Mathematical Biosciences and Engineering

MBE, 19(8): 8621–8647.
DOI: 10.3934/mbe.2022400
Received: 23 March 2022
Revised: 26 May 2022
Accepted: 31 May 2022
Published: 14 June 2022

http://www.aimspress.com/journal/MBE

Research article

An application of data mining algorithms for predicting factors affecting Big Data Analysis adoption readiness in SMEs

Nguyen Thi Giang[1,2] and Shu-Yi Liaw[3,*]

[1] Department of Tropical Agriculture and International Cooperation, National Pingtung University of Science and Technology, Taiwan
[2] Faculty of Economics and Rural Development, Thai Nguyen University of Agriculture and Forestry, Vietnam
[3] Director of Computer Centre, Department of Business Administration, National Pingtung University of Science and Technology, Taiwan

* **Correspondence:** Email: syliaw@mail.npust.edu.tw; Tel: +88687703202 (ext. 6140).

Abstract: The adoption of Big Data Analysis (BDA) has become popular among firms since it creates evidence for decision-making by managers. However, the adoption of BDA continues to be poor among small and medium enterprises (SMEs). Therefore, this study adopted the Technology-Organization-Environment (TOE) framework to identify the drivers of readiness to adopt BDA among SMEs. Chi-square automatic interaction detection (CHAID), Bayesian network, neural network, and C5.0 algorithms of data mining were utilized to analyze data collected from 240 Vietnamese managers of SMEs. The evaluation model identified the C5.0 algorithm as the best model, with accurate results for the prediction of factors influencing the readiness to adopt BDA among SMEs. The findings revealed management support, data quality, firm size, data security and cost to be the fundamental factors influencing BDA adoption readiness. Moreover, the results identified the service sector as having a higher level of readiness toward the adoption of BDA compared to the manufacturing sector. The findings are imperative for the enhancement of the decision-making process and advancement of comprehension of the determinants of BDA adoption among SMEs by researchers, managers, providers and policymakers.

Keywords: Bayesian networks; big data analysis; CHAID; C5.0; data mining; neural network; SMEs

1. Introduction

The vigorous advancement of internet technology continues to generate large volumes of data through several sources, such as media, the cloud, the Web, the Internet of Things and databases [1]. The aggregation of these sources is referred to as big data, and companies are looking to process and analyze these huge data sets to extract benefits [2]. As a result, numerous studies have shown the benefits of BDA in organizations. Specifically, the use of BDA enhances the prediction of future product development trends, which improves the decision-making process [2–5], and enhances supply chain systems [6]. BDA prediction is also paramount in the promotion of firm performance [7–9], improvement of marketing efficiency [5,10] and prediction of market trends [5,11]. The momentousness of BDA adoption culminates with the development of a sustainable dynamic economic system that takes advantage of current contextual demands [12]. Evidence shows that firms that succeed in implementing BDA primarily graduate into major cross-national corporations. Examples of such corporations include Google, Apple, Twitter, Uber, Walmart, Amazon, IBM Watson, Rolls-Royce, Toyota and others [13]. Despite the benefits of BDA for firms and economic performance, numerous companies still encounter an assortment of barriers that inhibit the adoption of BDA, especially by SMEs [14–16]. For most developing economies, SMEs are pivotal in economic development and validation of BDA implementation. However, Coleman et al. [17] indicated that SMEs are still slow in implementing BDA, as they are faced with several barriers in the application of big data [17,18]. Del Vecchio et al. [19] pointed out the challenges and benefits of big data for SMEs. Noonpakdee et al. [20] presented barriers when Thailand SMEs adopted big data. Similarly, Chuah and Thurusamry [21] mentioned the challenges of SMEs in Malaysia using BDA. In addition, Mangla et al. [22] demonstrated the performance of SMEs' adoptions of BDA in India. Park and Kim [23] and Maroufkhani et al. [9] identified drivers of big data adoption among Korean and Iranian SMEs. However, the majority of these studies concentrate on the advantages and efficiency of BDA adoption, as well as the challenges that SMEs face when performing BDA. Previous research examining the factors influencing the use of BDA by SMEs is still scarce. With limited studies on BDA application by SMEs, such as in Vietnam, it becomes very difficult for SMEs to adopt BDA. The Technology-Organization-Environment (TOE) framework is composed of technology, organization and environment pillars [24]. It is considered to be the most comprehensive and flexible approach for examining company decisions on the adoption and implementation of information technology-based innovations [25]. Therefore, this study applies the TOE framework and four data mining algorithms (CHAID, Bayesian networks, neural networks and C5.0) to identify the predictors of readiness to adopt BDA by SMEs. The study was guided by the following objectives:

1) To identify the best model for the predicting factors' influences on the readiness to adopt BDA among SMEs and

2) To predict the key factors that affect the readiness to adopt BDA in SMEs.

The findings will be useful for managers, policymakers and providers to understand the influences of BDA adoption readiness. Managers can, therefore, build competitive strategies to enhance company performance through the use of BDA. Additionally, the study proves new techniques that can be used to predict the factors influencing enterprise readiness to adopt BDA.

