applied sciences

MDPI

Article

Visualization and Data Analysis of Multi-Factors for the Scientific Research Training of Graduate Students

Yanan Liu [1], Guojun Li [1], Yulong Yin [1] and Leibao Zhang [2,*]

[1] School of Information Management & Artificial Intelligence, Zhejiang University of Finance & Economics, Hangzhou 310018, China
[2] School of Business, Zhejiang University City College, Hangzhou 310015, China
* Correspondence: zhanglb@zucc.edu.cn

Abstract: With the change of graduate education from quantity expansion to quality promotion, how to improve the quality of graduate cultivation has aroused wide concern. However, existing scientific quantitative methods tend to investigate the results of graduate training, with a lack of attention to the multidimensional data during the training process. Thus, exploratory analysis of multidimensional data in the graduate training process and accurate grasp of the key process factors affecting graduate academic competence is an indispensable task for achieving the stated goals of graduate education. In this paper, a visual analytic system of graduate training data is proposed to help users implement in-depth analysis based on the graduate training process. First, a questionnaire is designed about the training process to identify multidimensional data timely and accurately. Then, a series of data mining methods are utilized to further detect key factors in the training process, which will be used to make academic predictions for first-year graduates. Meanwhile, an interactive visual analytic system has been developed to help users understand and analyze the key factors affecting the graduate training process. Based on the results of the visual analysis, effective suggestions will be provided for graduate students, supervisors, and university administrators to improve the quality of graduate education.

Keywords: graduate education; data mining; visual analytics; key factors; improvement suggestions

Citation: Liu, Y.; Li, G.; Yin, Y.; Zhang, L. Visualization and Data Analysis of Multi-Factors for the Scientific Research Training of Graduate Students. *Appl. Sci.* **2022**, *12*, 12845. https://doi.org/10.3390/app122412845

Academic Editor: Giacomo Fiumara

Received: 27 October 2022
Accepted: 12 December 2022
Published: 14 December 2022

1. Introduction

With the increasing demand for talents, especially high-level academic talents, graduate education has become a national top-level form of education. The universities have become centers for the training of high-level talents and research bases for scientific innovation. As the transformation from scale expansion to quality improvement of China's graduate education, the contradiction between the quantity and quality has become increasingly prominent, which makes it an important task to improve the training quality of postgraduates. One of the core indicators of the quality of graduate training is the ability to conduct innovative academic research. However, due to the lack of appropriate quantitative management measures for graduate participation in academic research, it is difficult for administrators to effectively carry out real-time and effective planning and management for the research training process of graduate students. Thus, the exploration and analysis of the process of graduate education and training will help administrators understand the key factors that influence the level of academic competence of graduates, and then optimize and guide graduate supervisors and graduate students to achieve efficient graduate training and the established talent training goals.

The application of questionnaires and data mining techniques brings a new perspective to quantitative management of graduate education. Questionnaire is an effective data collection tool and can provide a multi-perspective view of graduate academic engagement and training process, which is important for further factor exploration. Data mining methods are used to identify and extract valid information from graduate survey data, in order

to accurately analyze and predict the factors affecting graduate research training. However, quantifying graduate research management and improving graduate quality still face the following challenges: **CH1.** The application of existing data mining techniques usually focuses on the results of graduate training, such as the scientific research achievements, graduation planning and employment, etc, but lacks the tracking, acquisition and analysis of multidimensional data in the process of graduate training. **CH2.** It is also difficult to explore the key factors closely related to the quality of graduate training through data mining and machine learning algorithms from the multidimensional data. **CH3.** Another challenge is how to provide a systematic tool to accomplish visual analysis tasks based on multidimensional data and training effects of graduate students, and help administrators to quickly identify and analyze the influencing factors of graduate training results.

In order to address the above challenges, we first design a questionnaire based on multi-dimensional data of the graduate training process. By distributing and collecting the questionnaires, it can timely and accurately identify the real-time situation of the current graduate training process in universities (**CH1**). In order to further explore the key factors in graduate training, we utilize a set of data mining methods based on the questionnaires. The model is derived from the training process of senior graduate students and then used to comprehensively predict the training results of junior graduates (**CH2**). Finally, we design a series of visual interfaces based on the model results to further help administrators understand and analyze the influencing factors affecting the graduate training process through interactive visual analysis (**CH3**). The major contributions of our work are summarized as follows:

- A questionnaire is designed based on multi-dimensional data of the graduate training process to achieve refined tracking of graduate training data.
- Data mining methods are combined with multi-dimensional questionnaire data to identify the key factors in the graduate training process.
- A set of interactive visual analytics tools integrating visualization methods and human–computer interactions are provided to assist administrators in further understanding and exploring the key factors affecting the graduate training process.

The rest of the paper is organized as follows: Section 2 introduces the related work. Section 3 describes data information and summarizes the requirement tasks and the system overview. In Section 4 the visual designs and interactions are introduced, and the algorithms for feature selection and training classifiers used in this paper are then described. The results and suggestions are discussed in Section 5. Lastly, we draw conclusions from this work and summarize the future work in Section 6.

2. Related Work

2.1. Higher Education

In recent years, with the expansion of domestic graduate education enrollment, the sudden increase in the number of students has led to an increasing number of problems in the quality of graduate training management. A great deal of current work has been applied to analyze the factors influencing student development. Excellent campus academic atmosphere, a good infrastructure, and a front-end academic communication platform have a very important impact on students' academic research. Calma et al. [1] conducted a practical study of 22 universities in the Philippines and found that the low quality of students' scientific research results is related to several factors, among which the lack of adequate infrastructure resources for learning is the most important factor. Komarraju et al. [2] found that actively promoting close contact between students and teachers was crucial to developing students' academic self-concept and improving their academic achievement. In addition to an excellent campus environment, the experience, direction, and advice of supervisors also play a significant role in the academic outcomes of students. Lechuga et al. [3] described in detail the role of supervisors for students through the three roles of Allies, Ambassadors and Master–Teachers, emphasizing that the role of supervisors was crucial in guiding students in their research studies. Wheeler et al. [4] found that frequent seminars

held by teaching assistants led to significant improvements in students' content knowledge throughout the semester, with quantitatively examining the content of their seminars and their students' learning. Acker et al. [5] obtained through semi-structured data collection that the student–supervisor relationship is a key factor in student progress, i.e., students need constructive guidance from their teachers, and a good instruction also makes students feel confident. The role of personal subjective factors in academic results is also indispensable, and there are many studies that have been conducted in different directions to analyze. Amida [6] used structural equation modeling in R language to analyze the fact that students' active factors include many components, such as time management, motivation, and future ambitions, among which self time management is a very important factor, and the lack of self time management can lead to academic delays. Barattucci [7] used Biggs' 3P learning model and a correlational design to derive a direct correlation between personal factors (motivation and self-efficacy) and academic performance. In this context, this paper conducts an in-depth analysis of the research outcomes of graduates and their influencing factors from both supervisors and students, to facilitate the follow-up of targeted improvements for the students themselves, their supervisors and relevant departments and institutions.

2.2. Education Data Mining

With the rapid development of information technology, data mining is increasingly linked to the field of education, and a lot of research work has been conducted to explore such issues. Data mining explores the information generated by online courses and digs deeper into the relevance of the information to obtain various types of demand forecasts and effective recommendations. Kardan et al. [8] used a neural network approach in data mining techniques to analyze the factors that influence students' online course selection and predict that their course selection needs to help universities make optimal course arrangements. Wong et al. [9] analyzed student forums by extracting forum threads and discussion posts through data mining to understand how students interact with different discussion groups. Nilashi [10] used data mining, machine learning, and statistical data to analyze the information generated from the portal in order to explore the level of learner satisfaction with the MOOC and investigate the factors to improve student satisfaction. Data mining can be applied not only to predict outcomes, but also to identify the factors that influence each other in graduate groups and individuals to make targeted learning recommendations. Onwuegbuzie et al. [11] predicted the achievement of 26 graduate students in collaborative groups and found that initiating different ideas from individual students in collaborative learning groups could promote graduates' motivation in research. Chen et al. [12] combined fuzzy set theory with data mining techniques to identify important factors affecting student learning outcomes from infrequent data. In addition to forecasting and advising, it also has a wide range of applications in school education system management. Pardos et al. [13] extracted the structural information of the course through data mining and helped users understand the chapters in the course knowledge by building a course tree. Yin [14] proposed to apply data mining methods to student information systems to extract useful student information through data mining to optimize the management of school. Gu [15] introduced a data mining method which can mine the effectiveness index data and complete the construction of the effectiveness model of vocational education model reform. However, most of the current research work on data mining mainly focuses on undergraduate studies and employment, but lacks attention to the management of graduate training. Therefore, against this background, this paper focuses on the use of data mining techniques to explore the attributes of graduates' academic research capabilities throughly.

