Building Research Productivity Framework in Higher Education Institution

Ahmad Sanmorino[1]
Faculty of Engineering
Universitas Sriwijaya, Palembang, Indonesia

Ermatita[2]*, Samsuryadi[3], Dian Palupi Rini[4]
Faculty of Computer Science
Universitas Sriwijaya, Palembang, Indonesia

***Abstract*—The purpose of this study is to build a framework for improving research productivity in higher education institutions. The research begins by collecting data and defining candidate variables. The next process is to determine the selected variable from the candidate variable. Variable selection is carried out in three stages, univariate selection, feature importance, and correlation matrix. After the variable selection stage, eight input variables and one target variable were obtained. The eight input variables are Article (C), Conference (CO), Grant (GT), Research Grantee (RG), Rank (R), Degree (D), IPR, and Citation (C). The target variable is Research Productivity (RP). This selected variable is used to build the framework. The next step is to test the framework that has been built. The testing process involves four data mining classifiers, Support Vector Machine, Decision Tree, K-Nearest Neighbor, and Naïve Bayes. The classification results are tested using confusion matrix-based testing, accuracy, precision, sensitivity, and f-measure. The testing results show the proposed framework is able to obtain high accuracy scores for each classification algorithm. It means the proposed framework is relevant to use.**

Keywords*—Framework; research productivity; variable selection; data mining classifier

I. Introduction

Lecturers are the main research actors in a higher education institution. Lecturers are required to conduct research which is one of the three main functions, besides teaching and serving the community. The research achievement target is in accordance with the research scheme chosen by the lecturer. Research results are the targets achieved by researchers from a research activity at the end of the period. Research does not only talk about the quantity of research productivity but also shows the quality of research in a higher education institution [1]. Therefore, the increase in research productivity, both quantity and quality must be measured, in order to know the extent of research progress in a higher education institution [2].

The increasing research productivity is strongly influenced by the environment and the involvement of stakeholders who have an interest in research [3]. This involvement is better known as collaboration. Research collaborations are carried out between one researcher or a group of researchers with other researchers. Each researcher comes from the same or different disciplines, or even different universities [4]. On a wide scale, research collaboration happens between countries, because distance is not a problem now [5]. In recent years, data mining-based knowledge management has been used as the best approach to achieve the goals of an organization with a focus on knowledge creation [6]. One mechanism to increase research productivity is to use a knowledge-sharing approach that involves the role of academics in higher education [7]. The results of this study indicate that the involvement of academics in higher education in research productivity has a variance of 22.6 percent. This shows that the character of academics such as education degree, academic rank, and experience has a considerable influence on research productivity.

In higher education institutions, the data mining approach is the right solution for the analysis of very large research data. Through a data mining approach, researchers know which variables are significant in research productivity. These variables are then used as constructs to build a mechanism for increasing research productivity. The mechanism for increasing research productivity is formulated in the form of a model or framework. The framework development process starts from the preprocessing stage, by selecting the variables to be used. The role of the data mining approach in this case is as a tool for analysis or testing of the framework that has been built. Tests are carried out to determine the performance of the proposed framework. The analysis and testing process involves several data mining algorithms. Furthermore, a comparison of the test results using several data mining algorithms is carried out in order to obtain the best results. The results of this test also show the framework's performance from various points of view, because each data mining algorithm used in testing has different characteristics and approaches. Next is a discussion on research related to research productivity in higher education institutions, followed by an explanation of materials and methods, then results - discussion, and conclusion.

II. Related Work

The framework is defined as mutually supporting parts to achieve a goal. The framework is analogous to a skeleton in the human body that is interconnected, mutually supportive, influencing one another. The framework has a clear direction of achievement, usually illustrated by an arrow to a point. Many researchers have developed frameworks for various needs. In the previous study, researchers built a research productivity framework by combining knowledge sharing and gamification-based variables [8][9]. Another example of developing a framework using knowledge sharing in a higher education environment has been carried out by some researchers [7]. Research productivity is used to determine the

*Corresponding Author

position of higher education institutions on a national and international scale. A mechanism is needed to optimize research productivity. Sample data were taken from tutors to professors in Malaysia with a ratio of 50:30:20 for senior lecturers: assoc. professor: professor.

Through the proposed knowledge management framework (KS), researchers have succeeded in proving that the role of academics, which 12 constructs, has a positive effect on research productivity. The 12 constructs used are commitment, social network, management support, social media, attitude, subjective norm, intention, and behavior, perceived behavior control, facilitating conditions, trust, and research productivity. The results showed that academic productivity has a variance of 22.6 percent. This suggests the academic behavior of KS has a large impact on research productivity. The academic attitude, academic commitment, trust, and social network explain the variance of 36.4 percent. Management support has a variance of 38.7 percent for subjective norms while facilitating conditions and social media have a variance of 26.5 percent for perceived behavioral control. Academics KS intention and KS behavior explain the variance of 62.1 and 47.1 percent, respectively.

The framework is composed of variables that are related to each other. The variable selection process starts with the selection of features from the dataset that has been collected. There are several studies and publications related to research productivity (Table I).

