Vol. 10, No. 1, 2024, pp. 243-249
DOI: https://doi.org/10.29210/1202423794

Contents lists available at Journal IICET

Jurnal EDUCATIO (Jurnal Pendidikan Indonesia)
ISSN: 2476-9886 (Print) ISSN: 2477-0302 (Electronic)

Journal homepage: https://jurnal.iicet.org/index.php/jppi

The application of naive bayes method for final project topic selection within the project-based learning framework in the data mining course

Denny Kurniadi
Universitas Negeri Padang, Indonesia

Article Info

Article history:

Received Feb 15th, 2024
Revised Mar 16th, 2024
Accepted May 25th, 2024

Keyword:

Naive bayes method,
Final project topic,
Project-based learning,
Data mining,
Students

ABSTRACT

This study aims to explore the potential application of the Naive Bayes Method in predicting the determination of final project topics for students within the context of project-based learning (PBL) in the Data Mining course. Adopting an observational quantitative approach, the research involved 34 students from the Educational Technology Informatics Study Program at Padang State University between 2013-2015 as research subjects. The sampling technique employed was stratified random sampling, aligning with the research findings. The study's results indicate that the Naive Bayes Method is proficient in providing accurate predictions concerning the determination of final project topics, with significant accuracy, precision, and recall values. Moreover, the integration with project-based learning (PjBL) approach effectively enhances the authenticity and relevance of the learning experience for students, further substantiating the effectiveness of Project-Based Learning (PjBL).

Corresponding Author:

Denny Kurniadi,
Universitas Negeri Padang
Email: dennykurniadi@ft.unp.ac.id

Introduction

In the realm of higher education, the process of determining topics for final projects stands as a crucial pillar in the academic development of students (Alamri et al., 2020; Ibarra-Sáiz et al., 2020; Santos et al., 2020). It is a juncture where students often face challenges, as highlighted by studies from the Association for Computing Machinery (ACM), which underscore that over 70% of students encounter difficulties in deciding their final project topics (Du et al., 2020). Consequently, there arises a pressing need for a structured and effective approach to aid students in navigating this decision-making process (Gao & Yu, 2020; Secinaro et al., 2021).

Project-based learning, which promotes learning through real-world problem-solving, is gaining increasing favor within higher education (Rosenkranz, 2022). Specifically, within the Data Mining course, students engage in analyzing vast datasets to uncover patterns and trends (Hung et al., 2020; Namoun & Alshanqiti, 2021). Here, the integration of the Naive Bayes Method in predicting final project topics assumes particular importance (Ali et al., 2020). Through this method, students can be guided towards selecting topics aligned with their interests and needs, while instructors or academic advisors can provide more targeted assistance (Klempin & Lahr, 2021).

However, it is essential to delve deeper into the purpose, significance, and novelty of this research. Exploring the potential application of the Naive Bayes Method in predicting the determination of final project topics for students within the project-based learning framework in the Data Mining course is the central aim of this study (Charitopoulos et al., 2020)(Bošnjaković & Đurđević Babić, 2023)(Ikegwu et al., 2023). In doing so, it is imperative to contextualize the importance of final project topics in shaping students' academic trajectories, necessitating an overview of the current state of the field under study.

The research hypothesis posits that the Naive Bayes Method can provide accurate predictions regarding the determination of final project topics (Egwim et al., 2021)(Shafiq et al., 2023), supported by previous studies indicating its effectiveness in this domain (Kukkar et al., 2023)(Egwim et al., 2021). While direct studies on this specific topic may be limited, existing literature in the fields of Data Mining and project-based learning offers a robust foundation for this research (Zotou et al., 2020). Anticipated primary findings of this study promise to offer fresh insights into the application of the Naive Bayes Method in higher education, particularly in supporting academic decision-making (Alyahyan & Düştegör, 2020)(Yağcı, 2022).

The significance of engaging with this manuscript lies in gaining a deeper understanding of how the Naive Bayes Method can be harnessed within higher education, specifically in predicting final project topics for students in the Data Mining course. Through this manuscript, readers are poised to gain insights into the contributions of this research to the field of higher education and its potential to streamline the process of selecting final project topics for students.