2.3. Education Data Visualization

Visualization is a theory, method, and technology that uses computer graphics and image processing technology to convert data into graphics or images displayed on the

screen and to process them interactively, which has been used extensively in several fields in recent years such as machine learning [16,17], deep learning [18], radio signals analysis [19], anomaly detection [20], graphical perception [21] and sampling [22], and the contrastive dimensionality reduction [23]. In addition, the applications in the field of higher education are manifested in three main areas, namely the impact of visualization on students, higher education institutions, and university faculties.

The impact of visualization on students involves the visual presentation of content and the communication of useful teaching information. Kumar et al. [24] integrated two techniques, narrative and text visualization, to convert text into a narrative format in order to help students understand the course in an easy way. Fahd et al. [25] implemented semantic visualization with domain-specific knowledge through knowledge maps, next-generation graph data storage, and modeling of unstructured data, which has facilitated students' deep conceptual understanding of the literature and the discovery of hidden knowledge from large digital repositories and their associations. Ida et al. [26] proposed a visualization method for textual information of higher education courses to analyze the structure of higher education and provide useful educational information for teachers. Zhao et al. [21] formulated three hypotheses about the role of visual focal areas, graph structure recognition, and mental model formation on graph perception, and used real-world graphs with background stories to assess these hypotheses, allowing students to improve long-term memory and classroom engagement with graphs through familiarity with background stories. In terms of impact on higher education institutions, visualization analysis can help university management make sound decisions. Chong [27] and others used data metrics visualization techniques and learning behavior analysis methods to provide targeted support and interventions for adult learners in higher education so that schools can understand adult learners and anticipate their needs. Vílchez [28] combined bibliometrics and information visualization to identify clusters of academic networks through co-citation analysis which could help managers of higher education institutions to make decisions that lead to useful academic outputs. Ngo et al. [29] proposed a unified data framework that allows for aggregating high-demand data sources into a single research resource relevant to higher education research, while developing a set of analytical tools based on the data framework for helping educational researchers with data mining and synthesizing visual analysis of complex data sources. Choo et al. [30] built a web-based deep learning library, ConvNetJS with Deepvis toolbox, to provide effective interactive visualization for interpreting deep learning models and help educators achieve their teaching tasks through deep learning techniques with visualization modules. Schwab et al. [31] adapted a real-world course into a wed-based educational system, boc.io, and aided students' understanding of course knowledge by linearly or nonlinearly adding educational concepts and materials such as lecture slides, book chapters, videos, and LTI. With respect to impact on teachers, Wei et al. [32] promoted personalized education by predicting students' interactive performance in an online problem pool and recommending different learning resources to students with different needs. Sundgren [33] provided quick and accurate estimates of students' online writing collaboration strategies by exploring the visualization of Google Doc revision history for online collaborative writing documents. However, most of the visual analytic work has been used currently in teaching and learning, mainly in the undergraduate field, to help students understand the course more intuitively and tap into the hidden knowledge in the course through knowledge visualization. Few studies have explored the training and academic aspects of graduates through visualization. This paper takes advantage of visualization technology to explore the training of graduates and future academic research projections, to help students, instructors, and universities understand the cultivation of graduates better and improve the quality of graduate teaching.

3. Requirement Analysis and System Overview

3.1. Data Description

The questionnaire was designed by professionals from the graduate school of a university in Zhejiang Province, China, and its design was reasonable. It was distributed to all grades of graduate students (majors not counted) enrolled in the university at that time, and 677 questionnaires were collected in a timely and effective manner.

In order to better discover the value of the data and facilitate the processing of the model, it is necessary to pre-process the data. Firstly, since some attributes cannot describe the intrinsic distribution pattern of the sample, such as the way and time of filling in the questionnaire, the independent variables were deleted to remove factors that had no effect on the results. Meanwhile, manual answer standardization was carried out for irregularities and implied information to turn them into data that can be processed normally. Secondly, for outliers and missing values in the questionnaire, we used the mean value of the corresponding attributes to fill in, while duplicate values were deleted or replaced with the mean value as appropriate. Finally, in order to better implement the data mining algorithm based on the questionnaire data, we classified the attributes into five modules, including basic information of student, academic participation of students, basic information of supervisors, academic guidance of supervisors, and academic achievements (evaluation indicator). Each module contains the corresponding attributes. By vectorizing the options of each attribute with the numerical replacement of A-1, B-2, C-3, D-4, and E-5, the questionnaire data were converted into feature vectors required by the data mining algorithms. More detailed data will be shown in Tables 1 and 2.

Table 1. The descriptive results of questionnaires.

Module	Factors	Value	All		Third Year		Second Year		First Year	
			No.	PCT	No.	PCT	No.	PCT	No.	PCT
Students' basic information	Gender (Sbasic3)	Male	225	33.23%	45	29.41%	89	33.09%	91	35.69%
		Female	452	66.77%	108	70.59%	180	66.91%	164	64.31%
	Recent graduates (Sbasic4)	Yes	495	73.12%	123	80.39%	191	71.00%	181	70.98%
		No	182	26.88%	30	19.61%	78	29.00%	74	29.02%
	Bachelor's degree from our university (Sbasic5)	Yes	155	22.90%	27	17.65%	62	23.05%	66	25.88%
		NO	522	77.10%	126	82.35%	207	76.95%	189	74.12%
	Plan after graduation (Sbasic6)	PH D	78	11.52%	8	5.23%	28	10.41%	42	16.47%
		Civil Servant	192	28.36%	43	28.10%	80	29.74%	69	27.06%
		Staff	366	54.06%	96	62.75%	143	53.16%	127	49.80%
		Others	41	6.06%	6	3.92%	18	6.69%	17	6.67%
	Participation of academic competition (Sbasic7)	None	409	53.59%	82	53.90%	145	71.37%	182	60.41%
		School	185	29.41%	45	31.23%	84	21.96%	56	27.33%
		Province	42	13.07%	20	4.83%	13	3.53%	9	6.20%
		National	41	3.92%	6	10.04%	27	3.14%	8	6.06%

