Optimizing Learning: Predicting Research Competency via Statistical Proficiency

Tarid Wongvorachan 1,*, Siwachoat Srisuttiyakorn 2 and Kanit Sriklaub 2

1 Measurement, Evaluation, and Data Science, Faculty of Education, University of Alberta, Edmonton, AB T6G 2G5, Canada
2 Faculty of Education, Chulalongkorn University, Bangkok 10330, Thailand; siwachoat.s@chula.ac.th (S.S.); kanit.s@chula.ac.th (K.S.)
* Correspondence: wongvora@ualberta.ca

Abstract: In higher education, the cultivation of research competency is pivotal for students’ critical thinking development and their subsequent transition into the professional workforce. While statistics plays a fundamental role in supporting the completion of a research project, it is often perceived as challenging, particularly by students in majors outside mathematics or statistics. The connection between students’ statistical proficiency and their research competency remains unexplored despite its significance. To address this gap, we utilize the supervised machine learning approach to predict students’ research competency as represented by their performance in a research methods class, with predictors of students’ proficiency in statistical topics. Predictors relating to students’ learning behavior in a statistics course such as assignment completion and academic dishonesty are also included as auxiliary variables. Results indicate that the three primary categories of statistical skills—namely, the understanding of statistical concepts, proficiency in selecting appropriate statistical methods, and statistics interpretation skills—can be used to predict students’ research competency as demonstrated by their final course scores and letter grades. This study advocates for strategic emphasis on the identified influential topics to enhance efficiency in developing students’ research competency. The findings could inform instructors in adopting a strategic approach to teaching the statistical component of research for enhanced efficiency.

Keywords: research competency; statistics; supervised machine learning

1. Background of the Study

In higher education programs, the development of research competency is a major milestone in the students’ curriculum due to its potential to instill critical thinking and transition students into the professional workforce [1,2]. A pivotal aspect of research competency development lies in statistics, given its fundamental role in analyzing data to support research findings [3]. For example, researchers need descriptive statistics to summarize characteristics of the data, while inferential statistics makes inferences of the population from the sample at hand to test hypotheses and answer research questions [3,4]. At a more advanced level, structural equation modeling (SEM), a multivariate statistical technique, is often utilized to examine relationships between variables of interest [5]. These examples show that statistics is essential in developing research skills in undergraduate students [2]. Despite its significance, statistics components are often perceived as challenging by students [6]. This difficulty is attributed to the mathematical components inherent in statistics, which might pose comprehension challenges compared to content that focuses on facts and ideas [7,8].

In fact, a considerable number of students who learn statistics belong to majors outside mathematics or statistics, such as political science or psychology [9]. The challenges faced by these students in statistics courses could induce subject-related anxiety, highlighting the
struggle in grasping statistical concepts [6,10]. It is established that students’ proficiency in statistics, as represented by their formative assessment scores, can determine their overall performance in the statistics course [10,11]. However, whether students’ statistical proficiency can predict their research competency, as reflected by their final scores in a research methods course, has yet to be examined. In response to this gap, we perform an investigation to identify specific statistical topics that can predict students’ final course grades in a research methods course. This investigation seeks to identify areas of importance, providing a foundation for a strategic approach to emphasize and refine relevant topics in the statistics course.

To fulfill the aim of our research, we employ a supervised machine learning model to identify influential predictors of students’ performance in a research method class. The overarching research question is as follows: “How is the predictability of skills in statistics to students’ research competency?” The predictor variables are students’ formative scores in each topic from a statistics course, as well as their learning behavior in the statistics course as auxiliary variables. The outcome variable is students’ learning performance in the research method class. Results from the analysis include a predictive regression model for students’ final course score in a research method course, a predictive classification model for students’ success in the research method course, and lists of important predictors to the targeted variable as well as their influence on the prediction. While predicting students’ statistical competency is more directly related to the predictors, predicting students’ research competency may allow us to extrapolate the results to examine how well students can practically apply statistical concepts to real-world research scenarios, especially in research-oriented professions [12].

Instead of relying on traditional statistical analysis for retrospective inference, we employ a machine learning approach to predict students’ research competency. This approach offers an algorithm and predictors as a guideline to inform instructors in developing their course designs on the research methods topics [13]. Machine learning technology has been extensively employed in educational settings, particularly in the field of learning analytics. In this context, data on students’ learning performance, interactions with course materials, and behaviors such as assignment submission times are leveraged to inform decision-making for both students and instructors [14]. Furthermore, this study contributes to the body of knowledge by identifying topics in statistics that are crucial in determining students’ research competency. Ideally, instructors should ensure that students understand every topic of the course material. However, it is impractical to deliver the entire course content at a detailed yet slow pace, considering the time limit of a standard program. Such a program typically allows a maximum of three teaching hours per day over a 16-week semester [15]. This research could highlight the topics in statistics that need more emphasis to increase efficiency in developing students’ research competency. By investing time and resources into enhancing the accessibility of these topics, instructors could enhance students’ background knowledge in statistics and consequently their competency in the research methods course.

2. Literature Review

2.1. Foundations of Research Competency in Higher Education

Based on the researcher skill development framework (RSD), research competency is defined as the ability to formulate or respond to a research question, employing rigorous methodologies, and effectively disseminating findings to various audiences [16]. The RSD framework comprises six aspects that contribute to research competency: purpose, data acquisition, credibility evaluation, data organization, knowledge synthesis, and findings dissemination. Although the RSD framework primarily focuses on skills, we argue that its aspects can also be construed as constituting components to research competency. This is because the six components align with the definition of competency, which is the ability to produce observable performance, possess knowledge, and adhere to standards in executing successful research [16,17]. Many higher education programs aim to develop
research competency in students to foster critical thinking, enabling students to contribute new knowledge to the academic community [1,18]. In fact, research methods courses at the university level often mandate extensive prerequisite courses for senior students, ensuring that they can translate theoretical knowledge into practical applications through research [19].