Moreover, this manuscript succinctly outlines the significant contributions of this research, emphasizing its innovative aspects that pave the way for further developments in the field. By synthesizing existing knowledge and proposing novel applications, this study endeavors to advance both theoretical understanding and practical implementation within the realm of higher education.

Beyond its academic significance, this research holds practical implications for educational institutions, instructors, and students alike. By leveraging the Naive Bayes Method to predict final project topics, educational institutions can enhance the overall learning experience for students, ensuring that projects align closely with their interests and career aspirations. Additionally, instructors can utilize the insights provided by this research to offer more tailored guidance and support throughout the project selection process, thereby fostering a more enriching educational environment.

Furthermore, the adoption of the Naive Bayes Method in predicting final project topics can contribute to the optimization of resource allocation within educational institutions. By accurately predicting students' topic preferences, institutions can allocate resources more efficiently, ensuring that students have access to the necessary tools and support systems to succeed in their projects. This, in turn, can lead to higher levels of student satisfaction and academic achievement, ultimately benefiting the institution as a whole.

In addition to its practical implications, this research also contributes to the broader theoretical understanding of project-based learning and data mining within the context of higher education. By exploring the intersection of these two fields and proposing a novel methodology for predicting final project topics, this study expands the theoretical framework within which educators and researchers can approach curriculum development and pedagogical practices.

Moreover, by highlighting the potential of the Naive Bayes Method in predicting final project topics, this research opens up new avenues for future inquiry and exploration. Subsequent studies may seek to further refine and validate the predictive model proposed in this research, exploring its applicability across different educational contexts and disciplines. Additionally, researchers may investigate the impact of predicted project topics on student learning outcomes and academic performance, providing further insights into the efficacy of this approach.

Overall, this research represents a significant contribution to the fields of higher education, data mining, and project-based learning. By addressing an important gap in the existing literature and proposing a novel methodology for predicting final project topics, this study advances our understanding of how data-driven approaches can enhance the educational experience for students. Through its practical implications, theoretical contributions, and potential for future research, this research has the potential to shape the way educators and institutions approach project-based learning in the digital age.

Method

This research adopts an observational quantitative approach, which involves systematic observation and the collection of numerical data without direct intervention from the researcher (Creswell & Creswell, 2018). In this

study, student grade data from the Educational Technology Informatics Study Program at Padang State University during the period 2013-2015 is gathered using this method. The observational quantitative approach allows researchers to directly observe phenomena and collect data on the observed variables (Wahyudi et al., 2022). By employing this approach, the study aims to provide an objective analysis of student performance in the Data Mining course, without influencing the natural environment or behavior of the participants.

Participant

The research focuses on recommending final project topics from students enrolled in the Educational Technology Informatics Study Program at Padang State University. The sample comprises 34 students enrolled between 2013-2015, encompassing attributes such as transcripts of grades from semesters 1-4 and field concentration categories. The sampling technique employed is stratified random sampling to ensure a balanced representation of various academic levels and student backgrounds.

PjBL

PjBL is a learning method that provides direct experience through real projects (Almulla, 2020; Nurhidayah et al., 2021). Although effective in enhancing motivation and learning relevance, PjBL also faces challenges such as the need for adequate resources and difficulties in assessment (J. Chen et al., 2021; Sukackė et al., 2022; Ngereja et al., 2020). The implementation strategies of PjBL in this research include: (a) project planning, (b) project definition, (c) student roles, (d) mentoring and support, (e) outcome evaluation, and (f) reflection and learning (Saad & Zainudin, 2022; Fajra & Novalinda, 2020; Maryati et al., 2022).