Table 1. *Cont.*

Module	Factors	Value	All		Third Year		Second Year		First Year	
			No.	PCT	No.	PCT	No.	PCT	No.	PCT
Students' academic information	Frequency of academic lectures organized by your college (Sacademic1)	Uncertain	31	4.58%	8	5.23%	11	4.09%	12	4.71%
		1-2/semester	37	5.47%	5	3.27%	23	8.55%	9	3.53%
		1-2/month	296	43.72%	77	50.33%	112	41.64%	107	41.96%
		Per week	313	46.23%	63	41.18%	123	45.72%	127	49.80%
	Frequency of participating in academic lectures (Sacademic2)	Hardly	13	1.92%	1	0.65%	8	2.97%	4	1.57%
		Once/semester	133	19.65%	33	21.57%	66	24.54%	34	13.33%
		Once/week	111	16.40%	18	11.76%	25	9.29%	68	26.67%
		Twice/week	420	62.04%	101	66.01%	170	63.20%	149	58.43%
	Frequency of participating in academic training (Sacademic3)	None	80	11.82%	11	7.19%	33	12.27%	36	14.12%
		1–2 times	300	44.31%	68	44.44%	106	39.41%	126	49.41%
		More than 3	297	43.87%	74	48.37%	130	48.33%	93	36.47%
	Frequency of reading papers (Sacademic4)	1–5/semester	35	5.17%	9	5.88%	11	4.09%	15	5.88%
		1–5/month	255	37.67%	76	49.67%	95	35.32%	84	32.94%
		1–5/week	327	48.30%	59	38.56%	135	50.19%	133	52.16%
		1/day	60	8.86%	9	5.88%	28	10.41%	23	9.02%
Supervisor's basic information	Supervisor's guiding way (Tbasic1)	Single tutor	623	92.02%	146	95.42%	248	92.19%	229	89.80%
		Tutor group	54	7.98%	7	4.58%	21	7.81%	26	10.20%
	Supervisor's title (Tbasic2)	Lecturer	9	1.33%	0	0.00%	5	1.86%	4	1.57%
		Associate professor	207	30.58%	36	23.53%	81	30.11%	90	35.29%
		Professor	461	68.09%	117	76.47%	183	68.03%	161	63.14%
	Number of students for each Supervisor (Tbasic3)	No more than 3	545	80.50%	98	64.05%	208	77.32%	239	93.73%
		No more than6	130	19.20%	55	35.95%	60	22.30%	15	5.88%
		More than 7	2	0.30%	0	0.00%	1	0.37%	1	0.39%
Supervisor's information about guidance	Communication frequency (Tguidance1)	None	10	1.48%	5	1.96%	2	0.74%	5	1.96%
		1-2/semester	60	8.86%	25	9.15%	21	7.81%	25	9.80%
		1-2/month	320	47.27%	96	53.59%	142	52.79%	96	37.65%
		1-2/week	287	42.39%	129	35.29%	104	38.66%	129	50.59%
	Communication way (Tguidance2)	Face-to-face	541	79.91%	211	77.78%	211	78.44%	119	82.75%
		Telephone	15	2.22%	4	2.61%	7	2.60%	4	1.57%
		E-mail	14	2.07%	3	2.61%	7	2.60%	4	1.18%
		Message	107	15.81%	37	16.99%	44	16.36%	26	14.51%
	Participation of supervisor's projects (Tguidance3)	None	339	50.07%	73	54.90%	126	46.84%	140	47.71%
		One	231	34.12%	55	32.55%	93	34.57%	83	35.95%
		More than two	107	15.81%	25	12.55%	50	18.59%	32	16.34%

Table 1. *Cont.*

Module	Factors	Value	All		Third Year		Second Year		First Year	
			No.	PCT	No.	PCT	No.	PCT	No.	PCT
	Supervisor's requirements for paper publication (Tguidance4)	None	180	26.59%	31	20.26%	82	30.48%	67	26.27%
		Request	46	6.79%	11	7.19%	14	5.20%	21	8.24%
		Guidance	451	66.62%	111	72.55%	173	64.31%	167	65.49%
	Frequency of academic discussion (Tguidance5)	None	105	15.51%	16	10.46%	38	14.13%	51	20.00%
		Once/semester	108	15.95%	32	20.92%	44	16.36%	32	12.55%
		Once/month	231	34.12%	61	39.87%	97	36.06%	73	28.63%
		Once/week	233	34.42%	44	28.76%	90	33.46%	99	38.82%
	Supervisor's guidance degree (Tguidance6)	None	28	4.14%	5	3.27%	11	4.09%	12	4.71%
		Little	58	8.57%	13	8.50%	21	7.81%	24	9.41%
		Much	591	87.30%	135	88.24%	237	88.10%	219	85.88%
	Supervisor's disadvantages (Tguidance7)	Too many students	77	11.37%	30	19.61%	28	10.41%	19	7.45%
		Too busy	141	20.83%	36	23.53%	53	19.70%	52	20.39%
		High demand	179	26.44%	35	22.88%	78	29.00%	66	25.88%
		Invalid interaction	105	15.51%	20	13.07%	43	15.99%	42	16.47%
		None	175	25.85%	32	20.92%	67	24.91%	76	29.80%

Table 2. The descriptive results of questionnaires about evaluation basis.

Module	Factors	Value	All		Third Year		Second Year		First Year	
			No.	PCT	No.	PCT	No.	PCT	No.	PCT
Evaluation basis	Number of published papers (paper1)	0	448	66.17%	45	29.41%	167	62.08%	236	92.55%
		1	130	19.20%	55	35.95%	60	22.30%	15	5.88%
		2	64	9.45%	31	20.26%	30	11.15%	3	1.18%
		3	26	3.84%	16	10.46%	9	3.35%	1	0.39%
		4	6	0.89%	5	3.27%	1	0.37%	0	0.00%
		5	3	0.44%	1	0.65%	2	0.74%	0	0.00%

3.2. Requirement Analysis

In this study, the research focus was derived from the discussions with two domain experts (E1 and E2). E1 is a higher education administrator whose work involves the management of higher education student groups. He has extensive experience in the academic development of graduate students. E2 is a scholar in higher education management and has been working in the field of graduate education for eight years. He focuses on the cultivation of academic competence in graduate groups and has a strong demand for visualization and visual analysis of the factors affecting graduate academic competence. We interviewed the experts and reviewed the literature to understand that the ultimate expression of the quality of graduate training is in the form of academic publications and graduation planning of the graduate students, but the key factors that can influence academic competence often occur in the process of the cultivation of the graduate students during the period of research. In addition, these multidimensional factors do not affect the academic competence of graduates in isolation, but in a synergistic way. Thus, it is an urgent requirement to accurately identify the multidimensional influencing factors in the graduate training process. In addition, the experts specified that providing analysts with

an intuitive visual comparison of the impact of different factors on academic ability is necessary for analysts to understand the factors and further predict the academic attainment of a research group. In the form of structured interviews, a list of requirements was distilled and driven by the discussions with experts:

R1. Multidimensional Data Tracking and Acquisition for Graduate Training Processes. The key to quantitative analysis of graduate training quality lies in the accurate interpretation of training data. However, most existing studies focus on the training results of graduate students, such as the number of published papers and the future plans and employment of graduates, but lack attention to the academic ability training of graduate students at the current stage. Therefore, it is necessary to track and acquire multidimensional data on the actual training process, so as to establish a data foundation for the extraction of key factors for subsequent training management.

R2. Extraction of Key Factors for Graduate Training. Based on the data of the graduate training process, it is still a tough task to discover the latent key factors. The extraction of key factors can help educators explore and observe the graduate training process from multiple perspectives and multi-dimensional data, grasp the core factors affecting the academic ability of graduates, and provide important management ideas for administrators to improve the quality of education training.

R3. Visual Evaluation of Graduate Training Factors. A comprehensive and intuitive understanding of the multidimensional data of the graduate training process is essential to accurately analyze and grasp the key factors of training process and improve the quality of training. A visual analytics system will assist education experts in evaluating the key factors by visualizing the multidimensional factors of the graduate training process and comparing the differences between the different factors to analyze and explore the training process.

R4. Interactive Assessment and Recommendations for Graduate Training. When displaying the visual comparison between features of different factors and the subsequent prediction of key factors affecting academic publication, users need to manually select appropriate feature selection and prediction algorithms according to the visual displays to achieve the final evaluation results. To this end, an interactive analysis system will help experts to evaluate key factors in the graduate training process, while enabling a comprehensive presentation of the final academic research predictions, so as to provide a reasonable management direction for education experts for the cultivation of academic ability of graduate students.

3.3. System Pipeline

According to the literature analysis related to the quality of education, the questionnaire can be designed from two aspects: students and supervisors. By putting forward various attributes related to students' scientific research, the data mining methods are used to find out which attributes are related to whether students can publish academic achievements, so as to obtain the main factors affecting the quality of students' academic ability training. The workflow of this paper is shown in Figure 1. Firstly, according to the established research objectives, we designed a questionnaire with 20 influencing factors in five modules: basic information of students, academic participation of students, basic information of supervisors, and academic guidance of supervisors and academic achievements (evaluation basis) (**R1**). After pre-processing the questionnaire data, the algorithms of Lasso, Elastic Net, and Orthogonal Matching Pursuit are used to select the features of the original data of the second-year and third-year graduate students, to find out the most important factors affecting the publication of paper (**R2**). Then, we use the classification algorithms such as support vector machine, naive Bayes, logistic regression, random forest, and multi-layer perceptron to train the classifier to verify the effectiveness of feature selection while predicting the future academic publications of first-year graduate students (**R4**). In addition, we provide a set of visual and interactive interfaces that allow users to visually compare algorithm results and interactively conduct analysis to assess suggestions for improving graduate academic ability (**R3,R4**).