The six aspects of research competency outlined in the RSD framework align with the typical components of a research design, which involve the background of the topics under investigation, literature review, methods of data collection and analysis, and findings interpretations [3]. These elements enable researchers to identify issues within a chosen topic, formulate research questions, collect pertinent data, and draw conclusions from the analysis results [4]. Students’ mastery of these components is often evaluated through a research project, underscoring that research competency is a skill cultivated through practical experience [3,19]. Such a skill can be represented by their final course scores in a research method course. Developing this competency equips students to integrate the research process mindset into their professional practice, enhancing their effectiveness in addressing challenges in their future careers or advanced academic pursuits [1]. The transferable skills derived from research, such as critical thinking, data analysis, and information organization, can enhance students’ preparedness for a transition to the professional workforce [20].

2.2. The Role of Statistics in Developing Research Competency: Components and Challenges

Within the components of a quantitative research design, statistics serves as a critical link between the research question and the conclusions, playing an instrumental role in developing students’ research competency, particularly in data acquisition, credibility evaluation, data organization, knowledge synthesis, and findings dissemination (five out of six) [16]. In the aspect of data acquisition, statistics aids in determining an appropriate sample size through power analysis, estimating the smallest sample size necessary to achieve the required statistical power for hypothesis testing in an experiment [21]. During credibility evaluation, confirmatory factor analysis, a technique within SEM, assesses the quality of data collection instruments, ensuring alignment between the theoretical structure and empirical data gathered from the pilot study phase [5]. In the data organization aspect, descriptive statistics, such as central tendency analysis and data distribution summaries, reveal patterns that may be unavailable through mere observations [21]. In the knowledge synthesis, inferential statistics, like independent/dependent sample t-tests or analysis of variance (ANOVA), tests research hypotheses, examining statistical significance and effect size [3,21]. Lastly, in findings dissemination, data visualization is crucial for conveying results in an easily understood manner, showcasing the significance of the knowledge in statistics on key metrics that should be reported [22]. These instances underscore the pivotal role of statistics in shaping students’ research competency.

However, students in higher education, especially those outside mathematics or statistics majors, find statistics difficult to grasp; this struggle could hinder the development of their research competency, given the pivotal role of statistics in this context [6,23]. Specifically, the concepts, rules, and formulas in statistics can be both complex and counter-intuitive, discouraging students and inducing anxiety during exams [6,23]. Some students do not achieve learning at the conceptual level, as they merely memorize rules and formulas; this learning approach could make their learning less effective [23]. This problem is exacerbated by the fact that some statistics classes are large, making it impractical for instructors to engage students in hands-on projects at the process level [24]. Additionally, students with limited experience in statistics may feel nervous in their practice due to the disorganization of the real-world data that require skills to handle [23]. To address these challenges, instructors can offer opportunities for students to develop statistical thinking skills through project-based learning, supplemented by innovative teaching materials like web-based platforms for simulating statistical problems [23,24]. However, this solution is more viable in small classes where instructors can provide individualized supervision.
In larger classes, covering each student comprehensively becomes impractical. While emphasizing important topics for efficiency is a potential approach, the literature lacks clarity on which statistical topics significantly influence students’ research competency, highlighting a critical gap that necessitates investigation for a deeper understanding of students’ overall research skills development.

2.3. The Application of Machine Learning in Education

Machine learning has increasingly been utilized in educational research to predict student outcomes and enhance learning processes. Studies have demonstrated that machine learning models can effectively predict student performance by analyzing various indicators of student learning, such as time spent on assignments, formative assessment scores, and self-efficacy measures [11,25]. These models provide data-driven insights to stakeholders in education—students, teachers, parents, and principals—thereby informing decision-making processes.

For instance, Guo et al. [26] employed the neural network, a type of machine learning model, to predict the learning outcomes (i.e., outstanding, good, average, pass, and fail) of over 100,000 students based on their background and demographic data (e.g., gender, age, and health status), past academic performance (e.g., GPA from previous educational levels), school data (e.g., school type and school ranking), learning performance data (e.g., formative and summative scores), and personal data (e.g., attention and psychological background). Additionally, language-based machine learning technologies have been used to assess the quality of feedback provided by instructors, distinguishing between high- and low-quality feedback [27].

These examples illustrate the widespread implementation of machine learning in educational research. With the availability of educational data at both local levels (e.g., classroom data) and global levels (e.g., large-scale databases such as the Programme for International Student Assessment—PISA [28], and the Trends in International Mathematics and Science Study—TIMSS [29]), it is likely that machine learning technology will play a crucial role in current and future educational research.

3. Current Study

To answer the research question of “How is the predictability of skills in statistics to students’ research competency?”, this study utilized predictors comprising students’ performance in key statistical areas from a statistics course such as analysis methods selection or output interpretation, as well as their learning behavior in the course. These predictors were used to predict students’ research competency as indicated by their learning performance in a research methods course. The study undertook both regression and classification tasks: the former predicted students’ final course scores, while the latter predicted their course success, defined as achieving over 80% of the total course score—equivalent to a grade of B or higher. We also incorporated students’ background variables such as time taken to complete the assignment, cheating behavior in statistics assessments, and students’ quiz performance to examine the influence of students’ learning behavior in their statistics class in addition to their statistics skills. Both regression and classification algorithms were fine-tuned for optimal performance and evaluated using metrics appropriate to their respective tasks. The findings of this study offer insights into statistical topics that are influential to students’ research competency, which could inform the development of higher education curricula.

4. Methods
4.1. Dataset

The dataset utilized in this study encompasses undergraduate students’ profiles from both a statistics course and a research methods course at a Thai university, totaling N = 385 participants. All students enrolled in the research method course during the semester following their completion of the statistics course. The statistics course serves as a
prerequisite for enrollment in the research methods course. Both courses are instructed by various faculty members, all of whom are affiliated with the Department of Research and Psychology. Approximately half of the instructors teach both courses.