This project is conducted independently by students. The following are the steps used in the self-directed project-based learning approach (Pan et al., 2021; Dewi & Mirah, 2020; erovnik & Nančovska Šerbec, 2021): 1) Define Learning Objectives: Students should identify their own learning objectives related to understanding the concepts of Naive Bayes, data analysis techniques, and interpreting prediction results; 2) Dataset Selection: Students should choose a dataset that aligns with their interests and needs, ensuring it has sufficient data for training and evaluating Naive Bayes models; 3) Self-Study on Naive Bayes: Students should engage in self-study on the basic concepts of Naive Bayes, including Bayes' Theorem, feature independence assumptions, and its application in classification. They should also learn how to implement the Naive Bayes algorithm using RapidMiner; 4) Project Proposal Development: Students should develop a project proposal outlining their approach to predicting final project topics. The proposal should include dataset selection, data preprocessing steps, Naive Bayes implementation, and model evaluation techniques; 5) Data Preprocessing: Students should independently preprocess the data, including removing irrelevant data, handling missing values, and transforming necessary features; 6) Implementation of Naive Bayes Algorithm: Students should implement the Naive Bayes algorithm independently using RapidMiner. They should train the model, tune parameters, and evaluate their model's performance; 7) Feature Selection and Engineering: Students should perform feature selection and engineering independently to improve the performance of their Naive Bayes classifier. They should analyze the importance of different features and experiment with various feature combinations; 8) Model Evaluation: Students should evaluate the performance of their Naive Bayes model independently using appropriate evaluation metrics. They should interpret the evaluation results and understand the strengths and weaknesses of their model; 9) Project Presentation and Documentation: Students should prepare a final presentation and report documenting their project findings, methodology, results, and conclusions independently. They should present their findings and share their documentation with others; 10) Reflection: Students should reflect on their learning experience, including the challenges they faced, the lessons they learned, and how they would approach similar projects in the future.

Naïve Bayes Algorithm

The Naive Bayes algorithm is a straightforward prediction method based on Bayes' Theorem, assuming independence between predictor variables. Its strengths include simplicity, efficiency, and the ability to handle large datasets with high dimensions (Siva Subramanian et al., 2022). One advantage is its ability to make good predictions even with limited training data, making it useful for tasks with scarce data (Adadi, 2021). Additionally, Naive Bayes is computationally efficient and suitable for real-time applications on large datasets (P. H. Tran et al., 2022). However, it assumes conditional independence between features, which may not always hold true in real-world datasets (Chen et al., 2020), leading to inaccurate predictions, especially when features are highly correlated (Kurniawan et al., 2020). Furthermore, Naive Bayes is considered "naive" because it assumes all features contribute equally to prediction, which may not always be the case (Chen et al., 2020), resulting in poor performance when features have varying levels of relevance to the target variable (Dixit et al., 2020). Despite these limitations, Naive Bayes remains popular for prediction tasks, particularly in domains like text classification and spam filtering, due to its simplicity, speed, and ability to handle large datasets (Wickramasinghe & Kalutarage, 2021). Here are the parameters used for prediction in this project: 1) Data Collection and Preprocessing : Student grade data from the Computer Engineering and Informatics Education

Program from 2013 to 2015 were retrieved from the Academic Information System (SIA) of Padang State University. Data preprocessing involved selection, cleansing, and transformation for analysis using the Naïve Bayes algorithm; 2) Data Selection : Data obtained from the SIA database of Padang State University were selected based on their relevance to the research topic. Selected attributes for this study included transcripts of semester 1 to 4, as well as several related courses such as Programming Algorithms, Graphic Design, and others. A total of 294 student data records from the Computer Engineering Program for the years 2013, 2014, and 2015 will be used as training and testing data; 3) Data Cleansing and Transformation : This process aimed to ensure the consistency and validity of the data before further processing. It involved removing duplicate data, filling in empty attributes, and creating datasets based on selected attributes for use in data mining using RapidMiner; 4) Naïve Bayes Algorithm Classification (Selection of Training Data: The training data consisted of 114 student records from the years 2013-2015 from the Computer Engineering and Informatics Education Program at Padang State University that had been transformed; Calculation of Probability for Each Feature and Class: The initial step involved calculating the probability of each feature for each class by seeking the class probability of the field of interest or final project topic. Five classification classes were formed based on the training data; Calculation of Probability for Each Feature: Probability calculations were performed for continuous attributes such as grades from various courses)

Results and Discussions

The test results using RapidMiner tool indicate attribute distribution models for each field of interest.