TABLE I. RELATED STUDY WITH RESEARCH PRODUCTIVITY

Author	**Variable Selection Mechanism**	**Algorithm**
Henry *et at.*[10]	Chi-Square, Nagelkerke R Square	Logistic Regression
Ramli *et al.* [11]	Not mentioned	Logistic Regression, Decision Tree, Artificial Neural Network, SVM
Nazri *et al.* [12]	Spearman Rho Correlation	Decision Tree, PART, J-48, C4.5
Wichian *et al.* [13]	Chi-Square, Cronbach Alpha, R-Square	Neural Network Analysis (Back Propagation)
Sanmorino *et al.*	Chi-Square, Extra Tree, Pearson Correlation Co.	SVM, Decision Tree, K-NN, Naïve Bayes

There are several studies related to optimization of research productivity [14]. One of them discusses the gap in the number of professors against other academics, students, or faculty members. In other words, students and faculty members need to be involved in research. The proposed model increases research productivity in higher education institutions. The idea of this model is to involve students and faculty members in intensive research through a curriculum design that focuses on research, which enables students and faculty members to participate in research projects sponsored by the industrial world.

Apart from research, the performance of a lecturer is measured based on the quality of teaching and service to the community. Research related to teacher performance has been conducted [15]. Through this research, several factors associated with teacher performance were tested. The factors that influence teacher performance are currently unclear, so testing is needed to determine these factors. After the various factors are known, they are used to improve the quality of teacher performance in schools.

Researchers propose data mining-based classification and association models, such as decision trees, rule induction, K-NN, and Naïve Bayes to evaluate teacher performance in providing educational services in schools. Some of the attributes used in the test are teacher name, course, class, workspace, training, number of training, and several questions related to teacher performance in schools. The next step is the measurement of accuracy for the data mining method used. In addition to the use of data mining as previously stated, a data mining classifier is also used for various problem solutions such as performance prediction [16][17], performance improvement [18], or decision support system analysis which has been carried out by several researchers [19][20].

III. MATERIAL AND METHOD

The dataset in this research has been collected by the Ministry of Research and Technology Republic of Indonesia through the Science and Technology Index (SINTA) platform. SINTA was launched and has been actively used by academics since 2017. SINTA provides access to citations and scientific expertise in Indonesia. On its official website, SINTA is referred to as an information system used to measure the performance of researchers, including lecturers, and scientific journals in Indonesia. Apart from that, SINTA is a web based platform which is very easy to use. Another reason is because SINTA as an online database accommodates research data from lecturers from all over Indonesia, which is needed to carry out this research. The SINTA platform is equipped with a rating system for researchers and journals in Indonesia [21].

The framework testing process will use a data mining approach. In this study, data mining algorithms were used to measure the performance of the conceptual framework. Another goal is to find patterns and relationships between variables in the dataset. To accommodate the testing stage, this study applied the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [22]. There are six stages in CRISP-DM [23], shown in Fig. 1.

A. Business Understanding

Business understanding is the first stage in CRISP-DM. At this stage, knowledge of business objects is required, an understanding of the scope of the problem, and how to obtain data. Activities undertaken in the business understanding stage include: (1) clearly defining goals and specifications, (2) translate goals and specifications, and (3) determine the boundaries of data mining problems. The next step is to prepare an initial strategy to achieve the goals.

B. Data Understanding

The data understanding stage begins with data collection, identifying data types, qualitative or quantitative, and measurement levels such as nominal, ordinal, binary, and interval [24]. At this stage an understanding of the dataset is needed, to determine properties such as variables or attributes used in modeling.

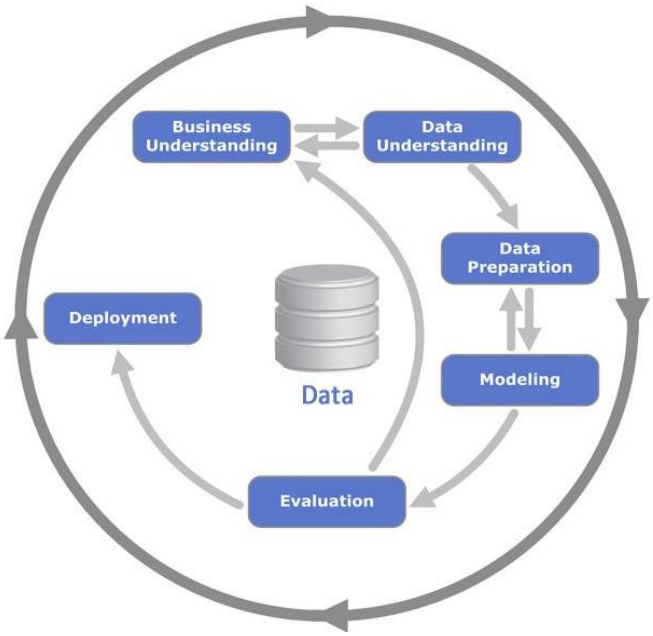

Fig. 1. CRISP-DM Methodology.

C. Data Preparation and Modeling

This stage begins with the identification of the variables used to build the framework. This process focuses on identifying significant variables toward the target variable and removing irrelevant or less important variables from the dataset. Irrelevant variables have a negative impact on the overall model performance. The details of the data preparation and modeling stages are shown in Fig. 2.