Both the statistics and research methods courses incorporate traditional lectures along with lab or tutorial sessions. Students are required to submit their assignments electronically via a learning management system. In the statistics course, assignments include both individual and group work. However, only individual assignments are considered in calculating the variable for cheating behavior. Both statistics and research methods courses are taught with both traditional lecture and lab/tutorial. Students were instructed to submit their assignments electronically through a learning management system. Data preprocessing and predictive model development primarily relied on the R programming language [30]. Data on the variables of interest were collected as part of their learning performance and behavior during the two courses. Assessments were conducted in Thai, the official language and medium of instruction at the university. Table 1 describes a list of variables utilized in this study and their code. The dataset utilized in this study is classified as secondary data due to its anonymity to the primary researchers. This ensures minimal ethical concerns, as there exists no feasible method for re-identifying the participants.

Table 1. List of utilized variables.

<table>
<thead>
<tr>
<th>Variable Code</th>
<th>Variable Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResComp</td>
<td>Students' research competency.</td>
</tr>
<tr>
<td>Interpret</td>
<td>Students' statistics interpretation skill.</td>
</tr>
<tr>
<td>Concepts</td>
<td>Students' understanding of statistical concepts.</td>
</tr>
<tr>
<td>ChooseMethod</td>
<td>Students' skills in statistical method selection.</td>
</tr>
<tr>
<td>SubmitRate</td>
<td>Students' assignment submission rate.</td>
</tr>
<tr>
<td>LearnPerform</td>
<td>Students' post-lecture quiz performance</td>
</tr>
<tr>
<td>AvgTimeSubmit</td>
<td>Students' time taken to complete the assignment.</td>
</tr>
<tr>
<td>CheatingBehavior</td>
<td>Students' cheating behavior in assignments.</td>
</tr>
</tbody>
</table>

4.2. Feature Selection

For feature selection, statistics topics were categorized into three main categories: 1. The interpretation category includes topics involved in translating and summarizing the analysis results from data (i.e., describing the data distribution, analyzing the relationship between variables using statistical measures and data visualizations, and interpreting results from hypothesis testing such as \(t\)-test or ANOVA. 2. The concept category comprises topics involving essential theories and principles of statistical methods such as sampling distributions, estimation, and hypothesis testing. This category also covers the understanding of statistical assumptions for statistical tests, the rationale behind different types of data scales (nominal, ordinal, interval, and ratio), and the conceptual framework for choosing appropriate statistical tests based on research questions and data characteristics. 3. The method selection category involves the practical application of statistical techniques and decision-making processes to choose the most suitable methods for data analysis based on the nature of data and research objective. Specifically, this category includes selecting the correct types of \(t\)-tests (one-sample, independent, or paired-sample), choosing between parametric and non-parametric tests based on data distribution and sample size, deciding on the appropriate correlation coefficients (Pearson, Spearman, or Cramer’s V) to examine the strength and direction of relationships between variables, and selecting the appropriate regression model for predicting outcomes or explaining the relationship between multiple variables. These three main variables were calculated based on students’ final exam scores as outlined in the test blueprint. The final exam weights 30% of the overall course grade.

Predictors that account for students’ background comprised three variables as follows: first, the students’ time taken to complete the assignment, represented by their average time of submission collected through a learning management system; second, the students’ submission rate (whether students submitted the assignment); and third, the students’
post-lecture quiz performance, represented by the average score of the post-class exercise in each lecture. These exercises were administered electronically via an online learning platform. They were designed to assess students’ understanding of the lecture, providing a measure of the students’ attention to the concepts taught in class. The content of the quizzes are questions that mirror the content taught earlier in the class such as statistical concepts or analysis results interpretation. Since this course is taught in the Faculty of Education, the students’ technical knowledge, such as mathematical formulas or equations, is not assessed unless necessary. The accuracy of the students’ responses does not contribute to their overall class grade. However, the completion of the quiz within the specified time frame is considered part of their class participation score, which accounts for 10% of the overall class grade. Data on these three variables were collected and extracted through a learning management system.

Finally, students’ cheating behavior is indicated by the median cosine similarity among their open-ended homework responses. A value close to 1.00 means a student’s work is very similar to others’, hinting at possible plagiarism. While assignments in the statistics course include both individual and group components, only individual work is used in computing the cheating behavior variable. These individual assignments typically comprise open-ended tasks that prompt students to express their opinions or respond to questions based on analysis results or given scenarios. Given the open-ended nature of these tasks, variations in students’ responses are expected. To detect potential cheating behavior, the researcher calculated cheating behavior scores by evaluating assignment similarity using cosine similarity between students’ responses. A high cosine similarity between responses may imply cheating behavior in their open-ended assignments. In total, seven variables served as predictors in this study. All of these variables were continuous.

All predictors were chosen for their relevance to students’ performance in statistics. The three categories of statistical skills were chosen because they represent the fundamental skills taught in the statistics course and align with the RSD framework. Additionally, predictors regarding students’ background were included to provide context for their statistical skills. These auxiliary variables reflect students’ effort in tasks, which can impact their learning of statistical skills. The outcome variable is the students’ final course score in the research method course for the regression task. For the classification task, the score was categorized into two classes, with class 1 representing students who achieve 80% and above in the final research method course grade and class 0 representing students who achieve below 80% in the final research method course grade. All predictors were examined with a correlational analysis to ensure their relationships among each other and relationships to the outcome variable.

4.3. Illustrative Tasks and Questions for Measured Variables

The task example and question example of the measured variables (i.e., Interpret, Concepts, Choose method, Learn perform, and Rescomp) are as follows.

4.3.1. Interpret

Tasks and questions under this variable focus on students’ ability to understand and explain statistical results. A task example of this component is “Interpret the output of a correlation analysis. Describe what the coefficients indicate about the relationship between two variables.” A question example would be “Data analysts have found that insufficient rest is correlated with students’ exam performance, with a correlation coefficient of $r = -0.886$. Please interpret the result of this analysis appropriately.” Finally, this component accounts for approximately 33% of the tasks and questions in the statistics course.