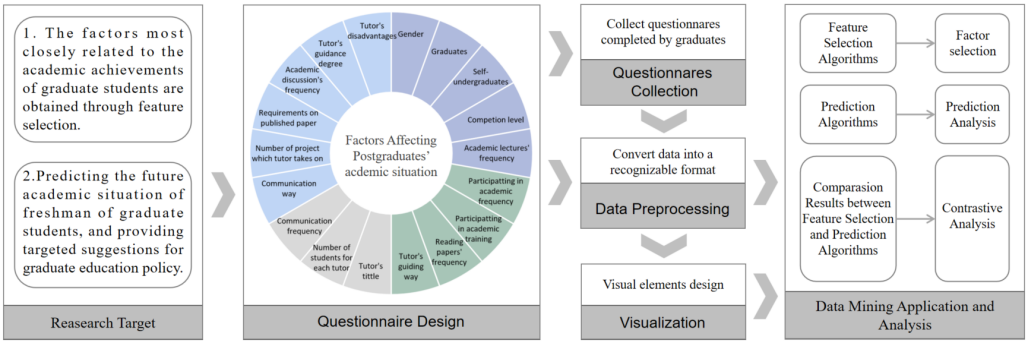

Figure 1. The pipeline of our system for data mining and visual analysis of graduates' training process.

4. Data Mining Methods for Graduates' Cultivation

4.1. Visualization and Interaction

A graduate questionnaire system has been developed that integrates a rich visual and interactive interface for visual comparison and analytical evaluation of feature selection and classification prediction results. The system interface is shown in Figure 2. In the control panel, firstly, we load our pre-processed questionnaire data into the system by selecting the data. In the Factor Comparison View (Figure 2a), the factors in each module are presented in the form of a bubble diagram, where the outermost bubble represents each module, the middle the factors in this module, and the innermost the options within this factor. The bubble plots enable a visual comparison of the individual factors and the percentage of options within them. The size of the bubbles is determined by the weights, with the innermost circle weight calculated from the percentage of people in each option, and the middle and outermost ones by feature selection to determine the importance of each factor comprehensively. In the Ranking View (Figure 2b), the top ten factors in terms of importance are calculated using three feature selection algorithms. The factors in the same bar chart are ranked according to their significance and mapped by different colors. When users want to see how the ranking of a factor changes between different bars, they can click on it and the change in ranking will be highlighted. Meanwhile, the "Feature Method Selection" button provides the user with the ability to combine features for different feature selection algorithms. The final combined ranking result is shown in Ranking View via a Nightingale rose chart (Figure 2c), where the size of each sector indicates the level of importance after the combined ranking and the importance level is mapped by colour. In addition, users can click on the "Classification Forecasting Method" button and choose to view the ROC curves of different classification algorithms (Figure 2e) to evaluate the best prediction by comparison. In Figure 2f, the AUC scores for each classification method with different feature selection algorithms and original factors are shown in a radar plot. For ease of observation, we set the inner boundary of the radar plots to 0.5 and the outer boundary to 0.7 as a way of expanding the differences in the AUC values of the methods for comparison. Finally, once the user has selected the feature synthesis method and classification prediction algorithm they wish to adopt, they can click on the "Comprehensive Result Prediction" button to predict the future academic publications of first-year students based on the current survey data of second-year and third-year students. The prediction results will be displayed in the Factor Comparison View (Figure 2d), where the user can select specific numbers by clicking on the legend above for more detailed analysis of prediction.

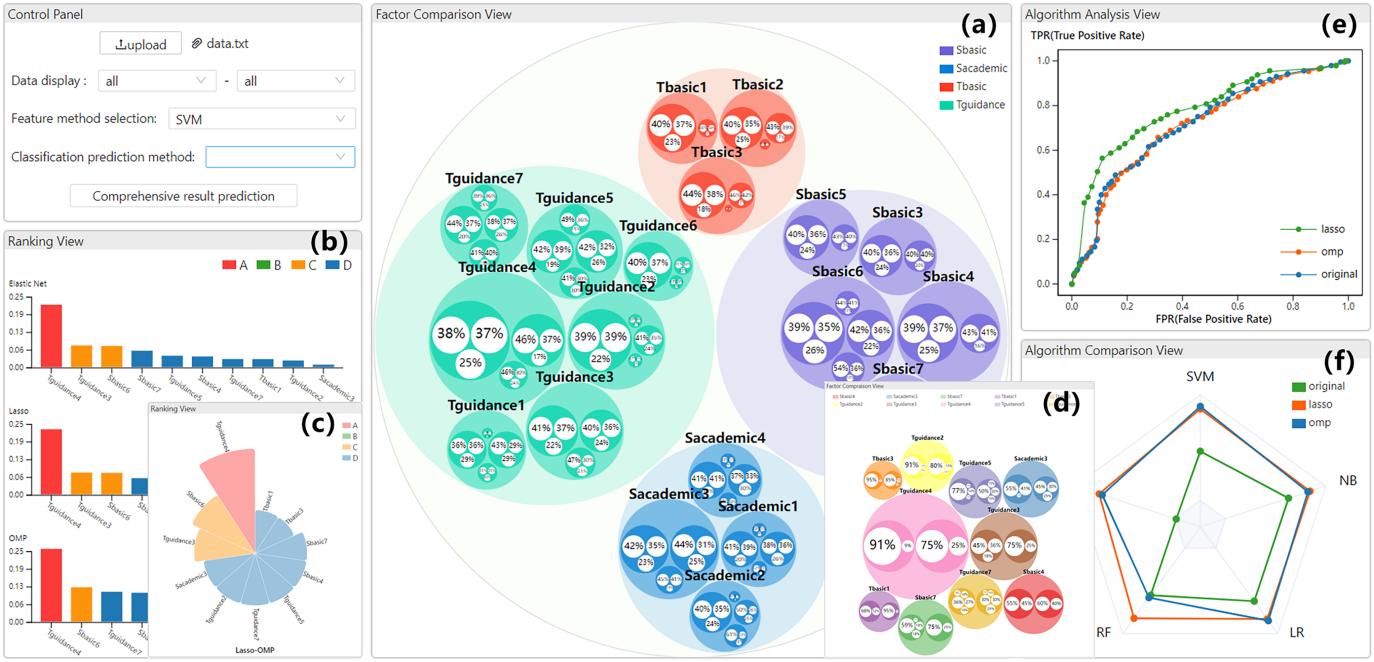

Figure 2. Visual analysis interface for graduate training exploration system. The Factor Comparison View details data of the graduate training process based on the questionnaire (**a**) and shows the predicted results of factors influencing the publication of first-grade graduate students (**d**). The Ranking View shows the comparison of the ranking of the results derived from the three factor selection methods (**b**) and the ranking results of the synthesis of the factor selection methods (**c**). The Algorithm Analysis View shows the ROC curves for the five classifications based on different factor selection methods (**e**). The Algorithm Comparison View compares the AUC values of the five classification methods based on different factor selection methods (**f**).

4.2. Factor Selection and Classification

4.2.1. Feature Selection Algorithms

In order to better uncover the key factors influencing graduate training, classical algorithms such as Lasso, Elastic Net, and Orthogonal Matching Pursuit are chosen to select features for the questionnaire data, respectively. The following is a brief introduction to these feature selection algorithms.

Least Absolute Shrinkage and Selection Operator (Lasso) [34] is a method for compressing the estimated coefficients. It was proposed by statistician Tibshirani in 1996. The basic idea is to construct the L1 penalty function, so that some coefficients with no significant influence after compression are compressed to zero, thus realizing variable selection. Specifically, under the constraint that the sum of the absolute values of the regression coefficients (i.e., the L1 norm) is less than a constant, the sum of the squares of the residuals is minimized, so that some regression coefficients strictly equal to 0 can be generated, which can be explained.

Elastic Net (EN) [35] is based on Lasso to add the L2 norm as a regression model for a priori regular term training. This combination allows for learning a model with only a few parameters that are non-zero sparse, just like Lasso, but because the impact of the penalty of L2 norm, it still maintains some regularity like Ridge regression (Hastie). Elastic networks are useful when multiple features are related to another feature. Lasso tends to choose one of them at random, while elastic networks prefer to choose two.