Variable selection is one of the core concepts which greatly affect the performance of the data mining model. Some of the advantages obtained by doing variable selection are: (a) reducing overfitting, (b) reducing training time, and most importantly, (c) increasing accuracy. There are three stages of variable selection carried out in this study: (a) univariate selection, (b) feature importance, and (c) correlation matrix.

In the univariate selection stage, the Chi-Square statistical test is used. Chi-Square is used to test the relationship between two variables. In other words, Chi-Square is used to measure how strong the relationship between variables [25][26]. In this study, the relationship tested is between the input variables and the target variables. Variables with a significant relationship value are used for the constructs of the framework. The character of Chi-Square always has a positive value. The formula for Chi-Square is:

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Where, c = degrees of freedom, O = observed value(s), and E = expected value(s). If data from two variables are given, the observed number (O) and the expected number (E) are obtained. Chi-Square measures the deviation between the expected number E and the observed number O.

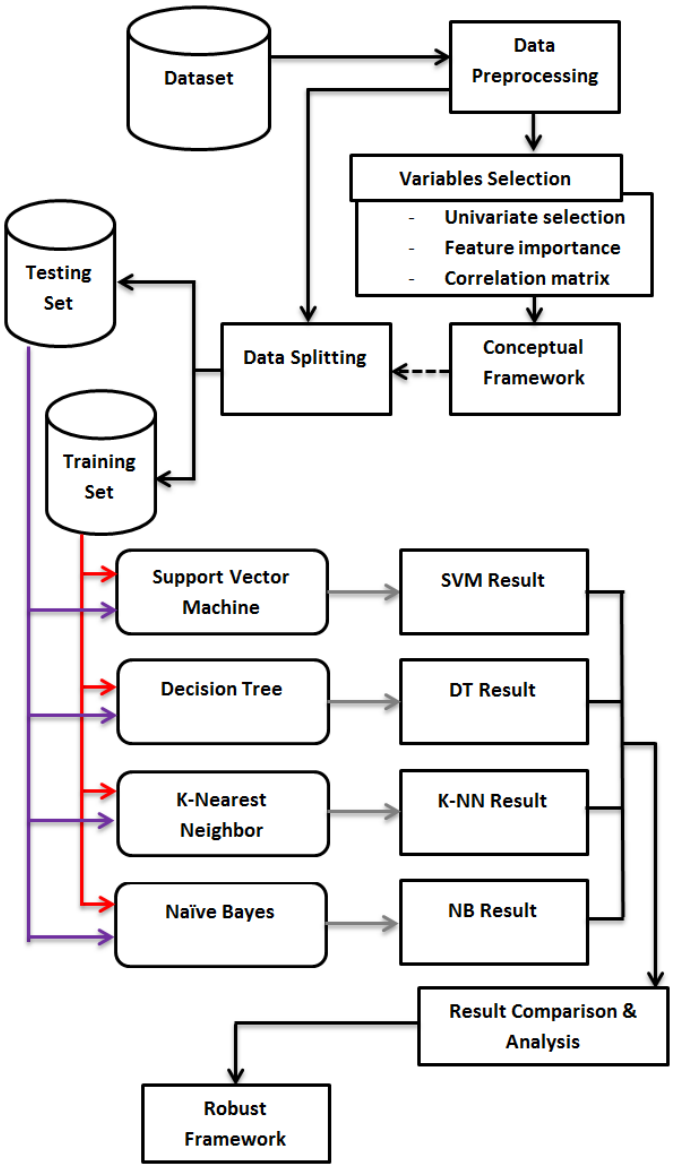

Fig. 2. Data Preprocessing and Modeling.

After the univariate selection stage, it is followed by the feature importance stage. Feature importance is similar to information gain, which extracts the information level (weight) of a feature or variable [27][28]. The results of the selection using feature importance show the score for each variable. The higher the score of a variable, the more relevant or important it is to the target variable. Feature importance uses a Tree-based classifier. In this study, the Extra Tree Classifier used to extract the important variables from the prepared dataset [29]. The correlation matrix shows the correlation between input variables with other input variables or input variables to the target variable. Correlation can be positive if an increase in the Input variable has an impact on an increase in the target variable, or conversely, an increase in the input variable decreases the target variable. Unlike univariate selection (Chi-Square), the correlation matrix can be negative. The correlation matrix test is usually visualized with a heat map. The heat map shows the variables most

related to the target variable and vice versa. After obtaining the relevant variables, the next steps are broken down into two stages: (a) building a conceptual framework, (b) dividing the sample. The sample will be divided into two parts, with a ratio of 70:30, 70 percent for training, and 30 percent for validation. Training and testing data are used as modeling input. This modeling stage is a test for the conceptual framework. In this testing phase, four data mining algorithms are used, Support Vector Machine (SVM), Decision Tree, K-Nearest Neighbor, and Naïve Bayes. This testing phase shows the framework's performance from various points of view because each data mining algorithm used in testing has different characteristics and approaches. The next step is to compare the test results to get the best results.