4.3.2. Concepts

This variable assesses students’ understanding of fundamental statistical concepts. A task example of this component is “Explain the concept of multicollinearity problems in regression analysis.” A question example can be seen in Figure 1.
Which of the following indicates that the estimated regression equation might have a multicollinearity problem?

A. The VIF (Variance Inflation Factor) values are very low, and the Tolerance values are very high.
B. The correlation between the independent variables and the dependent variable is positive, but the regression coefficients are negative.
C. The correlation between the dependent variable and the independent variables is high, but the regression coefficients are not statistically significant.

Figure 1. A question example of the ‘Concept’ component.

This component also accounts for approximately 33% of the tasks and questions in the statistics course.

4.3.3. ChooseMethod

This variable evaluates students’ ability to select appropriate statistical methods for various research scenarios. A task example of this component is “Given a research scenario where you need to compare the means of two dependent groups, which statistical test would you choose and why?” A question example can be seen in Figure 2.

Figure 2. A question example of the ‘ChooseMethod’ component.

Similarly to the other two components, this component also accounts for approximately 33% of the tasks and questions in the statistics course.

4.3.4. LearnPerform

As an auxiliary variable from the statistics course, the LearnPerform variable was assessed through post-lecture quizzes designed to evaluate students’ immediate understanding and retention of the lecture material. Each quiz consisted of a mix of multiple choice, true/false, and short answer questions. However, the students’ response correctness does not contribute to the overall course grade; only their participation in the quiz does. Therefore, students who participate in all quizzes will receive a full 10% toward their final grade in the statistics course regardless of their response correctness. Students who partially complete the quiz will receive a prorated quiz score. See Figure 3 for an example of a short answer question. See Figure 4 for an example of a true/false question. See Figure 5 for an example of a multiple choice question.

4.3.5. ResComp

Aside from the three categories of students’ statistical proficiency, the dependent variable of students’ research competency was assessed through a combination of practical assignments, mid-term exams, and the final project. The final grade for the research methods course was determined by a weighted average of these components.

- Assignments: Practical assignments required students to apply statistical methods to analyze datasets, interpret results, and write reports. An example of the assignment...
task is “Analyze the provided dataset using an appropriate statistical test to determine if there is a significant difference in test scores between two teaching methods. Submit a report detailing your analysis, results, and conclusions.”

- **Midterm Exam:** The mid-term exam consisted of theoretical questions and practical problems requiring statistical analysis and interpretation. An example of a midterm exam question is “Describe the assumptions of ANOVA and perform an ANOVA test on the given dataset. Interpret the results.”
- **Final Project:** The final project involved a comprehensive research study, where students formulated a research question, collected and analyzed data, and presented their findings. The overall direction of the final project in the research method course is “Conduct a research study on a topic of your choice, using appropriate statistical methods to analyze the data. Prepare a report that includes your research question, methodology, analysis, results, and conclusions.”

### Linear Regression

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R²</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.769</td>
<td>0.591</td>
<td>286.452</td>
<td>2</td>
<td>397</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.292</td>
<td>1.607</td>
<td>2.049</td>
<td>&lt;0.0411</td>
</tr>
<tr>
<td>Hour</td>
<td>3.873</td>
<td>0.179</td>
<td>21.639</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Class.Behav</td>
<td>4.510</td>
<td>0.414</td>
<td>10.906</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

A student who reviews their lessons for 12 hours per week and exhibits excellent study behavior is likely to score approximately how many points on the exam? (Answer to three decimal places)

**Figure 3.** A question example of a short answer question in the ‘LearnPerform’ component.

An educational researcher aims to create a data visualization with the objective of comparing the average mathematics scores of students with different learning styles (learn.style), which include auditory learning, kinesthetic learning, and reading-writing learning, within public and private schools.

The resulting data visualization developed by the researcher to address the above objective is shown below. (The numbers in the diagram represent the average mathematics scores.)

![Data Visualization](image)

Is the data visualization developed by the researcher considered an appropriate visualization?

A. TRUE
B. FALSE

**Figure 4.** A question example of a true/false question in the ‘LearnPerform’ component.
Consider the following statements:
a. The distribution of mathematics achievement scores is right-skewed.
b. No student scored more than 8 points in mathematics.
c. More than 50% of students scored less than 5 points in mathematics achievement.

Which of the following is correct?

A. a and b
B. a and c
C. b and c
D. a, b, and c

Figure 5. A question example of a multiple choice question in the ‘LearnPerform’ component.

The students’ final grade in the research method course was determined as follows: assignments, 30%; midterm exam, 30%; and final project, 40%.

4.4. Data Preprocessing

In terms of data preprocessing, one case exhibited missing values, which were addressed using the bootstrap aggregating trees imputation technique via the recipe package [31]. We conducted the train–test split procedure using the "initial split" function from the rsample package [31], with a split ratio of 80% for training data and 20% for testing data. For the classification task, there was a 60:248 discrepancy between the number of instances in class 1 and class 0, respectively, indicating a moderate class imbalance issue. To mitigate this, we employed the Synthetic Minority Oversampling Technique (SMOTE) from the themis package [32]. SMOTE synthesized additional instances of the minority class (class 1), resulting in a balanced class proportion of N = 248 for each class in the final dataset used for classification. The final dataset for the classification task comprised 419 instances for training and 77 for testing datasets. For the regression task, the final dataset consisted of 308 instances for training and 77 instances for testing datasets.