Orthogonal Matching Pursuit (OMP) is a classical reconstruction algorithm in signal processing and compression sensing domain [36]. It decomposes the signal on the complete dictionary library. The essence is also the optimization problem of the L1 norm. It can effectively and quickly find the sparse solution and realize the selection of variables and features.

4.2.2. Classification Algorithms

In order to verify the validity and rationality of the feature selection results, and improve the accuracy of our predictions, we then use a variety of machine learning methods to train classifiers using the 11 features selected by Lasso and OMP methods and all of the features of the original data, respectively. The classification prediction methods selected in this paper cover four categories: decision tree, regression, neural network, and statistical methods, including Support Vector Machines (SVM), Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP). These classification methods are presented as follows:

Support Vector Machines (SVM) was first proposed by Cortes and Vapnik in 1995. It shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition, and can be applied to other machine learning problems, such as function fitting [37] . It is based on the VC dimension theory of statistical learning theory and the principle of structural risk minimization, based on the complexity of the model information (i.e., the accuracy of learning for specific training samples) and learning ability (i.e., without error). Find the best compromise between the ability to identify any sample in order to obtain the best promotion. Essentially, the SVM algorithm is a class classifier, a hyperplane that separates different classes of samples in the sample space. This means that, given a number of labelled training samples, the SVM algorithm outputs an optimally separated hyperplane.

Naive Bayesian is a classification method based on Bayes' theorem and feature condition independent hypothesis [38]. The basic idea is to solve the items to be classified under the conditions of this occurrence. The probability, which is the largest, is considered to be the category to which the item to be classified belongs.

Logistic Regression, although with regression label, logical regression is a classification algorithm that fits the data to a logit function (or logistic function) to predict the probability of an event occurring [39]. It starts with a linear regression with a continuous value that has practical significance, but linear regression has no way to accurately and robustly segment the problem of classification. Therefore, an algorithm such as logistic regression is designed, and its output is characterized. The probability that a sample belongs to a category.

Random Forest is a classifier that contains multiple decision trees. It is an important integrated learning method based on Bagging, which can be used for classification and regression. Bagging's strategy is to select n samples from the sample set through resampling (Bootstrap has put back repeated samples); train a weak classifier for these n samples (can be ID3, C4.5, CART, SVM, LR). The method is repeated; the above two steps are repeated m times to obtain m weak classifiers; the data are placed in the m classifiers, and the prediction results of the data are determined according to the voting results of the m classifiers. The random forest is an improvement based on Bagging. The corresponding strategy is to select n samples from the sample set using Bootstrap sampling. Select k attributes from all the attributes, and then use the information gain and Gini index method to find the most. The good segmentation attribute establishes the CART decision tree (also SVM, LR, etc.), where k controls the degree of randomness introduction, repeats the above process to establish m classifiers, uses these trees to form random forests, and obtains predictions by averaging results.

Multi-Layer Perceptron (MLP) is a feed forward artificial neural network model that maps a set of input vectors to a set of output vectors [40]. It is an important artificial neural network. The network consists of an input layer composed of sensing units, one or more layers of computing nodes to form a hidden layer, and one layer of computing nodes to form an output layer. The input signal propagates forward through the network on a progressive basis, so it is called a multilayer perceptron. Each neuron model in a multilayer perceptron network includes a nonlinear activation function. A common form of application that satisfies nonlinear requirements is the sigmoid nonlinear function defined by the logistic function.

5. Evaluation

In this section, the final results of the questionnaire data are analysed in terms of feature selection, classification and prediction. Then, suggestions are made for the three groups of postgraduate students, supervisors and university administrators based on the results of the analysis.

5.1. Selection Results

We used Lasso, Elastic Net, and OMP methods to filter out the top 10 factors of influence on publication from the 19 factors in the data (except for paper publication), respectively. In order to facilitate the next analysis and comparison, we adopted the interval equivalence division to divide the results into A, B, C, D four grades for ranking and mapped with different colors (A-red, B-green, C-orange and D-blue).

As shown in Figure 3a, the 10 factors selected by each of the three algorithms are not identical. The ranking order of the factors selected by Lasso and Elastic Net is roughly the same, with the only difference from the order of Tguidance5 and Sacademic4 (Figure 3b). Both factors were ranked as D (mapping in blue) in the ranking process, so the difference in ranking had little impact on the final composition. Therefore, we focused on the analysis of Lasso and OMP results. The order of the factors filtered by Lasso and OMP differed more markedly. With the exception of factors Tguidance4 and Sbasic7, which do not change in ranking, all other factors increased (e.g., Tguidance7) or decreased (e.g., Tguidance3) (Figure 3c). In addition, Tbasic3 in OMP and Tbasic1 in Lasso are not reflected in the results selected by the other algorithm, respectively, which is the main difference between the two algorithms' filtering results.

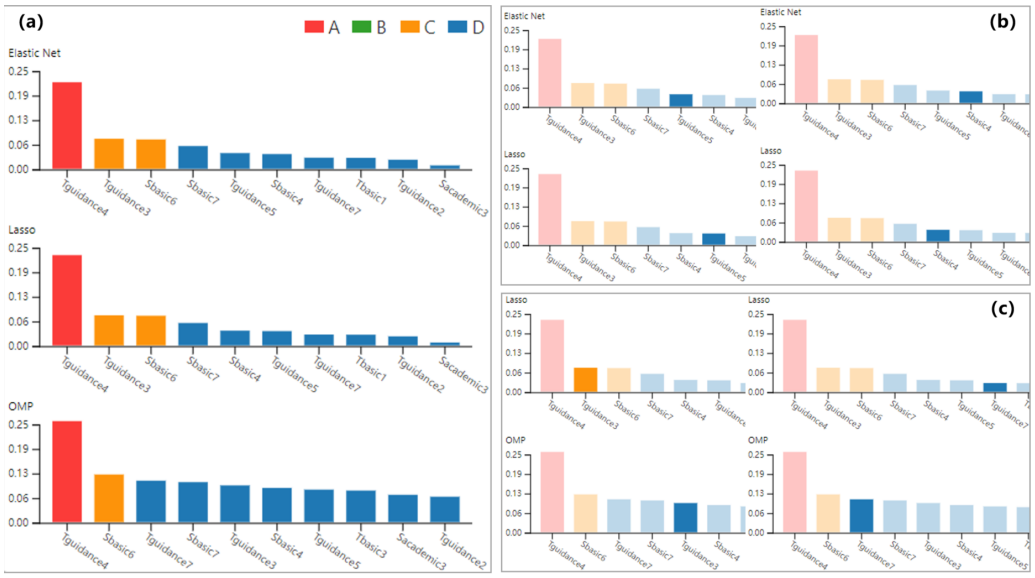

Figure 3. A presentation of the ranking results in three factor selection methods, i.e., Lasso, EN, and OMP (**a**). When the factor that users are interested in is selceted, the other factors' colour will be lightened to highlight the selected one. (**b**) shows the comparison of changes in factors ranking (Tguidance5 and Sbasic4) between Lasso and EN, which presents a small change in ranking. While, (**c**) shows the large changes in factors ranking (Tguidance3 and Tguidance7) between Lasso and OMP.

In order to combine the results of the two feature selection methods, the attribute set is added to the same degree, that is, in the Lasso and OMP, the attributes of A, B, C, and D grades are assigned 8, 6, 4, and 2 scores (0 points are not selected for filtering), and then the average value is obtained. For example, Tguidance3 has a score of 4 in Lasso, a score of 2 in OMP, and the final score of Tguidance3 is 3 on the average of 2 and 4. The weighted scores are then used to re-rank the importance of the factors. Thus, we obtain the following division [1, 2.75]-D, [2.75, 5.5]-C, [5.5, 6.25]-B and [6.25, 8]-A.