D. Evaluation and Deployment

The confusion matrix is used to determine the best model. By looking at the confusion matrix value, the accuracy of each model is known. Classification is included in supervised learning, which is a predictive model where the prediction results are discrete. The way to measure the performance of the classification model is to compare the actual value with the predicted value. The confusion matrix is a performance measurement for machine learning classification problems, where the output is two or more classes [30]. The Confusion Matrix is a table with four different combinations of predicted and actual values [31]. There are four terms that represent the results of the classification process in the confusion matrix, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on the Confusion Matrix, the formula for accuracy is obtained:

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (2)$$

Accuracy shows how accurate the model is in classifying correctly.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (3)$$

Precision shows the accuracy between the actual data and the prediction results displayed by the model.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (4)$$

Recall or sensitivity shows the success of the model in retrieving information.

$$F - Measure = \frac{(2*Recall*Precision)}{(Recall+Precision)} \quad (5)$$

F-Measure (f1-score) shows the weighted average comparison of precision and recall [32]. Accuracy is appropriate to use as a reference for the performance of the classification method if the dataset has a very symmetric amount of FN and FP data. However, if the numbers are not symmetric, it is suggested to use the F-Measure as a reference.

IV. Result and Discussion

At an early stage, candidates for the variables are defined as shown in Table II.

The next step is an analysis of the candidate variables. The analysis focuses on the relevance and ease of obtaining data for each variable. Based on the analysis, there are several variables that cannot be used: (1) Working hours, the obstacles faced are difficulties in getting information about working hours, (2) Marital status, because this information is personal, so researcher prefers not to use it, (3) SINTA's score, is optional because the calculation of the score comes from the number of articles and the number of citations, which the variables have been determined, (4) Experience, there is no valid data yet for research experience. After defining the variables, the next step is variable selection.

TABLE II. Candidate Variable

Variable Name	Variable Description	Measurement Level
Degree (D)	Lecturer education degree	Nominal (Master, Doctor)
Gender (G)	Lecturer's gender	Binary*(*Male, Female*)*
Working hours (WH)	Type of working hours	Ordinal (Part time, Full Time)
Nationality (N)	Nationality	Nominal (Indonesia, Non Indonesia)
Rank (R)	Lecturer's rank	Ordinal (Lecturer, Assist Prof, Assoc Prof, Full Prof)
Marital Status (MS)	Lecturer's marital status	Nominal (Single, Married, Widowed)
Conference (CO)	The total number of attended conferences	Ordinal (Never, Ever, Often)
Article (A)	The total number of published articles on Scopus	Ordinal (None, Very Few, Few, Enough, Much)
Citation (C)	The total number of citations for the published articles on Scopus	Ordinal (None, Few, Many, Very Much)
Intellectual Property Rights (IPR)	The total number of IPR registered	Ordinal (None, Few, Many)
Experience (E)	Research experience	Ordinal (Inexperienced, Short Time, Long Enough, Very Experienced)
Research grantee (RG)	Lecturers who receive research grants	Ordinal (Yes, No)
Grant (GT)	The total number of grants obtained	Ordinal (None, Few, Many, Very Much)
SINTA's score (SS)	Lecturer's SINTA score	Ordinal (Low, Medium, High, Very High).
Research Productivity (RP)	Target variable	Binary (Fulfilled, Not Fulfilled)

A. Univariate Selection

Univariate selection is used to select the variable with the strongest relationship toward the target variable. Chi-Square statistical testing shows the results of the selection in order, as shown in Table III.

The test results show that Article (A) is in the first rank. This shows Article (A) has the strongest relationship toward the target variable, followed by Citation (C), Conference (CO), Grant (GT), and others. The results of this test also show the number of articles and the number of citations,

which play an important role in measuring the research performance of a lecturer. Then, for the two lowest ranks, it turns out that Gender (G) has the weakest relationship toward the target variable. In other words, Gender (G) does not have a significant effect on research productivity. Nationality (N) is in the lowest rank, because all lecturers are from Indonesia. This variable does not make a significant difference to the target variable. Variables with a score below 1 are not used in building the proposed framework, so only eight input variables and one target variable remain.

TABLE III. UNIVARIATE SELECTION

No	Variable Name	Chi-Square Score
1	Article (A)	76.533603
2	Citation (C)	47.256279
3	Conference (CO)	45.680553
4	Grant (GT)	22.205091
5	IPR	5.002761
6	Rank (R)	4.027538
7	Research grantee (RG)	2.538671
8	Degree (D)	1.129551
9	Gender (G)	0.118249
10	Nationality (N)	0.000000

B. Feature Importance

Through the feature importance stage, it is possible to know the importance of each variable. The higher the score of a variable, the more relevant or important it is to the target variable. The results of feature importance using the Extra Tree Classifier are shown in Fig. 3.

The selection of feature importance shows Article (A) is in the first rank, with the value 0.3447, followed by Conference (CO) and Citation (C). This measurement shows Article (A) is the variable most relevant to the target variable. Overall, the test results using feature importance are not different from the univariate selection, where the two lowest ranks are Research Grantee (RG) and Nationality (N). These two variables are the least relevant to the target variable. There is a difference in the bottom two variables between univariate selection and feature importance. As a solution, the third step was carried out, the correlation matrix.

C. Correlation Matrix

Correlation can be positive if an increase in the Input variable has an impact on an increase in the target variable, or conversely, an increase in the input variable decreases the target variable. The correlation matrix results using heat maps are shown in Fig. 4 and Fig. 5. Heat maps showing the correlation between input variables with other input variables or input variables for the target variable. Like the two previous steps, Nationality (N) and Gender (G) have poor correlation with other variables. Even Nationality does not have a correlation (zero correlation) with other variables. Gender (G) still has a correlation. Although the correlation to the target variable is the lowest when compared to others. The correlation of Gender (G) to the target variable is 0.046. Even Gender (G) has a negative correlation with Article (A), Research Grantee (RG), and Grant (GT). For other variables, the correlation to the target variable is still > 0.2 (Table IV).