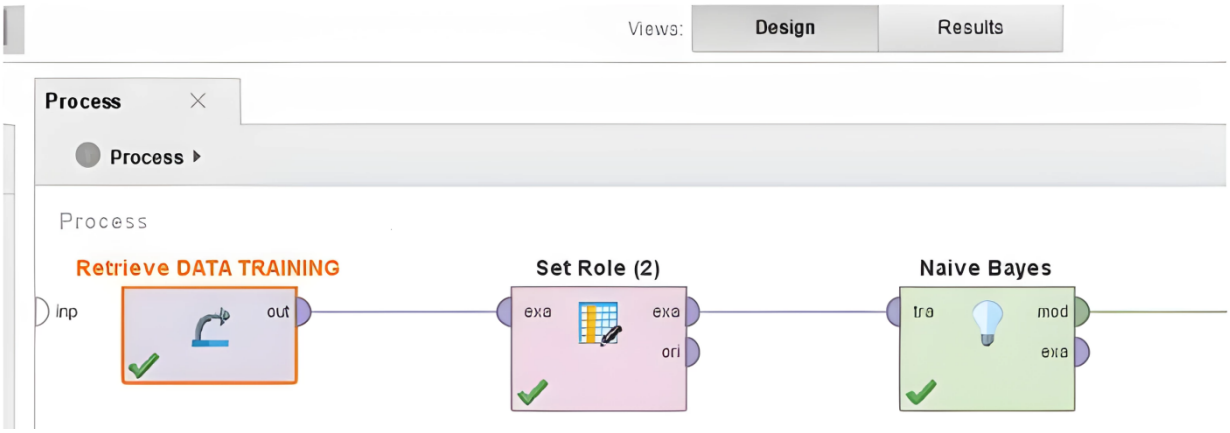

Figure 1. RapidMiner Design

Based on Figure 1, the obtained results are as follows:

SimpleDistribution

```
Distribution model for label attribute Minat

Class MM (0.167)
23 distributions

Class MP (0.237)
23 distributions

Class RPL (0.140)
23 distributions

Class TKJ (0.026)
23 distributions

Class WP (0.430)
23 distributions
```

Figure 2. Sample Distribution

From the Figure 2, the probabilities for each category of field of interest are as follows: "MM" 0.167, "MP" 0.237, "RPL" 0.140, "TKJ" 0.026, and "WP" 0.430. These findings are consistent with manual calculations of the probabilities for field of interest classes.

PerformanceVector

PerformanceVector:
accuracy: 52.63%
ConfusionMatrix:

True:	MM	MP	RPL	TKJ	WP
MM:	10	0	0	0	0
MP:	2	0	0	0	0
RPL:	3	0	0	0	0
TKJ:	0	0	0	0	0
WP:	4	0	0	0	0

kappa: 0.000
ConfusionMatrix:

True:	MM	MP	RPL	TKJ	WP
MM:	10	0	0	0	0
MP:	2	0	0	0	0
RPL:	3	0	0	0	0
TKJ:	0	0	0	0	0
WP:	4	0	0	0	0

Figure 3. Performance Vector

Based on the Confusion Matrix on the test data, accuracy, precision, and recall values are obtained. For example, for the category "MM", the accuracy is 52.63%, with precision and recall being 100% and 52.63% respectively. This means, out of 19 tested data items, 10 of them are correctly classified as "MM", 2 are classified as "MP", 3 are classified as "RPL", and 4 are classified as "WP". Based on the final testing, the results obtained are:

Table 1. **Result**

Category	Sample count	Accuracy	Precision	Recall
MM	19	52.63%	100%	52.63%
MP	27	62.96%	100%	62.96%
RPL	16	56.25%	100%	56.25%
TKJ	3	100%	100%	100%
WP	49	61.22%	100%	61.22%

Based on Table 1, the classification results using the Naïve Bayes Classifier algorithm are represented in the table, which includes accuracy, precision, and recall for each category. The classification model provides satisfactory results, with accuracy approaching the percentages listed in the table. The overall accuracy evaluation is conducted by calculating the proportion of data items classified correctly against the total tested data.