As can be seen from Figure 4, 11 relatively important factors were finally selected from the set of 19 factors, namely "Supervisor's requirements for paper publication", "Participation of supervisor's projects", "Plan after graduation", "Participation of academic competition", "Recent graduates", "Frequency of academic discussion", "Supervisor's disadvantages", "Communication way", "Participation of academic training", "Number of students for each supervisor" and "Supervisor's guiding way". As there is no B-level factor, the A-level factor "Supervisor's requirements for paper publication" plays a decisive role in whether the graduates can publish or not. For postgraduates intending to publish, they should select a supervisor who has a strong requirement for their papers. Four of the eight C-level factors are related to the students themselves, i.e., "Plan after graduation", "Participation of academic competition", "Recent graduates" and "Participation of academic training", which shows that the higher the level of participation in academic competitions, lectures and professional academic training, the higher the likelihood of scientific publication. The remaining four relate to supervisor guidance, which are "Participation of supervisor's projects", "Frequency of academic discussion", "Supervisor's disadvantages", and "Communication way", which reveal that the more frequently a graduate student is involved in supervisor's projects or academic discussions, the more likely he publish papers. Finally, there are two factors at D-level, both related to supervisor guidance, i.e., "Number of students for each supervisor" and "Supervisor's guiding way". This can be used to advise universities about graduate admissions without focusing too much on gender or whether they are their own undergraduates.

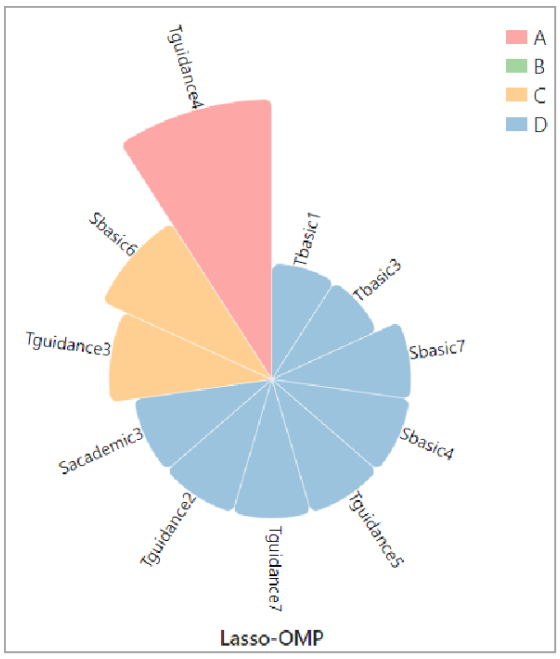

Figure 4. Ranking results from the synthesis of two factor selection methods (Lasso and OMP).

As a result, it is important for graduates to participate as much as possible in their supervisor' projects and academic activities to fully exercise their skills. At the same time, the problems existing in the supervisor's guidance, i.e., the excessive number of graduate students, the supervisor is too busy, the requirements are too high, and the problems in communication, etc., still have a greater impact on paper publication, which needs to be receive more attention by both supervisors and graduates.

5.2. Classification and Prediction

5.2.1. ROC Curve and AUC Value

We use the ROC curve and AUC value to further prove the rationality of our feature selection results, and find the prediction model that best fits our data structure in the comparison of various model effects.

The ROC (Receiver Operating Characteristic) curve is a curve drawn on a two-dimensional plane. The abscissa is the false positive rate (FPR) and the ordinate is the true positive rate (TPR). The ROC curve shows the trade-off between TPR and FPR. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). Classifiers that give curves closer to the top-left corner indicate a better performance.

Figure 5 shows the ROC curve results of five classification models based on 11 characteristics selected by Lasso and OMP as well as all the features of the original data, respectively. We can intuitively obtain two conclusions. Firstly, the performance of the ROC curve of the data through the feature selection is significantly improved than the one of the original data, which proves the rationality and validity of our feature selection results. Secondly, the classification performance of SVM is better than other models, which means that we could choose SVM to be the predictive model for the following research on the possibilities of paper publication by the first-year students.

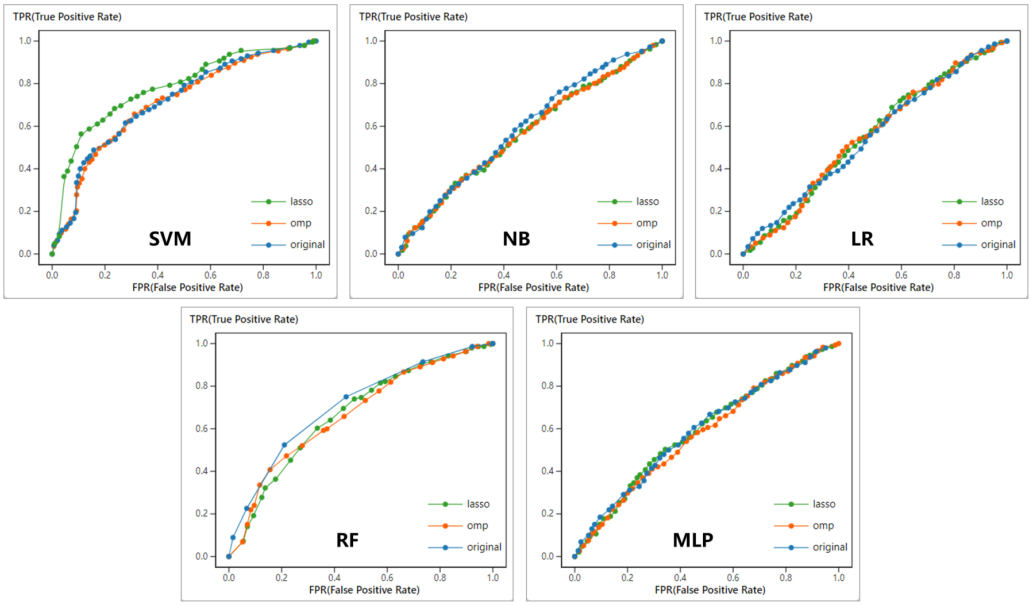

Figure 5. ROC curve for five classification methods.

Except for the intuitive observation of using SVM as the predictive model from the performance of the ROC curve, we also use the AUC value as another criterion for judging the quality of the models. AUC (Area Under Curve) is a standard used to measure the quality of a classification model. The AUC value is the area covered by the ROC curve. The larger the AUC, the better the model's classification effect. AUC = 1 is a perfect classifier. When using this predictive model, a perfect prediction can be made no matter what threshold is set. However, in most cases of prediction, there is no perfect classifier. When the AUC value is between 0.5 and 1, the prediction result is better than the random guess, and the classifier may have the value when the threshold is properly set.

Figure 6 shows the AUC values of our models as the criterion for classification accuracy. It can be seen that the original data have the smallest AUC value, which further verifies the effectiveness of our method for feature selection. In addition, from the AUC values of the five classification methods, SVM performs significantly better than others when classifying the features selected by Lasso and OMP. This indicates the same conclusion that it is more appropriate to choose the SVM model when predicting.

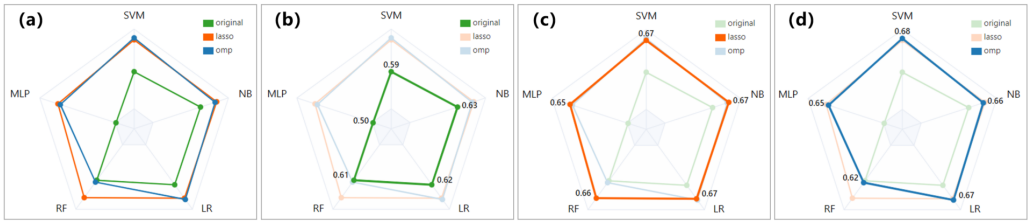

Figure 6. AUC values for five classification algorithms. (**a**) shows the composite comparison and (**b**–**d**) show the specific AUC values for the different methods, i.e., original, Lasso, OMP.

5.2.2. Prediction Results

Combining the above feature selection results with the classification results, we selected the SVM algorithm with the highest accuracy to predict whether first-year graduates would be able to publish papers by integrating the 11 features selected from the results of Lasso, Elastic Net and OMP. In the meantime, considering the aim of this study is to explore the factors influencing the academic situation of graduates and to use publication as a measure, it is more practical to focus on the group of graduates who are more likely to publish papers, i.e., first-year graduates whose plan after graduation is to obtain a PhD. Thus, we conduct a comparative analysis of the specific factors that predict whether or not a paper will be published for these graduates. If not, the reasons why they are unable to publish are analyzed and suggestions are proposed for them. The results are shown in Figure 2d. To further discuss the predictive effect, we select one typical factor in each of A, C and D grades for detailed analysis.