Article (A) has the highest correlation to the target variable, 0.77, followed by Citation (C) of 0.67. The average score of Article (A) on other input variables is very high, thus increasing its correlation to the target variable. The significant difference compared to the previous stage is that Degree (D) and IPR have a low correlation score. This happens because the correlation of Degree (D) and IPR for other input variables is very low so that it affects their correlation to the target variable. However, it is still fair to use as a construct for the proposed framework.

After getting the input and target variables, the next step are to build the framework. Fig. 6 shows the conceptual framework.

Framework consists of eight input variables and one target variable. The eight input variables are Article (C), Conference (CO), Grant (GT), Research Grantee (RG), Rank (R), Degree (D), IPR, and Citation (C). The target variable is Research Productivity (RP). The framework that has been built must be tested first. The data mining approach was chosen as a testing tool because it is in accordance with the characteristics of the dataset that has been prepared. In this testing phase, four data mining algorithms are used, Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (K-NN), and Naïve Bayes (NB). The classification results using data mining algorithms tested using confusion matrix-based measurement. The test results using the confusion matrix-based measurement (accuracy, precision, sensitivity, f1-measure) are shown in Table V.

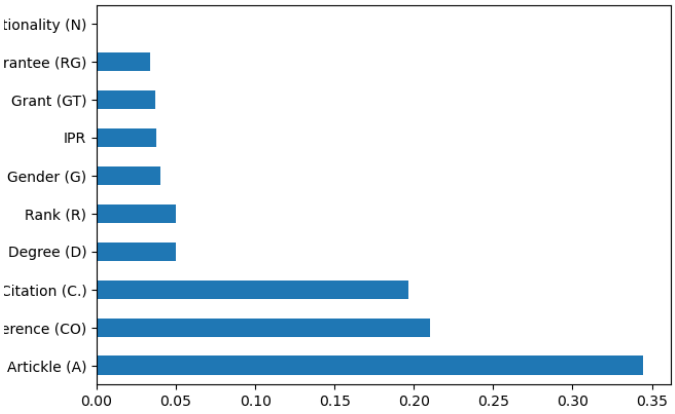

Fig. 3. Feature Importance.

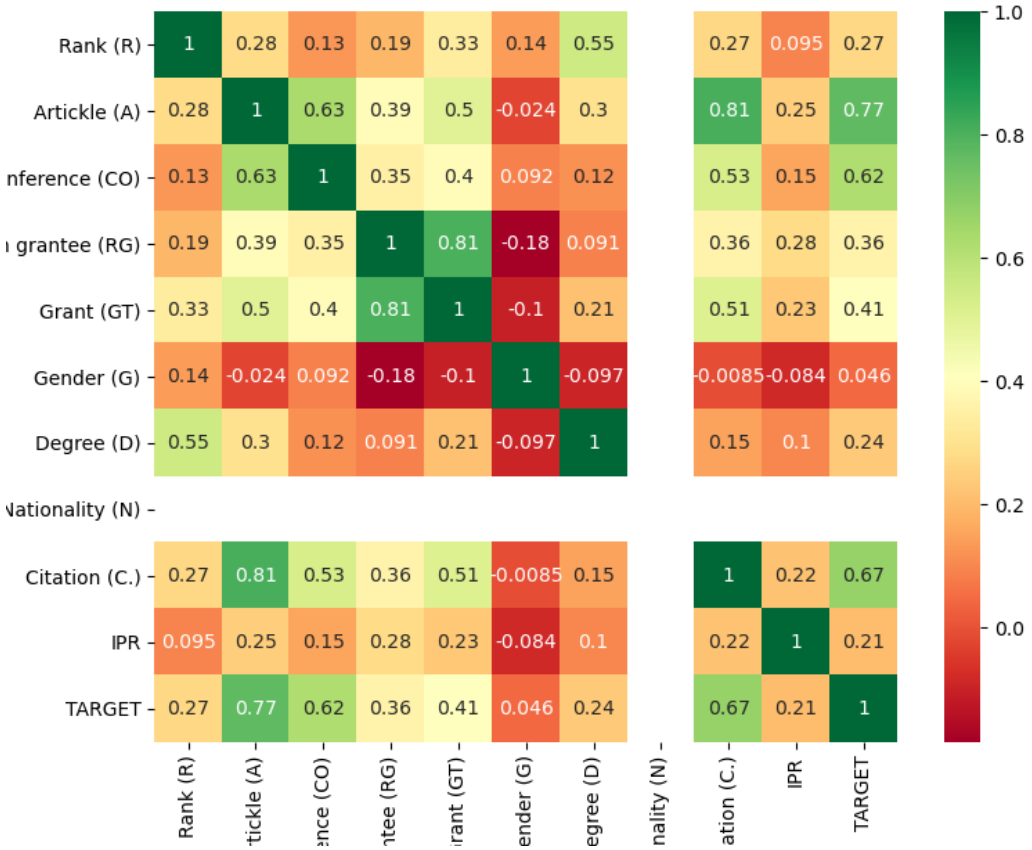

Fig. 4. Correlation Matrix (Include Gender and Nationality).