2. Literature review

2.1. Big data analytics adoption among SMEs

Big data includes both structured and unstructured large volumes of data, and their analysis requires specific processing. The key features of the big data process are categorized into 3 Vs: (i) volume, (ii) velocity and (iii) variety. In this case, volume depicts the amount of information in the dataset, while velocity refers to the rate at which data are created. Variety indicates the different forms of data that are created. Zhong et al. [26] added two more Vs, verification and value, to characterize big data as a "5Vs" data source. In this case, verification concerns bad data that need to be verified, whereas value addresses the economic and social costs of application. On the other hand, Saggi and Jain [27] classified big data features into volume, velocity, variety, valence, veracity, variability and value to produce the "7Vs" classification. The valence is related to the complexity of the data, and veracity reflects accuracy within the dataset, while the inconsistencies in all data are mostly responsible for variability.

Ideally, BDA involves two components, big data and business analytics [5]. The former provides the foundation for informational and technological analysis for business activities, whereas the latter provides valuable insights necessary for the improvement of the decision process in the business unit. This has a multidisciplinary benefit that promotes firm business performance [28]. For example, big data has been adopted in the manufacturing sector [9], the health care sector [29], the service sector [26] and the hospitality industry [30]. Dubey et al. [31] argued that BDA presents unequivocal and fundamental impact effects on the swiftness of supply chains and competitive advantage. Previous studies that presented benefits, challenges and performance applied big data in SMEs [17,19,20,22,32,33]. For example, Park and Kim [23] used the analytic hierarchy process and regression analysis and found that benefits received, technological abilities, financial abilities and data quality are the major factors predicting the intention to apply big data among Korean companies. Mangla et al. [22] applied structural equation modeling (SEM) to show that BDA increased project performance in Indian SMEs. Similarly, Maroufkhani et al. [9] and Lutfi et al. [34] also used partial least squares structural equation modeling (PLS-SEM) to identify the elements impacting the intentions of Iranian SMEs and Jordanian SMEs to use BDA. In addition, Sun et al. [35], Maroufkhani et al. [36] and Baig et al. [37] used a review of related articles to figure out drivers of an organization's inclination toward the utilization of big data for businesses purposes. Clearly, most of the previous studies on factors affecting the intentions of BDA adoption used latent variables. This leads to the limitation of independent factors [38]. The observed variables (e.g., demographic variables, sector, firm size) are rarely included in the research model. This research works to bridge this gap.

2.2. Theoretical background

2.2.1. Technology-Organization-Environment framework

The TOE framework is useful in revealing the drivers of decisions to embrace new information technology [24]. It is a threefold framework consisting of technology, organization and environment. The technology pillar defines factors associated with tools, software, IT infrastructure, etc. which affect decisions to apply big data by individuals and/or organizations. The organization pillar defines the

capacity of a firm to acquire competence in the employment of multiple resources required for the operation of information systems in firms. The environmental pillar consists of multiple industrial features, e.g., competitors and vendor support, directly or indirectly affecting the operations of enterprises. The TOE framework is considered to be flexible and is widely used in technology application studies amongst companies [39]. Some previous studies on BDA adoption have applied the TOE framework. Sun et al. [35] and Baig et al. [37] laid out a synopsis of the determinants of big data adoption using the TOE framework. Park et al. [40] and Park and Kim [23] applied the TOE framework to ascertain the drivers of big data adoption among Korean companies. Similarly, Lai et al. [41], used the TOE framework to identify the determinants of BDA adoption by Chinese firms. Maroufkhani et al. [9] applied the framework to find out the determinants of BDA application among SMEs in Iranian. However, previous studies evaluating factors affecting BDA mostly refer to latent variables without considering observed variables. Therefore, the present study extends the TOE framework to understand the drivers of BDA adoption. The research model of this study is shown in Figure 1.