Supervisor's requirements for paper publication (Tguidance4). As shown in Figure 7a, all of the graduate students who can publish papers are more or less required for paper publication from their supervisors, and 91% of whom have supervisor's guidance as well, whereas 25% of the graduates who cannot publish papers have no requirements from their supervisors. Thus, it can be concluded that the lack of requirements of supervisors is the key reason why these graduates cannot publish papers.

Participation of supervisor's projects (Tguidance3). As shown in Figure 7b, the majority of graduate students who are able to publish papers involved in their supervisor's projects, with only 18% not involved. In contrast, 75% of those who are unable to publish do not participate in their supervisor's projects, and the remaining 25% take part in only one of their supervisor's projects. This shows that deep participation in supervisor's research projects can effectively improve the academic ability of graduates.

Frequency of academic discussion (Tguidance5). As shown in Figure 7c, 77% of graduate students who are capable of publishing papers have discussions with their supervisors as frequently as once a week. Nevertheless, as many as 50% of those who are unable to publish have discussions with their supervisors less frequently than once a week, and 20% of them do not talk to their supervisors at all. This suggests that a high frequency of communication and discussion can be of great help in promoting the research progress of graduates.

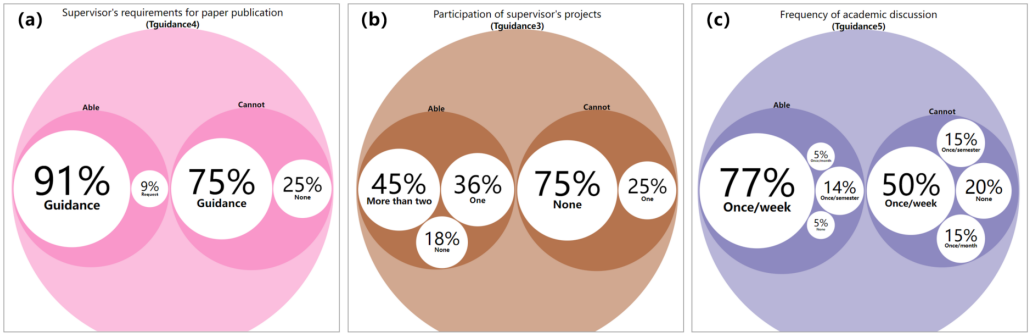

Figure 7. Predicted results of whether first-year graduate students will be able to publish papers. (**a**–**c**) show the results for typical factors in grades A, C and D, respectively, i.e., Supervisor's requirements

for paper publication (Tguidance4), Participation of supervisor's projects (Tguidance3) and Frequency of academic discussion (Tguidance5).

In summary, there are some specific suggestions for the first-year graduates. Firstly, choose a face-to-face interaction with your supervisor and be brave enough to ask questions as frequently as possible. Next, actively participate in your supervisor's projects, not too much, but in a way that allows you to fully exercise yourself and gain academic knowledge. Finally, take part in academic training and lectures; although there is not a strong positive correlation between this and publication, the more times the better in terms of the prediction results.

5.3. Advice for Graduates, Supervisors, and University Administrators

Combining the above analysis, we can make the following suggestions to the three groups of graduate students, supervisors, and university administrators.

5.3.1. For Graduate Students

It is important to face up to the academic participation. The supervisor and your original situations are not the primary factors affecting the publication of papers, and you should actively participate in academic research.

From graduates basic information: A plan after graduation is important for the development of graduates' academic skills, and it is more likely to publish papers for those who intend to pursue their studies for a PhD. Furthermore, as a student's level of participation in a disciplinary competition increases, it is more possible that the student will publish the paper. This seems to be contrary to the conventional wisdom that writing a paper takes a lot of time and energy, and so does competition. There are fewer people who do both, but this research shows that participating in discipline competitions can also promote the writing and publication of a paper to some extent. Participating in high-level competitions can help graduate students obtain good training in teamwork, document writing, practical ability, logical analysis, and other aspects. In addition, participation in academic training and lectures plays an important role in publishing as well. Academic training is a great way to become focused and in-depth cultivation about a certain category in a short period of time, which can strengthen learning from class to practice.

From the interaction between graduates and their supervisors: Firstly, be as proactive as possible in contacting your supervisors and getting involved in their projects or academic research. A deeper understanding of subject knowledge and an improvement in academic skills will be obtained in practice. Secondly, the results of the feature selection suggest that a single supervisor is beneficial for publication, since a single supervisor knows his graduate students more specifically and can arrange them in relation to their own circumstances better. Therefore, it is recommended that graduates who wish to pursue further study choose a single supervisor, and actively communicate with their supervisors.

5.3.2. For Supervisors

It is suggested that the supervisors give students more opportunities to participate in research projects and provide more guidance to students on their papers to develop their academic skills. The supervisor's requirements for graduates to publish papers also play a role in influencing graduates' academic research to a certain extent. Moreover, it is necessary to focus on the communication situation with graduates and hold regular laboratory meetings.

5.3.3. For University Administrators

Advice on admissions: Do not pay too much attention to gender and whether a student is an undergraduate at this university, which has little relevance to academic ability promotion. Instead, appropriate consideration can be given to whether or not the student is a recent graduate. In general, recent graduates find it easier to adapt to graduate life.

Advice on organizing activities: It is recommended that more academic training and more disciplinary competitions be organized to give graduates the opportunity to undertake academic training and to exercise their practical skills. In addition, more projects for supervisors should strive to give students more opportunity to practice their professional knowledge.

6. Conclusions and Future Work

Graduate education is a form of education in which students continue to pursue further study after an undergraduate degree. It belongs to the highest stage of higher education. The cultivation and improvement of graduate students' academic ability is a systematic project, which needs the joint efforts of graduate students, supervisors, colleges, and universities. This paper analyzes the basic attributes of graduate students and utilizes the data mining method to determine the relationship between the research results of students and these attributes, based on which the future publications and development can be predicted.The experimental results show that the classification results after feature selection are better than the raw data without processing, which indicates that the data mining methods can effectively find out the main factors affecting the cultivation of graduate academic ability.

There are still some limitations in our work, such as the selected range of samples being small. In addition, due to the randomness of the model itself, it also brings some errors, which will affect the final results. However, we firmly believe that our research has revealed some of the existing problems in higher education and can also provide feasible policy advice for graduate education. In the future work, we hope that the sample selection can be expanded more comprehensive, so as to avoid the inadequacy and make the research results more adequate.

Author Contributions: Conceptualization, Y.L. and L.Z.; methodology, Y.L. and G.L.; software, Y.Y.; formal analysis, Y.L. and G.L.; investigation, Y.L. and G.L.; resources, Y.L. and L.Z.; data curation, Y.Y.; writing—original draft preparation, G.L. and L.Z.; writing—review and editing, Y.L. and G.L.; visualization, Y.Y.; supervision, Y.L. and L.Z.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Public Welfare Project of Zhejiang Provincial Science and Technology Department No. LGF22F020034 and the Professional Development Program for Domestic Visiting Scholars in Universities of Zhejiang Province No. FX2020028.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Calma, A. Postgraduate research training: Some issues. *High. Educ. Q.* **2011**, *65*, 368–385. [CrossRef]
2. Komarraju, M.; Musulkin, S.; Bhattacharya, G. Role of student–faculty interactions in developing college students' academic self-concept, motivation, and achievement. *J. Coll. Stud. Dev.* **2010**, *51*, 332–342. [CrossRef]
3. Lechuga, V.M. Faculty-graduate student supervisoring relationships: Supervisors' perceived roles and responsibilities. *High. Educ.* **2011**, *62*, 757–771. [CrossRef]
4. Wheeler, L.B.; Maeng, J.L.; Chiu, J.L.; Bell, R.L. Do teaching assistants matter? Investigating relationships between teaching assistants and student outcomes in undergraduate science laboratory classes. *J. Res. Sci. Teach.* **2017**, *54*, 463–492. [CrossRef]
5. Acker, S.; Hill, T.; Black, E. Thesis supervision in the social sciences: Managed or negotiated? *High. Educ.* **1994**, *28*, 483–498. [CrossRef]
6. Amida, A.; Algarni, S.; Stupnisky, R. Testing the relationships of motivation, time management and career aspirations on graduate students' academic success. *J. Appl. Res. High. Educ.* **2020**, *13*, 1305–1322. [CrossRef]
7. Barattucci, M.; Zakariya, Y.F.; Ramaci, T. Academic Achievement and Delay: A Study with Italian Post-Graduate Students in Psychology. *Int. J. Instr.* **2021**, *14*, 1–20. [CrossRef]