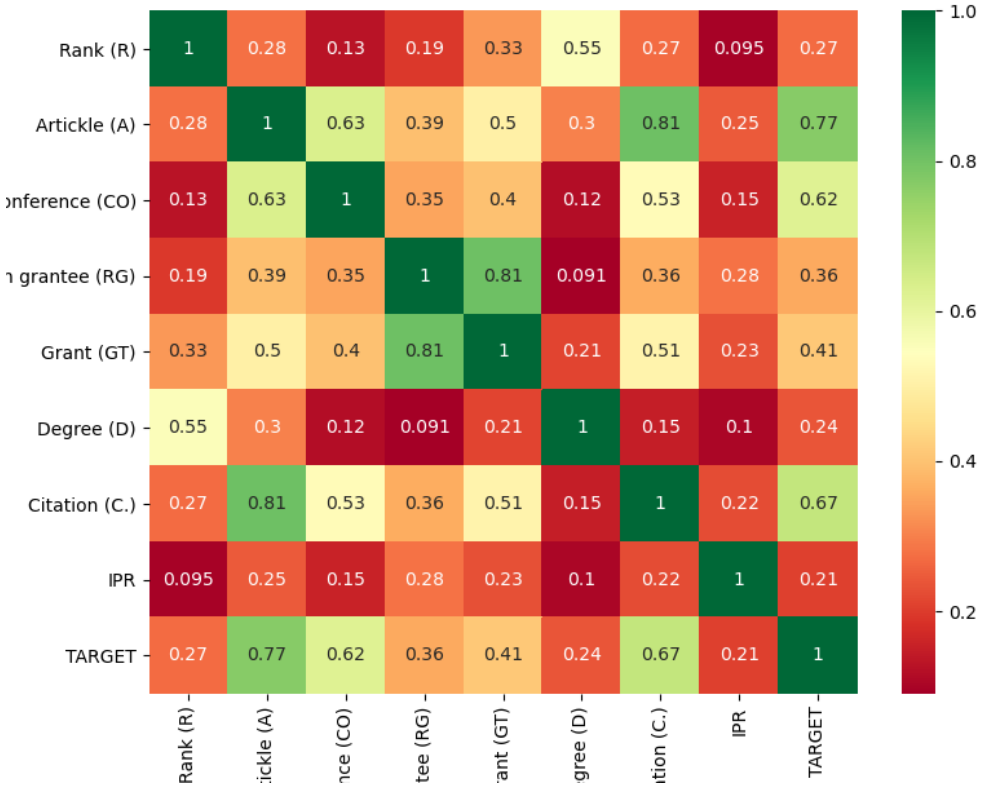

Fig. 5. Correlation Matrix (Exclude Gender and Nationality).

TABLE IV. THE CORRELATION OF INPUT VARIABLES TOWARD TARGET VARIABLE

No	Variable Name	Correlation Score
1	Article (A)	0.77
2	Citation (C)	0.67
3	Conference (CO)	0.62
4	Grant (GT)	0.41
5	Research grantee (RG)	0.36
6	Rank (R)	0.27
7	Degree (D)	0.24
8	IPR	0.21

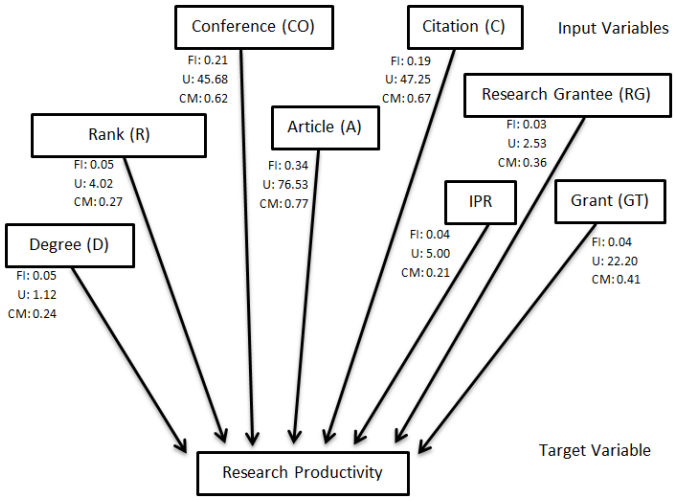

Fig. 6. Conceptual Framework.

TABLE V. ACCURACY, PRECISION, SENSITIVITY, F-MEASURE AND MISCLASSIFICATION RATE FOR 4 ALGORITHMS

Classifier	Accuracy	Precision	Sensitivity	f1-Score	Misclass. rate
SVM	78.26%	85.29%	77.27%	76.67%	21.74%
DT	86.95%	87.12%	87.12%	86.95%	13.05%
K-NN	95.65%	96.15%	95.45%	95.61%	4.35%
NB	86.95%	90.00%	86.36%	86.54%	13.05%

The testing result shows the proposed framework is able to obtain high accuracy scores for each classification algorithm. The highest accuracy score on the K-NN classification algorithm is 95.65 percent, followed by Decision Tree and Naïve Bayes, each with 86.95 percent; the last is Support Vector Machine at 78.26 per cent. Just like the accuracy score, for the measurement of precision, sensitivity, F-Measure, the K-NN algorithm is also the highest, with the lowest misclassification rate, only 4.35 percent. The results of confusion matrix-based testing prove that the proposed framework is relevant to use, with high accuracy scores and little misclassification rate. When compared with the results of other related research tests regarding research productivity, the position of the results of this test is (Table VI).