In this research, the identification of significant patterns or trends in the academic scores of students from the Informatics Engineering Education Program at Universitas Negeri Padang from 2013 to 2015 is crucial. This identification process aims to uncover relationships between the variables under study, such as grades in specific courses and students' fields of interest. Through this analysis, it is hoped that relevant patterns or trends can be revealed, which will then assist in addressing research questions regarding factors influencing students' field of interest concentrations and changes in their academic performance over time.

Hypotheses in this research objective are assumptions or predictions regarding the relationships between specific variables that are to be tested through data analysis. In the context of this research, hypotheses may focus on predictions about factors influencing students' fields of interest or final project topics in the Informatics Engineering Education Program at Universitas Negeri Padang.

The methods used in this research have advantages in gathering accurate and relevant data and are capable of analyzing the relationships between variables carefully. This supports the validity of the research findings by ensuring that the generated findings are trustworthy and statistically meaningful. However, limitations may arise in terms of generalizing the results due to sample limitations or potential biases in the data collection or analysis process. Therefore, while these methods can provide valuable insights, it is important to consider these limitations in interpreting the research results.

The practical implications of this research are highly relevant for educational practices and policymaking in the field of education. The research findings can provide valuable insights for policymakers in developing more effective and relevant curricula to meet students' needs. The recommendations generated can also aid in the development of training programs for educators to enhance the quality of teaching and learning. Thus, this research not only provides a better understanding of the researched topic but also makes a tangible contribution to improving the overall quality of education.

Furthermore, the findings of this research have significant implications for educational practices and policymaking, especially in curriculum development and educator training. By comparing these findings with previous research, we can gain a deeper understanding of the dynamics within the academic environment of the Informatics Engineering Education Program at Universitas Negeri Padang.

One notable aspect revealed by this study is the identification of significant patterns or trends in academic scores over a specific timeframe. These findings support earlier research in the field, indicating that certain factors such as students' fields of interest and performance in specific courses play pivotal roles in shaping their academic trajectory. However, our research contributes further by delving into the nuances of these relationships within the context of our institution, offering valuable insights for curriculum designers and educators.

The analysis conducted in this study utilized the Naïve Bayes Classifier algorithm, resulting in satisfactory classification results. While this approach has proven effective, it's essential to acknowledge its limitations. One such limitation is the reliance on a specific algorithm, which may not capture all intricacies of the data. Additionally, the generalization of results may be hindered by sample size constraints or biases inherent in the data collection process.

To strengthen the discussion, it's pertinent to provide the author's commentary on the findings. From the author's perspective, these results underscore the importance of a nuanced approach to curriculum development that takes into account individual student interests and performance patterns. Moreover, they highlight the need for ongoing professional development initiatives for educators to adapt teaching strategies effectively.

Moving forward, future research endeavors should aim to address the limitations identified in this study. Expanding the sample size and employing diverse analytical techniques could yield more robust findings. Furthermore, longitudinal studies tracking students' academic trajectories over extended periods would offer valuable insights into the long-term impact of various educational interventions.

In conclusion, while this research contributes significantly to our understanding of academic performance trends within the Informatics Engineering Education Program, it also underscores the complexity of the educational landscape. By critically examining the findings in comparison with existing literature and offering the author's perspective, this discussion provides a comprehensive evaluation of the research outcomes and sets the stage for further inquiry and practical application.

Conclusions

In conclusion, the successful application of the Naive Bayes Method in predicting final project topics within the project-based learning framework for Data Mining courses underscores its significance in enhancing the educational experience. The method's demonstrated accuracy, precision, and recall values validate its effectiveness in guiding students towards relevant and engaging project topics. Moreover, the integration with project-based learning (PjBL) has not only enriched students' learning experiences but also bolstered their motivation and involvement in the educational process.

This research holds considerable implications for the advancement of higher education. By providing insights into effective pedagogical strategies, it offers valuable guidance for policymakers in the development of adaptive curricula tailored to meet the evolving needs of students. Additionally, the findings contribute to the ongoing discourse on educational methodologies, paving the way for further research and exploration in the realms of project-based learning and data mining algorithms.