8. Kardan, A.A.; Sadeghi, H.; Ghidary, S.S.; Sani, M.R.F. Prediction of student course selection in online higher education institutes using neural network. *Comput. Educ.* **2013**, *65*, 1–11. [CrossRef]
9. Wong, G.K.W.; Li, S.Y.K.; Wong, E.W.Y. Analyzing academic discussion forum data with topic detection and data visualization. In Proceedings of the IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Bangkok, Thailand, 7–9 December 2016. [CrossRef]
10. Nilashi, M.; Abumalloh, R.A.; Zibarzani, M.; Samad, S.; Zogaan, W.A.; Ismail, M.Y.; Mohd, S.; Akib, N.A.M. What Factors Influence Students Satisfaction in Massive Open Online Courses? Findings from User-Generated Content Using Educational Data Mining. *Educ. Inf. Technol.* **2022**, 1–35. [CrossRef]
11. Onwuegbuzie, A.J.; Collins, K.M.T.; Jiao, Q.G. Performance of cooperative learning groups in a postgraduate education research methodology course: The role of social interdependence. *Act. Learn. High. Educ.* **2009**, *10*, 265–277. [CrossRef]
12. Chen, T.; Chang, K.Y. A study on the rare factors exploration of learning effectiveness by using fuzzy data mining. *EURASIA J. Math. Sci. Technol. Educ.* **2017**, *13*, 2235–2253. [CrossRef]
13. Pardos, Z.A.; Kao, K. moocRP: An open-source analytics platform. In Proceedings of the Second (2015) ACM Conference on Learning, Vancouver BC, Canada, 14–18 March 2015. [CrossRef]
14. Yin, X.H. Construction of student information management system based on data mining and clustering algorithm. *Complexity* **2021**, *2021*, 4447045. [CrossRef]
15. Gu, J. An effectiveness model of vocational education mode reform based on data mining. *Int. J. Contin. Eng. Educ. Life Long Learn.* **2022**, *32*, 111–127. [CrossRef]
16. Yuan, J.; Chen, C.; Yang, W.; Liu, M.; Xia, J.; Liu, S. A survey of visual analytics techniques for machine learning. *Comput. Vis. Media* **2021**, *7*, 3–36. [CrossRef]
17. Shah, N.; Bhagat, N.;Shah, M. Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention. *Vis. Comput. Ind. Biomed. Art* **2021**, *4*, 9. [CrossRef]
18. Liu, M.; Shi, J.; Li, Z.; Li, C.; Zhu, J.; Liu, S. Towards Better Analysis of Deep Convolutional Neural Networks. *IEEE Trans. Vis. Comput. Graph.* **2017**, *23*, 91–100. [CrossRef]
19. Ying, Z.; L, G.; H, X.; G, B.; Z, Z.; Q, W.; Y, L.; Y, L.; F, Z. ASTF: Visual Abstractions of Time-Varying Patterns in Radio Signals. *IEEE Trans. Vis. Comput. Graph.* **2022**, 1–11. [CrossRef]
20. Wang, X.; Chen, W.; Xia, J.; Wen, Z.; Zhu, R.; Schreck, T. HetVis: A Visual Analysis Approach for Identifying Data Heterogeneity in Horizontal Federated Learning. *IEEE Trans. Vis. Comput. Graph.* **2022**, 1–10. [CrossRef]
21. Zhao, Y.; Shi, J.; Liu, J.; Zhao, J.; Zhou, F.; Zhang, W.; Chen, K.; Zhao, X.; Zhu, C.; Chen, W. Evaluating Effects of Background Stories on Graph Perception. *IEEE Trans. Vis. Comput. Graph.* **2021**, *28*, 4839–4854. [CrossRef]
22. Zhao, Y.; Jiang, H.; Chen, Q.; Qin, Y.; Xie, H.; Wu, Y.; Liu, S.; Zhou, Z.; Xia, J.; Zhou, F. Preserving Minority Structures in Graph Sampling. *IEEE Trans. Vis. Comput. Graph.* **2021**, *27*, 1698–1708. [CrossRef]
23. Xia, J.; Huang, L.; Lin, W.; Zhao, X.; Wu, J.; Chen, Y.; Zhao, Y.; Chen, W. Interactive Visual Cluster Analysis by Contrastive Dimensionality Reduction. *IEEE Trans. Vis. Comput. Graph.* **2022**, 1–11. [CrossRef] [PubMed]
24. Kumar, A.S.; Vijayalakshmi, M.N.; Koppad, S.H.; Dharani,A. Narrative and Text Visualization: A Technique to Enhance Teaching Learning Process in Higher Education. In Proceedings of the Data Visualization, Singapore, 4 March 2020. [CrossRef]
25. Fahd, K.; Venkatraman, S. Visualizing risk factors of dementia from scholarly literature using knowledge maps and next-generation data models. *Vis. Comput. Ind. Biomed. Art* **2021**, *4*, 19. [CrossRef] [PubMed]
26. Ida, M. Web service and visualization for higher education information providing service. In Proceedings of the 2010 IEEE International Conference on Software Engineering and Service Sciences, Beijing, China, 16–18 July 2010. [CrossRef]
27. Chong, S.; Lee, Y.H.; Tang, Y.W. Data Analytics and Visualization to Support the Adult Learner in Higher Education. In Proceedings of the 2020 The 4th International Conference on E-Society, E-Education and E-Technology, Taipei, Taiwan, China, 15–17August 2020. [CrossRef]
28. Vílchez-Román, C.; Sanguinetti, S.; Mauricio-Salas, M. Applied bibliometrics and information visualization for decision-making processes in higher education institutions. *Libr. Hi Tech* **2020**, *39*, 263–283. [CrossRef]
29. Ngo, L.; Dantuluri, V.; Stealey, M.; Ahalt, A.; Apon, A. An architecture for mining and visualization of us higher educational data. In Proceedings of the 2012 Ninth International Conference on Information Technology-New Generations, Las Vegas, NV, USA, 16-18 April 2012. [CrossRef]
30. Choo, J.; Liu, S. Visual analytics for explainable deep learning. *IEEE Comput. Graph. Appl.* **2018**, *38*, 84–92. MCG.2018.042731661. [CrossRef] [PubMed]
31. Schwab, M.; Strobelt, H.; Tompkin, J.; Fredericks, C.; Huff, C.; Higgins, D.; Strezhne, A.; Komisarchik, M.; King, G.; Pfister, H. booc.io: An education system with hierarchical concept maps and dynamic nonlinear learning plans. *IEEE Trans. Vis. Comput. Graph.* **2016**, *23*, 571–580. [CrossRef]
32. Wei, H.; Li, H.; Xia, M.; Wang, Y.; Qu, H. Predicting student performance in interactive online question pools using mouse interaction features. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany, 23–27 March 2020. [CrossRef]
33. Sundgren, M.; Jaldemark, J. Visualizing online collaborative writing strategies in higher education group assignments. *Int. J. Inf. Learn. Technol.* **2020**, *37*, 351–373. [CrossRef]

34. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2011**, *73*, 273–282. [CrossRef]
35. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
36. Pati, Y.C.; Rezaiifar, R.; Krishnaprasad, P.S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In Proceedings of the 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–3 November 1993. [CrossRef]
37. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–279. [CrossRef]
38. Kononenko, I. Semi-naive Bayesian classifier. In Proceedings of the European Working Session on Learning, Porto, Portugal, 6–8 March 1991. [CrossRef]
39. Hosmer, J.; David, W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*, 3rd ed.; John Wiley & Sons: Amherst, MA, USA, 2013; pp. 35–47.
40. Wu, T.F.; Lin, C.J.; Weng, R. Probability estimates for multi-class classification by pairwise coupling. *Adv. Neural Inf. Process. Syst.* **2003**, *5*, 975–1005.