4.5. Predictive Algorithm

For predicting the outcome variable, we employed the Elastic-net regularized generalized linear model (GLM) as our predictive algorithm [33,34]. The Elastic-net GLM model is similar to the commonly used ordinary regression procedure but incorporates additional features such as a penalty term. This term effectively reduces the coefficients of unimportant predictors (i.e., predictors that contribute minimally to the prediction), thereby preventing overly complex models and mitigating potential multicollinearity issues [34].
The Lasso method can shrink predictive coefficients of influential variables to exactly zero, effectively performing variable selection by retaining only the most important predictors for model simplification [35]. The ridge method, unlike lasso, does not set any coefficients to zero but rather reduces their coefficient to minimize the impact of less important features while keeping all variables in the model [35]. Elastic-net regularization introduces two penalty parameters: $\lambda$ and $\alpha$. The $\lambda$ parameter controls the overall strength of the penalty, while the $\alpha$ parameter determines the mix between ridge ($\alpha = 0$) and lasso ($\alpha = 1$) regularization [34]. This flexibility allows Elastic-net to perform well in various situations, balancing the benefits of both types of regularization.

This choice was made after comparing various predictive algorithms for both regression and classification tasks, including random forest, k-nearest neighbor, support vector machine, and extreme gradient boosting trees. These algorithms yielded comparable results. However, the Elastic-net GLM possesses an advantage of interpretability. Given its linear nature, this model allows for clear interpretation of how predictors influence the outcome variable directionally [33]. Additionally, linear models, like the Elastic-net GLM, have been shown to perform well with small sample sizes compared to ensemble models such as random forest [36]. This advantage was further enhanced by the quality of the data, as variables were meticulously selected based on their pairwise relationships [37]. By employing the Elastic-net GLM, our aim was to leverage its interpretability and effectiveness in modeling the relationship between predictors and the outcome variable.

4.6. Hyperparameter Tuning and Evaluating Metrics

To optimize both the regression and classification algorithms, the Latin Hypercube grid search method was utilized for its efficiency, offering comparable results to other approaches but at a lower computational cost [38,39]. The tuned hyperparameters were Elastic-net penalty terms and mixing parameters, both possessing the range of 0 to 1. Both algorithms had 50 sets of random hyperparameter values, through 10-fold cross validation with the 5 repetitions method ($5 \times 10$-fold CV), totaling 2500 number of trials. Following the identification of the optimal hyperparameter combination, both the regressor and classifier models underwent further training, testing, and validation using $5 \times 10$-fold CV to ensure optimal performance.

The evaluation of the regression algorithm’s effectiveness was based on regression metrics such as root mean squared error (RMSE) and R-squared. For the classification algorithm, classification metrics such as area under curve (AUC), precision, recall, F1 score, and accuracy were consulted. The regression metrics of RMSE and R-Squared are detailed as follows: RMSE is a measure of the differences between predicted and observed values. It is calculated as the square root of the average of the squared differences between the predicted and actual values [40]. RMSE is particularly useful for understanding the magnitude of errors in the predictions, with lower values indicating better model performance. R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data [40].

The classification metrics of AUC, precision, recall, F1 score, and accuracy are detailed as follows [40]: AUC is the area under the Receiver Operating Characteristic curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity). AUC provides a single value to evaluate the performance of a classifier across all threshold values. A higher AUC indicates better model performance, with a value of 1 representing a perfect classifier, and a value of 0.5 representing a model with no discriminative ability. Precision, also known as the positive predictive value, is the ratio of true positive predictions to the total number of positive predictions (both true positives and false positives). It reflects the accuracy of the positive predictions made by the model. Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total number of actual positive instances (both true positives and false negatives). It measures the model’s ability to correctly identify positive instances (i.e., students who achieve a grade of B or above). The
F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. Finally, accuracy is the ratio of the number of correct predictions (both true positives and true negatives) to the total number of predictions. It is a straightforward measure of how often the model is correct.

5. Results

5.1. Pairwise Correlation Results

Figure 6 presents the results of pairwise correlation analysis. Most variables share a significant positive correlation among each other and to the targeted variable (i.e., \( p < 0.05 \)), with some variables (i.e., average time of submission and cheating behavior) having significant negative relationships to the outcome variable. Specifically, the outcome variable “ResComp” has the highest positive correlation with the “Interpret” variable (0.568) and the “Concepts” variable (0.528). Additionally “ResComp” also shows significant positive correlation with “ChooseMethod” (0.323), “SubmitRate” (0.307), and LearnPerform (0.219). These findings suggested that the students’ research competency is positively associated with their performance across the three categories of statistical skills, as well as their completion of assignments and post-lecture quiz performance. On the contrary, variables such as “AvgTimeSubmit” (−0.290), and “CheatingBehavior” (−0.147) show a notable negative correlation with “ResComp”. This indicates that students who take a significantly longer time to finish assignments and engage in cheating tend to score lower in research competency.

![Pairwise correlation analysis](image)

5.2. Predictive Regression Analysis

After conducting a Latin Hypercube grid search tuning, the optimal hyperparameters for the GLM regressor model were the Elastic-net penalty term = 1.147 \( \times 10^{-5} \) and Mixing...
parameter \(= 2.657 \times 10^{-3}\). The model yielded an RMSE of 6.654 and an R-squared of 55.073%, indicating satisfactory performance. The RMSE of 6.654 indicates a small margin of error in our predictions, considering the possible range of research competency scores (up to 100). Furthermore, the R-squared value of 55.073% suggests that our model explains over half of the variation in student research competency. Figure 7 illustrates the relative importance of different factors in predicting students’ final scores in the research methods course. The analysis reveals that “Interpret” is the most influential factor, followed by “Concept” and “SubmitRate”. Interestingly, the time taken to complete assignments has the least influence on predicting the final scores. Regarding directionality, “Interpret”, “Concepts”, “SubmitRate”, “ChooseMethod” and “LearnPerform” demonstrated positive relationships with the outcome variable, suggesting that higher scores in these areas correspond to greater research competency as reflected in the students’ final course scores. Conversely, “CheatBehavior”, and “AvgTimeSubmit” show a negative relationship with the outcome variable, implying that factors related to it may impede students’ score in the research method course. These results align with the results from the correlational analysis.