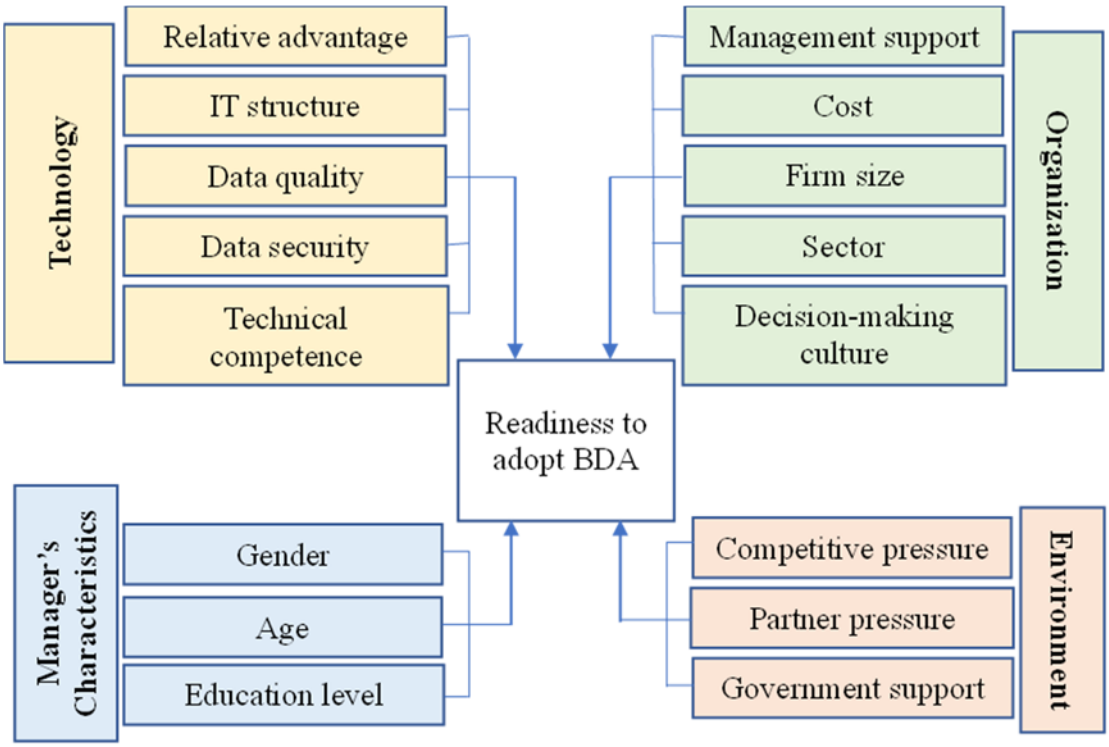

Figure 1. Research model.

2.2.2. Technology dimension

The technology pillar involves intra- and inter-organizational drivers that influence company decisions to embrace new information technology [42]. In this dimension, the first factor mentioned is the relative advantage, which outlines the level to which the new proposed technology provides greater benefit for firms [43]. According to Ghobakhloo et al. [44], SMEs are only willing to embrace new technology if the said advantages outweigh the performance of existing technology. IT infrastructure is salient for organizational competitiveness [30], reflecting a firm's ability to operationalize

information systems. However, SMEs often lack IT resources, undercutting their abilities for data collection and analysis [19]. According to Wang and Wang [32], the lack of IT specialists is a major drawback for most SMEs in attaining flexibility in IT infrastructure usage. Data quality is an important factor leading to the success of enterprises' BDA adoption.

Big data stockpiles could be structured, semi-structured or unstructured. Organizations must choose specific software to ensure the quality of the data as well as the efficiency of BDA [14]. Park and Kim [23] mentioned that data quality has a great influence on big data adoption decisions among Korean firms. The security issue is critical for firms' decisions to adopt BDA. Third parties are privy to personal and company information, thus exposing individuals and companies to cybercrime [45]. Therefore, data security is a key factor affecting the decisions of enterprises to adopt BDA [35]. Technical competence refers to expertise, which is a prerequisite for analyzing big data by employees. Yadegaridehkordi et al. [30] indicated that enough knowledge for staff to analyze information technology is an important factor affecting the application of innovation in organizations. Alharthi et al. [14] concluded that staff lack of BDA skills is a barrier when companies adopt BDA.

2.2.3. Organization dimension

The organizational dimension represents different organizational conditions that affect readiness toward the adoption of BDA. The first element is management support, which is critically vital in the adoption of an innovation [46]. If managers realize the benefits of BDA adoption, they can allocate the resources needed for implementation. By contrast, if management does not see the profits of BDA adoption, they will oppose the application of that data [47]. Second, the adoption of BDA is attached to a cost factor to maintain and develop the application of big data [35]. In this regard, company development-related costs are usually funded through support from the financial institution. Such support tends to be limited for SMEs compared to larger firms, thereby undermining the adoption of BDA by small companies [17]. Hence, firm size is considered an essential driver for the adoption of technological innovations [24]. The type of industry is another driver believed to influence the intentions to apply new technology in enterprises. Gangwar [48] pointed out that there was a significant difference between the manufacturing and the service sectors regarding BDA. Finally, decision-making culture is another factor that influences the adoption of BDA. More often than not, organizations that apply an evidence-based decision-making culture embrace big data analytics to develop evidence that enhances managers' competence for strategic decision-making, thereby improving enterprise profitability [35].

2.2.4. Environment dimension

Environmental factors include external factors that the organization may encounter [49]. Factors such as competition pressure, partner pressure and government support are perceived as external drivers of big data adoption by SMEs [23]. Competitive pressure outlines the extent to which competitors affect organizational decisions towards the adoption of new technologies [24]. The role of competition pressure is widely acknowledged in the literature on IT adoption [50,51]. Zhu et al. [52] revealed the importance of the pressure from trading partners in influencing company decisions to adopt and utilize new information technology. In addition, the government also plays a fundamental role in influencing the adoption of information technology. If the government exudes a strong political

will and ensures a good institutionally enabling environment for the enrollment of big data technology, firms are often encouraged to develop internal policies for the adoption and implementation of BDA. Such a positive relationship has been confirmed by numerous studies [35,41]. Government support and policy include the provision of public data, fostering of experts, protecting intellectual property and regulation for privacy and security that affect the use of big data by firms [53].