In essence, this study not only sheds light on the practical application of the Naive Bayes Method but also underscores its broader significance in the realm of educational innovation and improvement. As we move

forward, it is imperative to continue exploring novel approaches that enhance student engagement and foster meaningful learning experiences.

References

Adadi, A. (2021). A survey on data-efficient algorithms in big data era. *Journal of Big Data*, *8*(1), 24. https://doi.org/10.1186/s40537-021-00419-9

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 3. https://doi.org/10.1186/s41239-020-0177-7

Bošnjaković, N., & Đurđević Babić, I. (2023). Systematic Review on Educational Data Mining in Educational Gamification. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-023-09686-2

Charitopoulos, A., Rangoussi, M., & Koulouriotis, D. (2020). On the Use of Soft Computing Methods in Educational Data Mining and Learning Analytics Research: a Review of Years 2010–2018. *International Journal of Artificial Intelligence in Education*, *30*(3), 371–430. https://doi.org/10.1007/s40593-020-00200-8

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361. https://doi.org/https://doi.org/10.1016/j.knosys.2019.105361

Creswell, J. W., & Creswell, J. D. (2018). Mixed Methods Procedures. In *Research Defign: Qualitative, Quantitative, and Mixed M ethods Approaches*.

Dixit, A., Mani, A., & Bansal, R. (2020). Feature selection for text and image data using differential evolution with SVM and Naïve Bayes classifiers. *Engineering Journal*, *24*(5), 161–172. https://doi.org/10.4186/ej.2020.24.5.161

Egwim, C. N., Alaka, H., Toriola-Coker, L. O., Balogun, H., & Sunmola, F. (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications*, *6*, 100166. https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100166

Ikegwu, A. C., Nweke, H. F., & Anikwe, C. V. (2023). Recent trends in computational intelligence for educational big data analysis. *Iran Journal of Computer Science*. https://doi.org/10.1007/s42044-023-00158-5

Kukkar, A., Mohana, R., Sharma, A., & Nayyar, A. (2023). Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms. *Education and Information Technologies*, *28*(8), 9655–9684. https://doi.org/10.1007/s10639-022-11573-9

Kurniawan, Y. I., Cahyono, T., Nofiyati, Maryanto, E., Fadli, A., & Indraswari, N. R. (2020). Preprocessing Using Correlation Based Features Selection on Naive Bayes Classification. *IOP Conference Series: Materials Science and Engineering*, *982*(1). https://doi.org/10.1088/1757-899X/982/1/012012

Shafiq, M., Alghamedy, F., Jamal, N., Kamal, T., Daradkeh, & Shabaz, D. M. (2023). Scientific programming using optimized machine learning techniques for software fault prediction to improve software quality. *IET Software*, *17*, n/a-n/a. https://doi.org/10.1049/sfw2.12091

Siva Subramanian, R., Prabha, D., Aswini, J., & Maheswari, B. (2022). *Evaluation of Different Variable Selection Approaches with Naive Bayes to Improve the Customer Behavior Prediction BT - Inventive Computation and Information Technologies* (S. Smys, V. E. Balas, & R. Palanisamy (eds.); pp. 181–201). Springer Nature Singapore.

Tran, P. H., Ahmadi Nadi, A., Nguyen, T. H., Tran, K. D., & Tran, K. P. (2022). *Application of Machine Learning in Statistical Process Control Charts: A Survey and Perspective BT - Control Charts and Machine Learning for Anomaly Detection in Manufacturing* (K. P. Tran (ed.); pp. 7–42). Springer International Publishing. https://doi.org/10.1007/978-3-030-83819-5_2

Wahyudi, W., Nurhayati, N., & Saputri, D. F. (2022). Effectiveness of Problem Solving-based Optics Module in Improving Higher Order Thinking Skills of Prospective Physics Teachers. *Jurnal Penelitian Pendidikan IPA*, *8*(4), 2285–2293. https://doi.org/10.29303/jppipa.v8i4.1860

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Zotou, M., Tambouris, E., & Tarabanis, K. (2020). Data-driven problem based learning: enhancing problem based learning with learning analytics. *Educational Technology Research and Development*, *68*(6), 3393–3424. https://doi.org/10.1007/s11423-020-09828-8