![Figure 7. Feature importance metrics of GLM regressor.](image)

Table 2 provides a summary of regression coefficients obtained from a GLM regressor, optimized with hyperparameter tuning. The table presents estimated coefficients, both raw and standardized, along with their standard errors and the 95% quantile intervals, all derived from 50 bootstrap samples. Upon examining the 95% quantile intervals for the regression coefficients as outlined in the table, nearly all variables are significantly related to the research competency scores. However, “AvgTimeSubmit” is the only variable that lacks a significant relationship, as its quantile interval overlaps the zero region based on its range of \(-0.160\) to 0.015; this indicates uncertainty in its impact on the research competency score.
Table 2. Bootstrap regression coefficients of GLM regressor summary.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Raw Score Coefficient</th>
<th>Standardized Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard Error</td>
</tr>
<tr>
<td>Constant</td>
<td>42.360</td>
<td>5.201</td>
</tr>
<tr>
<td>Interpret</td>
<td>0.216</td>
<td>0.028</td>
</tr>
<tr>
<td>Concepts</td>
<td>0.142</td>
<td>0.031</td>
</tr>
<tr>
<td>SubmitRate</td>
<td>0.130</td>
<td>0.027</td>
</tr>
<tr>
<td>ChooseMethod</td>
<td>0.085</td>
<td>0.029</td>
</tr>
<tr>
<td>CheatBehavior</td>
<td>-4.706</td>
<td>2.303</td>
</tr>
<tr>
<td>LearnPerform</td>
<td>0.054</td>
<td>0.034</td>
</tr>
<tr>
<td>AvgTimeSubmit</td>
<td>-0.015</td>
<td>0.019</td>
</tr>
</tbody>
</table>

Note. Bootstrap sample = 50 samples.

5.3. Classification Analysis

For the classification model, the optimal hyperparameters for the GLM classifier model are Elastic-net penalty term = 0.383 and Mixing parameter = 0.122. The model achieves performance metrics of AUC = 0.811, precision = 0.407, recall = 0.733, F1 score = 0.524, and accuracy = 0.740. These metrics indicate that the model is effective at distinguishing between the two classes of students, particularly in accurately identifying students who achieve a higher score as indicated by a higher recall value. In Figure 8, feature importance metrics illustrate the predictors’ relative influence in predicting students’ likelihood of success in the research methods course, defined as achieving a final grade of B or above. Among these predictors, “Interpret” emerged as the most influential, followed by “Concept” and “SubmitRate”, respectively, while “AvgTimeSubmit” exhibited the least influence. These findings parallel those of the regression task.

Figure 8. Feature importance metrics of predictors in the GLM classifier.

Similar to the approach taken with the GLM regressor, Table 3 offers a summary of regression coefficients from a tuned GLM classifier. This table shows the estimated coefficients, encompassing both raw score coefficients and standardized coefficients expressed in terms of odds ratios, together with their respective standard errors. The table also includes
Trends High. Educ. 2024, 3

the 95% quantile intervals for each coefficient, all of which are derived from an analysis of 50 bootstrap samples. These results describe the predictive power of the predictors. Notably, “Interpret” and “Concepts” stand out as substantial positive predictors, each demonstrating an over 30% increase in the likelihood of the predicted outcome for every unit increase in their scores. This highlights the significant influence these variables have on research competency. Other variables such as “SubmitRate”, “LearnPerform”, and “ChooseMethod” also show positive associations, albeit more modest, indicating that enhancements in these areas can slightly improve research competency. In contrast, “CheatBehavior” and “Avg-TimeSubmit” exhibit minimal negative impacts, with the former showing an odds ratio close to 1.00, suggesting its negligible effect, and the latter indicating only a slight decrease in the likelihood of the predicted outcome with longer submission times. This analysis underscores the importance of students’ interpretative and conceptual skills in statistics in contributing to their research competency, while also acknowledging the lesser roles of other factors.

Table 3. Bootstrap regression coefficients of GLM classifier summary.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Raw Score Coefficient</th>
<th>Standardized Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>3.214</td>
<td>0.432</td>
</tr>
<tr>
<td>Interpret</td>
<td>0.016</td>
<td>0.003</td>
</tr>
<tr>
<td>Concepts</td>
<td>0.015</td>
<td>0.003</td>
</tr>
<tr>
<td>SubmitRate</td>
<td>0.099</td>
<td>0.003</td>
</tr>
<tr>
<td>LearnPerform</td>
<td>0.099</td>
<td>0.003</td>
</tr>
<tr>
<td>ChooseMethod</td>
<td>0.007</td>
<td>0.003</td>
</tr>
<tr>
<td>CheatBehavior</td>
<td>-0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>AvgTimeSubmit</td>
<td>-0.002</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Note. Bootstrap sample = 50 samples.

6. Discussion

This study aims to identify key predictors among undergraduate students’ statistical skills and learning behavior within a statistics course, with the goal of predicting their research competency as represented by their performance in a research methods course. The findings of this study align with various learning theories. Specifically, the statistics interpretation skills are positioned on the evaluating and analyzing levels on Bloom’s revised taxonomy because they involve the critical process of making sense of statistical outputs by analyzing results and assessing their quality [41]. These two levels on Bloom’s taxonomy require higher comprehension in the subject matter, and therefore it could be inferred that students who have mastered the statistics interpretation skills can apply their statistical knowledge more effectively in developing their research competency.

Conversely, students’ grasp of statistical concepts resides at the understanding level within Bloom’s framework, which involves the ability to describe the relationship between principles of statistical methods and their underlying assumptions [21]. Similarly, proficiency in selecting appropriate statistical methods operates primarily at the understanding and remembering levels, as students must match suitable data analysis techniques with the characteristics of their data. This skill may entail a lower level of comprehension compared to the understanding of statistical concepts, given its focus on the practical matching of data and methods rather than abstract conceptualization [21,41]. As a result, it is reasonable to infer that these latter two skills may exhibit comparatively less predictive power regarding students’ research competency, as they require a lower level of comprehension in the subject matter compared to statistics interpretation skills.