2.2.5. Manager's characteristics dimension

Rojas-Méndez et al. [54] demonstrated that demographic variables (gender, age, education level) are important factors for predicting people's willingness to adopt the technology. In this regard, the manager's level of education is the most important demographic characteristic affecting the application of technology [54]. Parasuraman and Colby [55] pointed out that there is a need for studies focusing on factors such as age, education level, occupation and demographic characteristics to assess the readiness to use the new technology of each person. For this reason, the manager's characteristics dimension is included to predict the determinants of big data adoption by SMEs.

2.3. Data mining

Data mining includes many different algorithms used mainly for classification purposes. CHAID analysis is an algorithm that develops a predictive model that merges predictors that best explain the response variable [56]. A Bayesian network is a probability-based graphical model that represents expertise about an uncertain domain, where individual nodes correspond to some random variable, and each edge represents the conditional probability for the corresponding random variables [56,57]. Neural networks are a set of connected input/output units where each connection has a distinct weight associated with each other [56]. One of the most often used decision tree inducers is the C5.0 model, which divides the sample according to the field that delivers the most information gained at each level.

The four algorithms have some differences. Neural networks are widely used because of their ability to produce results quickly, although their capacity for problem-solving is limited. The CHAID model uses simple predictions based on the frequency distribution of potential problems. The C5.0 model is considered an algorithm with outstanding performance and high accuracy [58].

The data mining technique is applied in research to collect data from questionnaires and predict factors affecting the research problem. For instance, Cortez and Silva [59] collected data from 788 students in a public school in Portugal by questionnaire. The questionnaire included 37 items that mentioned demographics, social and school information. Four algorithms, consisting of decision trees, random tree, neural networks and support vector machines, were used to predict students' mathematics and Portuguese grades in this study. Yukselturk et al. [60] predicted dropout students through four algorithms: k-nearest neighbor, decision tree, naive Bayes and neural network. In that study, data was collected from 189 students in Turkey. The questionnaire included ten variables to predict students who drop out of courses. Applying the data mining technique, the researcher can easily discover unexpected factors [61]. However, studies using the data mining technique to predict factors affecting the adoption of BDA have still not been found.

3. Methodology

3.1. Data collection and sample

The questionnaire was literature-based and collected comments from professionals and managers of SMEs. The questionnaire was partitioned into three sections. Section A used thirty-five items collecting data on determinants of readiness to implement big data among SMEs. Section B consisted of nine items assessing the readiness to apply BDA. The first two sections used a seven-point agreement Likert-scale, ranging from 1 for "Strongly Disagree" to 7 for "Strongly Agree." Section C collected data on the respondents' socio-economic characteristics.

The subjects of this study are SMEs involved in manufacturing and service provision. The manufacturing and service sectors are two areas that have important roles in the economies of each country [62]. Manufacturing refers to the activities of people using tools and machines to convert raw materials into finished products, transport them to suppliers and recycle used products [26,63]. Services include areas such as retail, finance, tourism, health, accommodation services, restaurants, etc., whereby the service sector provides services to consumers. The questionnaire was emailed to Vietnamese managers of SMEs that met the eligibility criterion of the study. A total sample of 240 managers of manufacturing and service provider companies participated in the study. The data were collected during the period from September to December 2020.

Table 1 shows the respondents' demographic analysis. The gender proportion showed that the majority of respondents were males (72.5%), followed by females (27.5%). Age distribution was such that the majority of the respondents were aged 30 to 45 (57.9%), with those aged ≥ 46 accounting for 29.2%, and those aged < 30 accounted for 12.9%. The descriptive statistics revealed that 46.7% of managers hold bachelor's degrees, 39.2% hold post-graduate degrees, and only 14.2% have college or vocational training. Firm size showed that the majority of participants were small enterprises (82.5%), and medium enterprises accounted for 17.5%. Among these firms, 50.8% were manufacturing firms, and 49.2% were service firms.

Table 1. Demographics of respondents (n = 240).

Variable	Type	Frequency	Percentage (%)
Gender	Male	174	72.5
	Female	66	27.5
Age	< 30	31	12.9
	30–45	139	57.9
	≥ 46	70	29.2
Education level	College education	34	14.2
	Bachelor's degree	112	46.7
	Master's degree or above	94	39.2
Role of respondent	Chief Executive Officer	85	35.4
	Executive management	91	37.9
	IT management	64	26.7
Sector	Manufacturing	122	50.8
	Service	118	49.2
Firm size	Small enterprise	198	82.5
	Medium enterprise	42	17.5

3.2. Reliability, validity analysis and coding of the readiness to apply big data analysis

In this study, each variable is measured by at least three items based on references. To be more specific, the variables are relative advantage (four items) [51], IT infrastructure (three items) [20], data quality (three items) [41], data security (three items) [64], technical competence (four items) [65], management support (three items) [66], cost (three items) [51], decision-making culture (three items) [35], competitive pressure (three items) [67], partner pressure (three items) [67], government support (three items) [26] and readiness to apply BDA in SMEs (nine items) [37,55,68].

