0.14.

4. Conclusion

The study has provided insights into the effectiveness of integrative teaching strategy through the application of statistical techniques such as statistical visualization, MANOVA, logistic regression, and random forest classification, decision tree analysis. We have uncovered significant differences in the performance of students across various Bloom's Taxonomy levels and subjects of probability. Our findings indicate that integrated teaching probability with R group A demonstrates the superior performance in comparing with the traditional strategy (without a programming language integrated in teaching these subjects). This suggests that the integrated teaching strategy is more effective in enhancing students' understanding and mastery of the educational material. The integrated teaching strategy as mentioned here definitely benefits IT engineering students in ICTU to attain the objectives and Learning outcome of the probability and statistics course. They can also not only build up their knowledge and skills in data analysis, understand the tools of probability theory and statistics in studying data, but also build up their skill in programming, using cutting-edge technologies in studying data. Teaching probability and statistics with R benefits IT engineering students in maintaining student engagement and motivation. These aspects contribute to better retention of knowledge and a more enjoyable learning experience, which are important for achieving the educational goals of the subject. The source code for this study is available on GitHub at https://github.com/tiep1144/ictu.

REFERENCES

[1] M. C. Tucker, S. T. Shaw, J. Y. Son, and J. W. Stigler, "Teaching Statistics and Data Analysis with R," *Journal of Statistics and Data Science Education*, vol. 31, no. 1, pp. 18-32, 2022, doi: 10.1080/26939169.2022.2089410.

[2] Goodman, D. Noah, and A. Stuhlmüller, "The design and implementation of probabilistic programming languages," 2015. [Online]. Available: http://dippl.org. [Accessed April 01, 2024].

[3] R. Dos Santos Ferreira, V. Y. Kataoka, and M. Karrer, "Teaching probability with the support of the R statistical software," *Statistics Education Research Journal*, vol. 13, no. 2, pp. 132-147, 2014, doi: 10.52041/serj.v13i2.286.

[4] L. Pavlenko, M. Pavlenko, V. Khomenko, and V. Mezhuyev, "Application of R Programming Language in Learning Statistics," *Proceedings of the 1st Symposium on Advances in Educational Technology* (AET 2020), vol. 2, pp. 62-72, 2020, doi: 10.5220/0010928500003364.

[5] I. Ross and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299-314, 1996, doi: 10.2307/1390807.

[6] D. Tim, "Modern Data Science with R," *Journal of Statistical Software*, vol. 80, 2017, doi: 10.18637/jss.v080.b02.

[7] M. Borovcnik and R. Kapadia, "Research and developments in probability education," *International Electronic Journal of Mathematics*, vol. 4, no. 3, pp. 111-130, 2009.

[8] L. V. Pavlenko, M. Pavlenko, V. H. Khomenko, S. V. Khomenko, and M. M. Skurska, "Innovative approaches to the study of statistics by future IT professionals based on the use of the R programming language," *Physics and Mathematics* Education, vol. 1, no. 23, pp. 97-105, 2020.

[9] L. W. Anderson and D. R. Krathwohl, *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Complete Edition. New York: Longman, 2001.

[10] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *Theory into Practice*, vol. 41, no. 4, pp. 212-218, 2002, doi: 10.1207/s15430421tip4104_2.

[11] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859-877, 2017, doi: 10.1080/01621459.2017.1285773.