From the methodological perspective, our study’s use of a supervised machine learning model to predict research competency from statistical proficiency is consistent with existing research that highlights the potential of machine learning in educational contexts. Prior studies have demonstrated that machine learning can provide actionable insights...
into student learning behaviors and outcomes, which can help educators tailor their instructional strategies [11,42]. Our findings, which identify key statistical skills that predict research competency, add to this growing field by highlighting the importance of targeted instruction in statistics to enhance research skills through the application of machine learning technique.

From the theoretical perspective, the findings of this study also align with the framework of feedback levels proposed by Hattie and Timperley [43], which distinguishes between task-level feedback (i.e., how tasks are performed) and process-level feedback (i.e., the cognitive processes necessary to execute tasks effectively). When instructing students on the selection of statistical analysis methods, the majority of the feedback may concentrate on the task level, emphasizing correct and incorrect answers based on factual knowledge [43]. For instance, instructors might guide students to choose ANOVA for comparing continuous variables across multiple categorical groups, citing its formula and applicability [21]. This task-oriented instruction pertains to concrete and surface-level knowledge, demanding primarily task-level feedback.

In contrast, teaching students about statistical principles and the interpretation of statistical results involves a more analytical approach. Here, students must connect underlying statistical principles with the context of their study to derive meaningful interpretations. For example, understanding the nature of an intervention is crucial for interpreting statistical significance between pretreatment and post-treatment data [21]. Such tasks necessitate process-level instruction and feedback due to the abstract nature of statistical principles and contextual variables involved. Consequently, skills related to understanding statistical concepts and interpreting statistical results may wield greater influence on students’ research competency, as they engage learners in deeper levels of understanding and cognitive processing [43].

In a broader context, the findings of this study, which indicate that statistical proficiency can predict students’ research competency, are consistent with similar studies conducted in different educational contexts. For instance, Marsan et al. [44] found that fostering students’ understanding of basic statistical concepts, such as hypothesis testing, experimental design, and interpretation of research findings, led to an increased confidence in their grasp of these concepts. Additionally, students began to appreciate the benefits of learning statistics by reporting that learning statistics is helpful. Similarly, Lateh [45] identified statistics as a crucial component in developing students’ research skills, which can subsequently be transferred into essential 21st-century skills such as creativity, innovation, critical thinking, and problem-solving. These skills are vital for enabling students to become competent professionals after graduation. Furthermore, Pudjiastuti [46] demonstrated that students’ statistical literacy enhances their critical thinking skills, which has broader applications in research. For example, students who engaged extensively with statistical coursework were better equipped to critically analyze research methodologies and data interpretation in various academic and professional settings. This ability to critically assess and apply statistical knowledge highlights the importance of statistical education in fostering overall research competency. This cross-study comparison highlights the significant role of statistical education in preparing students for the demands of the modern workforce and underscores its relevance in academic curricula.

The inclusion of behavioral aspects such as submission rates and cheating behavior as predictors of students’ research competency, even though auxiliary within the scope of this study, offers valuable insights into students’ learning behaviors and motivations as learning process data. Process data, which represent students’ problem-solving processes such as the time taken to complete assignments, not only predict the final learning outcomes but also indicate actionable steps students can take to improve their performance [47]. For example, in an online learning platform, detailed process data—including time spent on assignments, submission rates, clickstream data, resource access frequency, and revision patterns—can be analyzed to identify key predictors of student success. Educators might find that timely submissions and frequent engagement with supplementary resources correlate with better
research competency. This information can guide targeted interventions, such as time management workshops, submission reminders, and resource utilization tips, thereby supporting each student’s journey towards achieving competency.

Findings regarding these learning process data align with previous literature in a sense that formative learning activities can be used to predict students’ learning performance [11]. In fact, the negative relationship of cheating behavior and time taken to complete assignments to students’ research competency can be attributed to the concept of self-efficacy, which plays a crucial role in shaping students’ academic outcomes [48,49]. Individuals with low self-efficacy may exhibit reduced effort in their learning endeavors due to diminished motivation and a sense of lack of control over their academic success [48,50]. Consequently, they may perceive themselves as incapable of achieving high scores, leading to behaviors such as procrastination or resorting to academic dishonesty. Conversely, the positive association between students’ submission rates and post-lecture quiz performance reflects their intrinsic motivation and attention to learning [50]. Students with high self-efficacy levels are more likely to be driven by internal motivations to excel academically, resulting in greater engagement and ultimately enhanced proficiency in statistics that contribute to their research competency [50].

7. Conclusions

This study advances our understanding of the crucial role that statistical proficiency plays in the development of research competency among higher education students. Employing supervised machine learning techniques, the research performed both regression and classification tasks to predict students’ final course scores and likelihood of achieving a letter grade of B or higher, respectively. The guiding research question is, “How is the predictability of skills in statistics to students’ research competency?” To answer the research question, our findings indicate that three primary categories of statistical skills—namely, understanding of statistical concepts, proficiency in selecting appropriate statistical methods, and statistics interpretation skills—can be used to predict students’ research competency as demonstrated by their final course scores and letter grades. Additionally, factors related to students’ learning behavior, such as assignment submission rates, post-lecture quiz performance, and academic dishonesty, serve as supplementary predictors. Our analyses reveal that statistics interpretation skills emerge as the most influential predictor, followed by the understanding of statistical concepts and method selection proficiency, respectively. The results are consistent with the existing literature, as statistical interpretation is a key aspect that contributes to students’ statistics anxiety as measured by the Statistical Anxiety Rating Scale [51,52]. These insights hold implications for instructors seeking to enhance the design of research methods courses within higher education contexts.