To assess the reliability and validity of latent variables, Cronbach's α value, composite reliability (CR), average variance extracted (AVE) of all constructs and factor loadings of items are shown in Table A1. A preliminary dataset analysis of External Factor Analysis (EFA) was carried out. The KMO (Kaiser-Meyer-Olkin) value was 0.814, being greater than the critical value (0.7) [69], and the Bartlett sphericity test's significant value was $p = 0.000$, indicating that factor analysis is suitable for the original dataset. Cronbach's α value was computed to assess the reliability of the questionnaire. The reliability test indicated that the value of Cronbach's α for the latent variables ranged between 0.626 and 0.867. According to Hair et al. [70], if the Cronbach's α value is greater than 0.700 (0.600 acceptable), the questionnaire has good internal consistency. Therefore, the questionnaire for this study was found to be consistent and reliable.

Table 2. The description of the independent variables.

No.	Variable	Data type	Description
Technology dimension			
1	Relative advantage	Continuous	Mean value
2	IT infrastructure	Continuous	Mean value
3	Data quality	Continuous	Mean value
4	Data security	Continuous	Mean value
5	Technical competence	Continuous	Mean value
Organization dimension			
6	Management support	Continuous	Mean value
7	Cost	Continuous	Mean value
8	Firm size	Nominal	1 = "Small", 2 = "Medium"
9	Sector	Nominal	1 = "Manufacturing", 2 = "Service"
10	Decision-making culture	Continuous	Mean value
Environment dimension			
11	Competitive pressure	Continuous	Mean value
12	Partner pressure	Continuous	Mean value
13	Government support	Continuous	Mean value
Manager's characteristics dimension			
14	Gender	Nominal	1 = "Male", 2 = "Female"
15	Age	Nominal	1 = "< 30", 2 = "30–45", 3 = "≥ 46"
16	Education level	Nominal	1 = "High school, College/Vocational education", 2 = "Bachelor's degree", 3 = "Master's degree, or above"

All factor loadings (from 0.520 to 0.865) were higher than the acceptable limit (0.5) [69]. The CR of all constructs indicated good internal consistency, being higher than 0.7 [71]. All constructs, except for data quality (0.457) and management support (0.471), had AVE values higher than 0.5, indicating good convergent validity. Taking into consideration the Fornell and Larcker [72] proposal that an AVE value equal to 0.4 can be acceptable if the CR value is greater than 0.6, the data quality and management support variables were accepted in this study because they had a CR value high of 0.7. This proves that all latent variables in this study have acceptable convergent values.

To predict the factors' influences on BDA adoption readiness, the dependent variable (readiness to apply big data in SMEs) was divided into two options based on an average of nine items that identify the readiness to apply BDA among SMEs. The first option was coded "1 = Low readiness," with the mean values of the nine items < 6.0, and "2 = High readiness" was used with the mean values of the nine items ≥ 6.0. Table 2 and Table 3 present the sixteen independent (input) variables and the dependent (target) variable.

Table 3. The description of the dependent variable.

Category	Frequency	Percentage (%)
1 (Low readiness)	119	49.6
2 (High readiness)	121	50.4
Total	*240*	*100.0*

3.3. Data analysis

This study used four data mining algorithms that were run through the Statistical Package for Social Sciences (SPSS) 18 software (IBM, Armonk, NY, USA). The algorithms used for the prediction of factors' influences on the adoption readiness of BDA include CHAID, Bayesian networks, neural networks and C5.0. These algorithms are commonly applied in studies that analyze data collected from questionnaires.

CHAID algorithm

CHAID is one of the pioneer algorithms that partition data into multiple subgroups [73]. However, this method does not allow for data pruning. CHAID applies the chi-square independence test to identify the splitting rule for each node. This test performs an automatic split categorization of independent categorical variables from continuous variables. Super-classes are then produced through the merging of the input variables based on statistical analogy, maintaining them if they are statistically dissimilar. A comparative analysis between the super-classes and the target variable is done to assess dependency using the chi-square independence test. The super-class that shows the highest significance is then selected as the splitting criteria for the node.

Bayesian networks algorithm

The Bayesian network is popular, being used in multiple research fields [74]. This method combines qualitative and quantitative variables. A Bayesian network is a directed graph with an additional set of probability distributions. Here, the graph represents the qualitative aspect, whereas the probability distributions represent the quantitative part. In the graph, the nodes denote dubious factors, while the arcs address the presence of a causal connection between two factors. Bayesian networks are very effective in predictive studies. The structure makes inferences from Bayesian networks robust, reduces the differences of estimated parameters and is also robust against overfitting.

Neural network algorithm

Neural networks are modeled from brain functionality. They use numerous connected receptor units that accept messages from other units, processing them and conveying the new message to other units. However, the output of the neural network is difficult to retrace; hence, interpretation becomes hard. These disadvantages are overridden by the complexity and flexibility of the algorithm, transforming it into a robust and comprehensive discriminator that is applicable to resolve varied problems compared to other methods [56].

C5.0 algorithm

The C5.0 algorithm evolved from the C4.5 algorithm as formulated by Ross Quinlan [75]. The algorithm has the capacity to segment data into multiple subgroups. The C5.0 possesses pruning ability, selecting splitting rules through an impurity measure [56]. The pros of the C5.0 algorithm include its robustness in handling missing data points and several input columns. In addition, the method requires shorter training sessions for estimates and uses normal enhancement techniques to improve the accuracy of the classification function.