TABLE VI. THE COMPARISON OF VARIABLES, ALGORITHM, AND ACCURACY OF CLASSIFICATION TESTING RESULTS

Author Name	Variables Used	Algorithm Used	Accuracy
Henry *et at.*[10]	Age Cohort, Highest Qualification, Cluster, Lecturer Track, Achievement, Job Policy, Monthly Income, Research Leadership, Research Supervision	Logistic Regression	78.2%
Ramli *et al.* [11]	Age, Gender, Marital Status, Qualification, Experience, Position, Division, Citation, Article, Conference, and Target (Status of Research Performance)	Logistic Regression	80.31%
		Decision Tree	83.40%
		Artificial Neural Network	82.24%
		Support Vector Machine	80.47%
Nazri *et al.* [12]	Age, Designation, No. Research Grant, Gender, Performance Score, Marital Status, Working Status, Amount of Grant, Department, Administrative Post, No. PhD Student, Faculty, Invitation as Keynote Speaker, Article (Index)	Decision Tree	70.30%
		PART	75.00%
		J-48	75.30%
		C4.5	70.20%
Wichian *et al.* [13]	Age, Academic Position, Thinking, Research Mind, Volition - Control, Meeting of International, Research Skill – Techniques, Research Fund, Research Management, Communication, Networking and Teamwork, Institutional Policy, Library Expenditure, Computing Facility	Neural Network (Back Propagation)	90.72%
Sanmorino *et al.*(this study)	Article, Conference, Grant, Research Grantee, Rank, Degree, IPR, Citation, and Target (Reseach Productivity)	Support Vector Machine	78.26%
		Decision Tree	86.95%
		K-Nearest Neighbors	95.65%
		Naïve Bayes	86.95%

There are differences in the combination of algorithms, variables and the number of datasets used that affect the performance of the classification algorithm, but this study has proven that the framework designed based on the variable selection has a relevant good accuracy score. Researchers cannot say that the results of this test are better than other related studies. To prove the test results of a study are better than other studies, the same scenario must be used, in the sense of where to collect the data, the number of datasets, the mechanism for selecting variables, the number of variables must all be the same, because they can affect the test results.

V. CONCLUSION

The framework development process starts from collecting datasets and determining candidate variables. The next process is to determine the selected variable from the candidate

variable. Variable selection is carried out in three stages, univariate selection, feature importance, and correlation matrix. After the variable selection stage, eight input variables and one target variable were obtained. The eight input variables are Article (C), Conference (CO), Grant (GT), Research Grantee (RG), Rank (R), Degree (D), IPR, and Citation (C). The target variable is Research Productivity (RP). This selected variable is used to build the framework. The next step is to test the framework that has been built. The testing process involves four data mining classifiers. The classification results are tested using confusion matrix-based testing, accuracy, precision, sensitivity, and f1-measure. The testing results show the proposed framework is able to obtain high accuracy scores for each classification algorithm. It means the proposed framework is relevant to use. There are several things recommended for future work, such as increasing the number of datasets, using other variables relevant to research productivity, such as research collaboration, teamwork, or research facilities in a higher education institution.

ACKNOWLEDGMENT

The first author is a doctoral student at the Faculty of Engineering, Universitas Sriwijaya. The authors would like to thank Universitas Sriwijaya for their support in carrying out this research.

REFERENCES

[1] C. N. Tan, "Improving Research Productivity through Knowledge Sharing: The Perspective of Malaysian Institutions," no. October, pp. 701–712, 2015.

[2] G. Abramo, C. A. D. Angelo, G. Abramo, C. Andrea, and D'Angelo, "How do you define and measure research productivity? Scientometrics," no. November, 2014, doi: 10.1007/s11192-014-1269-8.

[3] G. Li et al., "Enhancing research publications and advancing scientific writing in health research collaborations: Sharing lessons learnt from the trenches," J. Multidiscip. Healthc., vol. 11, pp. 245–254, 2018, doi: 10.2147/JMDH.S152681.

[4] Z. Zuo and K. Zhao, "The more multidisciplinary the better ? – The prevalence and interdisciplinarity of research collaborations in multidisciplinary institutions," J. Informetr., vol. 12, no. 3, pp. 736–756, 2018, doi: 10.1016/j.joi.2018.06.006.

[5] D. Press, "How to set-up a long-distance mentoring program : a framework and case description of mentorship in HIV clinical trials," no. January 2013, 2014, doi: 10.2147/JMDH.S39731.

[6] C. Sassenberg, C. Weber, and M. Fathi, "A Data Mining based Knowledge Management Approach for the Semiconductor Industry," no. July, 2009, doi: 10.1109/EIT.2009.5189587.