The implication of this study highlights the importance for instructors of statistics courses to prioritize lessons and tasks aimed at cultivating students’ foundational understanding of statistical principles and their skills in interpreting statistical results. This implication aligns with Zaffar et al.’s [53] finding that highlights the role of machine learning in identifying influential predictors among a large number of variables to inform decisions made in the educational context. In the context of research methods courses, instructors could incorporate review lectures focusing on these areas to reinforce students’ proficiency and readiness for applying statistical concepts to research formulation. This approach has the potential to bolster students’ research competency by equipping them with a robust statistical foundation. In fact, recent studies indicate that employing a strategic teaching approach, which involves demonstrating the practical application of statistical theories to undergraduate students, effectively reduces anxiety levels in statistics [2,54]. This approach can enhance students’ performance in statistics courses and their subsequent research methodology skills.

To enhance the practical implications of statistical education, instructors may allocate a significant portion of class time to reinforce students’ understanding of statistical concepts through real-world examples. These examples could include applying sampling
distribution in market research surveys, utilizing hypothesis testing for quality control purposes, or exploring scales of measurement in psychological constructs. To effectively implement this strategy, guest lecturers or workshops could be invited to illustrate the practical applications of these concepts, thus motivating students to engage with foundational statistical principles. Moreover, instructors can foster students’ ability to interpret statistical data by integrating data visualization tools such as Tableau Public [55], RAWgraphs [56], or OpenRefine [57] into the curriculum. By incorporating these tools, students can gain hands-on experience in interpreting patterns and relationships within datasets. For instance, they could analyze relationships between variables using scatter plots or compare distributions using box plots. Additionally, the incorporation of data visualization tools could be used to facilitate case study discussions to provide students with opportunities to apply their statistical interpretation skills in real-world scenarios, further solidifying their understanding of these concepts.

Moreover, the implications of this study can be viewed through the lens of learning analytics, as it leverages the capability of machine learning alongside students’ learning data encompassing both performance metrics and learning activities [42]. Researchers and instructors can leverage these findings to develop predictive systems that inform teaching and feedback strategies. For instance, instructors could utilize such systems to monitor the students’ progress in statistical skills across the three categories of statistical skills and intervene proactively when students show signs of falling behind, thereby ensuring that students maintain a solid grasp of statistics essential for effective learning in research methods courses. Researchers could additionally create a live platform capable of predicting students’ final scores based on their current scores in three key categories. This platform would empower students to visualize their potential future performance, thereby fostering self-regulated learning behaviors and ultimately improving their academic outcomes through informed decision-making [58]. By integrating the mentioned strategies, instructors can foster a learning environment conducive to enhancing students’ research competency and overall academic success.

This study has limitations to be aware of. Firstly, the small sample size, while common in undergraduate-level courses like statistics and research methods due to the nature of supervision-based learning, may limit the generalizability of the findings, particularly in the context of machine learning studies. To address this limitation, future research should consider incorporating longitudinal data to track students’ progress over time. This approach would allow for a larger and more diverse sample size, thereby enhancing the robustness and generalizability of the predictive algorithms used in this study. Longitudinal data would provide long-term insights into how statistical skills develop and influence research abilities throughout students’ academic careers and beyond. Secondly, the constrained sample size also restricts the selection of predictive algorithms, precluding the use of more complex models such as neural networks or random forest ensembles in their most effective form. With a larger dataset, researchers could explore the application of these advanced algorithms, potentially yielding more reliable prediction outcomes suitable for developing predictive systems in educational settings. Lastly, future investigations could expand upon the variables considered, including factors like the implementation of problem-based learning approach. Such an approach could promote knowledge retention and the practical application of skills acquired in higher education, such as statistics and research, within real-world scenarios [59]. By incorporating these additional variables, future studies can provide a more thorough understanding of the factors influencing students’ research competency to inform the design of more effective educational interventions.

**Author Contributions:** Conceptualization, T.W. and S.S.; methodology, S.S.; validation, S.S.; formal analysis, S.S.; project administration, T.W.; resources, K.S.; data curation, S.S. and K.S.; writing—original draft preparation, T.W.; writing—review and editing, T.W.; visualization, S.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.
**Institutional Review Board Statement:** Ethical review and approval were waived for this study due to the absence of direct involvement with human subjects, thereby mitigating potential ethical concerns such as invasion of privacy, coercion, or harm to participants. The data utilized in this study are anonymous to the primary researchers, ensuring the protection of individual privacy with no means of re-identification.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset utilized in this study belongs to the university of one of the authors and therefore is unavailable to the public.

**Acknowledgments:** The authors acknowledge and appreciate the assistance of the Faculty of education, Chulalongkorn university for the provision of the dataset.

**Conflicts of Interest:** The authors have no conflicts of interest to declare that are relevant to the content of this article.

**Abbreviations**

The following abbreviations are used in this manuscript:

- SEM: Structural Equation Modeling
- RSD: Research Skill Development
- ANOVA: Analysis of Variance
- GLM: Generalized Linear Model
- RMSE: Root Mean Squared Error
- AUC: Area Under Curve

**References**

1. Bandaranaike, S. From research skill development to work skill development. *J. Univ. Teach. Learn. Pract.* 2018, 15, 7. [CrossRef]


25. Wongvorachan, T.; Bulut, O.; Liu, J.X.; Mazzullo, E. A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning. Information 2024, 15, 326. [CrossRef]


35. Ghatak, A. Machine Learning with R; Springer: Singapore, 2017. [CrossRef]


37. Xu, P.; Ji, X.; Li, M.; Lu, W. Small data machine learning in materials science. NPJ Comput. Mater. 2023, 9, 42. [CrossRef]


45. Lateh, A. Using research based learning in statistics course to develop the students’ research skills and 21st century skills. Int. J. Learn. 2017, 3, 23–28. [CrossRef]


48. Bandura, A. Self-efficacy mechanism in human agency. Am. Psychol. 1982, 37, 122–147. [CrossRef]


55. Ryan, L. *Visual Data Storytelling with Tableau*; Addison-Wesley Data and Analytics Series; Addison-Wesley: Boston, MA, USA, 2018.


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.