3.4. Measures for performance evaluation

This study sought to categorize response variables into two options (Low readiness and High readiness); then, a partition node was inserted to segregate the data into training (70%) and testing (30%) sets. The performance of models was assessed through the confusion matrix (Table 4). Next, the performance of models was analyzed using the attributes of accuracy, precision, recall, specificity, F-measure and area under the receiver operating characteristic (ROC) curve (AUC) and k-fold cross-validation.

In Table 4, true positive and true negative present the number of correct positive and correct negative samples predicted by the model. False positive and false negative stand for the number of wrong positive and wrong negative samples [76,77].

Table 4. Form of confusion matrix.

Confusion Matrix of Readiness Low or High		Predicted value	
		Low readiness	High readiness
Observed value	Low readiness	True Negative (TN)	False Positive (FP)
	High readiness	False Negative (FN)	True Positive (TP)

Accuracy is judging the overall correct rate, that is, that the actual category is consistent with the predicted category [76,77].

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \tag{1}$$

Precision is judging how much of the recall is true, that is, how much of the actual truth is accurately predicted to be true [76].

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

Recall is the proportion of true positives to the total number of true positives and false negatives [76,77].

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$1 - \text{Recall} = \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{4}$$

Specificity is the correct rate of judgment that is true, that is, the ratio of true to true among predictions [76].

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{5}$$

F-measure: The harmonic mean of the precision and precision performance measurements is used to calculate the precision recovery curve. A high F-measurement result suggests that the categorization quality is excellent [76].

$$\text{F-measure} = 2 \times \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \tag{6}$$

AUC: The ROC is a two-dimensional diagram of the false positive rate (FPR) on the horizontal axis versus the true positive rate (TPR) on the vertical axis. Based on Eqs (3) and (4), the TPR and FPR values of the cut-off points between 0 and 1 are calculated, and then the diagram is plotted by joining these data points. The area under the curve (AUC) is an appropriate measure if its value always varies between 0.5 and 1. The AUC is the standard to evaluate the model performance [78]. More specifically, the model performance is evaluated as acceptable ($0.7 <= \text{AUC} < 0.8$), good ($0.8 < \text{AUC} < 0.9$) or outstanding ($\text{AUC} \geq 0.9$) discrimination [79].

k-fold cross-validation: In a comparative analysis of various forecast models, the total collection data is commonly divided into training and testing subsets, and thoroughly expecting models are analyzed based on their precision in the test data set. By dividing the information into designing and testing datasets, a decision of doing a single split or multiple splits can be made, which is regularly called k-fold cross-validation. To estimate the performances of classifiers, a stratified 10-fold cross-validation approach is used. Empirical studies showed that 10 folds seem to be an optimal number [80]. In this study, each fold of data included 24 cases (240 cases/10 = 24 cases) and was used once to test the performance of the classifier.

To be clearer, the research process of this study is shown in Figure 2.

Data were collected from 240 managers of Vietnamese SMEs. A total of sixteen input variables (eleven latent variables and five observed variables) were analyzed through the data mining technique. The performances of prediction models were evaluated through the four classification models. The best performance was revealed by the C5.0 model, predicting readiness to apply big data with more accuracy. Therefore, C5.0 was employed to predict the five observed variables' (firm size, sector, gender, age and education level) impacts on the readiness to apply BDA. Finally, the C5.0 procedure was illustrated as a decision tree.