[7] M. A. Fauzi, C. T. Nya-Ling, R. Thursamy, and A. O. Ojo, "Knowledge sharing: Role of academics towards research productivity in higher learning institutions," VINE J. Inf. Knowl. Manag. Syst., vol. 49, no. 1, pp. 136–159, 2019, doi: 10.1108/VJIKMS-09-2018-0074.

[8] A. Sanmorino, Ermatita, and Samsuryadi, "The preliminary results of the kms model with additional elements of gamification to optimize research output in a higher education institution," Int. J. Eng. Adv. Technol., vol. 8, no. 5, pp. 554–559, 2019.

[9] A. Sanmorino, Ermatita, Samsuryadi, and D. P. Rini, "A Robust Framework using Gamification to Increase Scientific Publication Productivity," Proc. - 2nd Int. Conf. Informatics, Multimedia, Cyber, Inf. Syst. ICIMCIS 2020, pp. 29–33, 2020, doi: 10.1109/ICIMCIS51567.2020.9354319.

[10] C. Henry, N. A. Md Ghani, U. M. A. Hamid, and A. N. Bakar, "Factors contributing towards research productivity in higher education," Int. J. Eval. Res. Educ., vol. 9, no. 1, pp. 203–211, 2020, doi: 10.11591/ijere.v9i1.20420.

[11] N. A. Ramli, N. H. M. Nor, and S. S. M. Khairi, "Prediction of research performance by academics in local universities using data mining approach," vol. 040021, no. August, 2019, doi: 10.1063/1.5121100.

[12] M. Z. A. Nazri, R. A. Ghani, S. Abdullah, M. Ayu, and R. N. Samsiah, "Predicting Academic Publication Performance using Decision Tree .," no. 2, pp. 180–185, 2019.

[13] S. Na Wichian, S. Wongwanich, and S. Bowarnkitiwong, "Factors affecting research productivity of faculty members in government universities: LISREL and Neural Network analyses," Kasetsart J. - Soc. Sci., vol. 30, no. 1, pp. 67–78, 2009.

[14] P. S. Aithal, "Study on Research Productivity in World Top Business Schools," Int. J. Eng. Res. Mod. Educ. ISSN 2455 - 4200, vol. I, no. I, pp. 629–644, 2016.

[15] R. K. Hemaid and A. M. E.- Halees, "Improving Teacher Performance using Data Mining," Ijarcce, vol. 4, no. 2, pp. 407–412, 2015, doi: 10.17148/ijarcce.2015.4292.

[16] V. Vijayalakshmi, K. Panimalar, and S. Janarthanan, "Predicting the performance of instructors using Machine learning algorithms," no. December, 2020.

[17] Q. A. Al-Radaideh and E. Al Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," Int. J. Adv. Comput. Sci. Appl., vol. 3, no. 2, p. 8, 2012, [Online]. Available: www.ijacsa.thesai.org 144.

[18] Q. Zhang, W. Hu, Z. Liu, and J. Tan, "TBM performance prediction with Bayesian optimization and automated machine learning," Tunn. Undergr. Sp. Technol., vol. 103, no. June, p. 103493, 2020, doi: 10.1016/j.tust.2020.103493.

[19] B. Wah, S. Huat, N. Huselina, and M. Husain, "Expert Systems with Applications Using data mining to improve assessment of credit worthiness via credit scoring models," Expert Syst. Appl., vol. 38, no. 10, pp. 13274–13283, 2011, doi: 10.1016/j.eswa.2011.04.147.

[20] H. Jantan, N. M. Yusoff, and M. R. Noh, "Towards Applying Support Vector Machine Algorithm in Employee Achievement Classification," pp. 12–21, 2014.

[21] L. Lukman et al., "Case Study Proposal of the S-score for measuring the performance of researchers , institutions , and journals in Indonesia," vol. 5, no. 2, pp. 135–141, 2018.

[22] C. Industry et al., "Data Science Process," no. 1, pp. 19–37, 2019, doi: 10.1016/B978-0-12-814761-0.00002-2.

[23] K. Jensen, "IBM SPSS Modeler CRISP-DM Guide," 2016.

[24] G. M. Robinson, Statistics, Overview, Second Edition., vol. 13. Elsevier, 2020.

[25] K. Molugaram and G. S. Rao, Chi-Square Distribution. 2017.

[26] F. Girosi and G. King, "Model Selection," Demogr. Forecast., pp. 94–123, 2018, doi: 10.2307/j.ctv301hd6.12.

[27] I. Kareva and G. Karev, "Replicator dynamics and the principle of minimal information gain," Model. Evol. Heterog. Popul., pp. 129–154, 2020, doi: 10.1016/b978-0-12-814368-1.00008-4.

[28] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection," IEEE Access, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.

[29] K. Kaur and S. K. Mittal, "Classification of mammography image with CNN-RNN based semantic features and extra tree classifier approach using LSTM," Mater. Today Proc., no. xxxx, 2020, doi: 10.1016/j.matpr.2020.09.619.

[30] F. Evaluating, P. To, and P. Model, "Model Evaluation," 2015, doi: 10.1016/B978-0-12-801460-8.00008-2.

[31] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification : A measure driven view," Inf. Sci. (Ny)., no. xxxx, 2019, doi: 10.1016/j.ins.2019.06.064.

[32] L. Derczynski, "Complementarity , F-score , and NLP Evaluation," pp. 261–266, 2013.