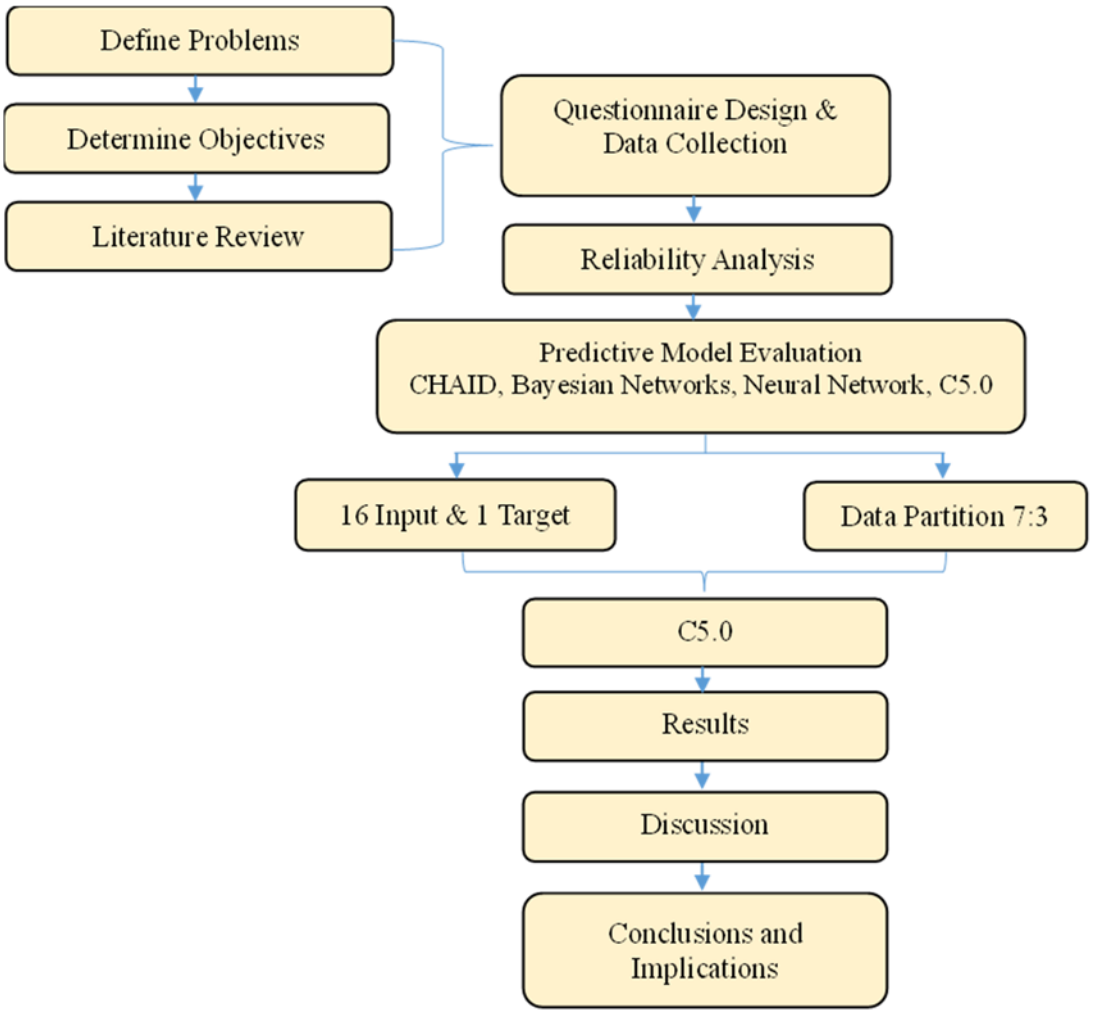

Figure 2. The study research process.

4. Results

4.1. Prediction accuracy of models

As shown in Table 5, the correctness values of the predictions of CHAID, Bayesian networks, neural network and C5.0 for training data were 83.32, 82.93, 81.10 and 87.20%, respectively. The results of the correct predictions on the testing data were 68.42, 85.53, 71.05 and 89.47%, respectively. Hence, these models have high prediction accuracy.

Moreover, the training data showed an AUC value range of 0.861 to 0.941, while the test set ranged from 0.747 to 0.939. Hence, the models were considered good in discriminating the predictors [79]. The stream of four models is shown in Figure 3.

The ROC curve is also used for the evaluation of the classification algorithms. The ROC curve visualizes the false positive rate against the true positive rate. The false positive rate result will change according to the classification threshold value, and the best classification result model can be selected according to the area under the ROC curve. A larger area means the model has a better classification effect. In Figure 4, the results show that on training data, Bayesian networks are the best model, and

for testing data, C5.0 is the best model.

Table 5. Evaluating the measurement results of four models.

Model type	Title	Training		Testing		AUC	
						Training	Testing
CHAID	Correct	135	83.32%	52	68.42%	0.891	0.747
	Wrong	29	17.68%	24	31.58%		
	Total	164		76			
Bayesian networks	Correct	136	82.93%	65	85.53%	0.941	0.910
	Wrong	28	17.07%	11	14.47%		
	Total	164		76			
Neural network	Correct	133	81.10%	54	71.05%	0.861	0.815
	Wrong	31	18.90%	22	28.95%		
	Total	164		76			
C5.0	Correct	143	87.20%	68	89.47%	0.893	0.939
	Wrong	21	12.80%	8	10.53%		
	Total	164		76			

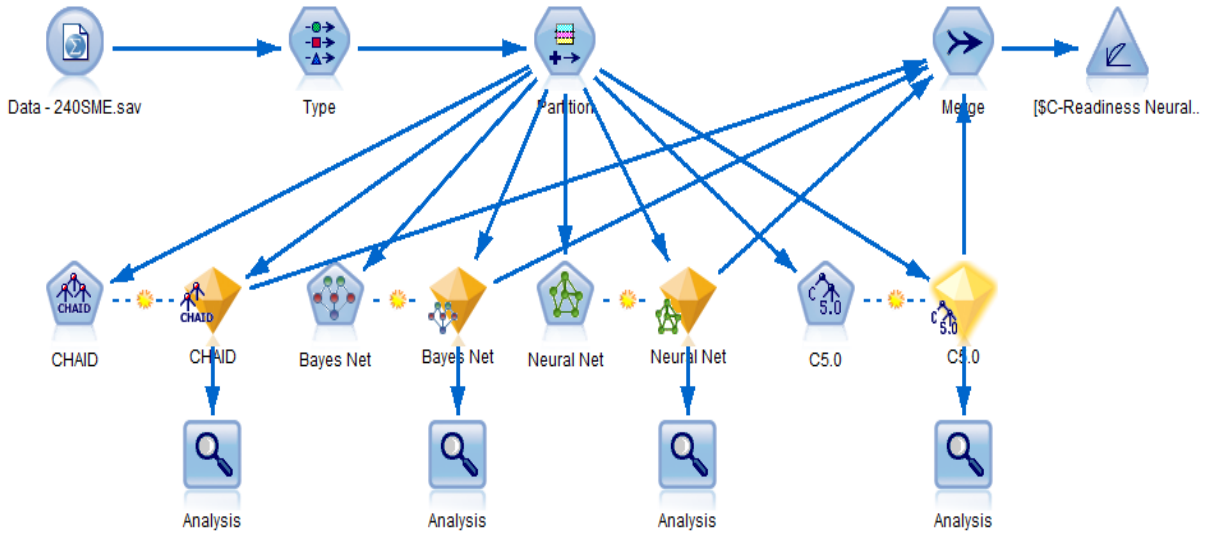

Figure 3. Stream of the four models with sixteen input variables.

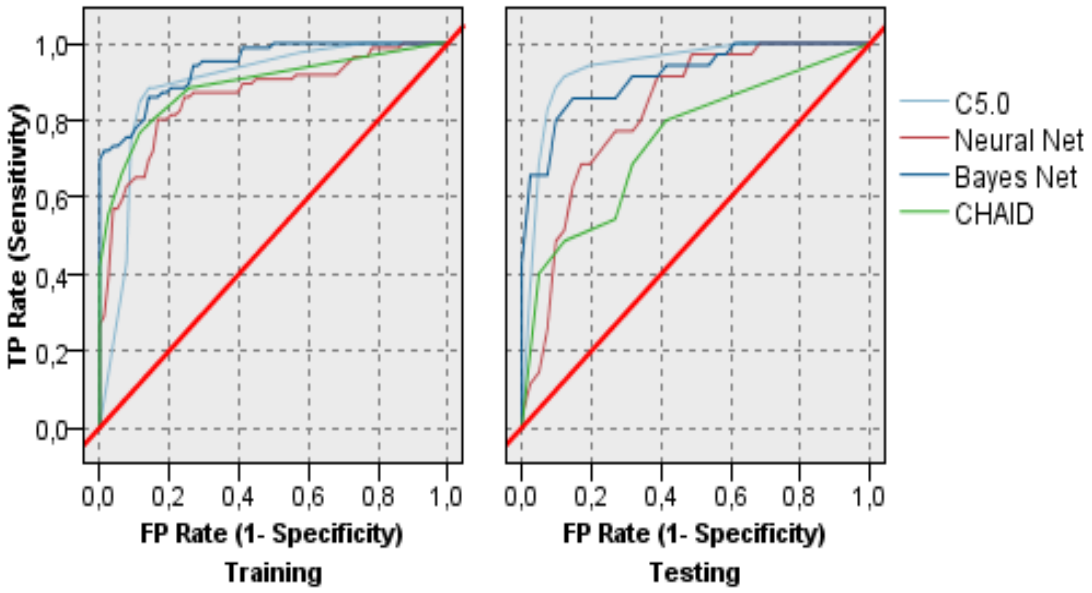

Figure 4. Graph of the ROC values of the four models.

The coincidence matrix and the evaluation results of the four models are shown in Table 6. It is clear that the resulting values in all four models for accuracy, precision, recall, specificity and F-measure were higher than 0.7, excepting accuracy, precision, recall, specificity and F-measure values on the testing data of the CHAID model and precision and specificity values on the testing data of the neural network model, which were approximately 0.7. This proves that the four models used in this study have good classification quality. Specifically, C5.0 is the model with the highest performance evaluation, followed by Bayesian networks, neural networks, and (the lowest) the CHAID model.

Similarly, the results of the 10-fold cross-validation for the four models are shown in Table A2, which indicated that C5.0 is the model with the highest average accuracy among the four selected tested models. Accordingly, the accuracy average for the 10-fold cross-validation of the C5.0 model is 0.885, followed by Bayesian networks (0.833) and neural networks (0.772), making the CHAID model (0.679) the lowest. This can be explained because C5.0 has higher memory performance than other algorithms and then can generate more precise rules. CHAID, an algorithm, applies the chi-square independence test that is suitable for categorical data. However, the input variables of this study are mostly continuous variables. To improve the precision of the model, this algorithm must perform by grouping data into categories. That is the reason why CHAID performs with the least precise predictions.

Table 6. Coincidence matrix and the evaluation results of the four models.

Model type	Partition	Title	Coincidence matrix		Accuracy	Precision	Recall	Specificity	F-measure
			Low readiness	High readiness					
CHAID		Low readiness	69	9					
	Training	High readiness	20	66	0.8232	0.8800	0.7674	0.8846	0.8199
		Total	*89*	*75*					
		Low readiness	28	13					
	Testing	High readiness	11	24	0.6842	0.6486	0.6857	0.6829	0.6667
		Total	*39*	*37*					
Bayesian networks		Low readiness	70	8					
	Training	High readiness	20	66	0.8293	0.8919	0.7674	0.8974	0.8250
		Total	*90*	*74*					
		Low readiness	37	4					
	Testing	High readiness	7	28	0.8553	0.8750	0.8000	0.9024	0.8358
		Total	*44*	*32*					
Neural network		Low readiness	65	13					
	Training	High readiness	18	68	0.8110	0.8395	0.7907	0.8333	0.8144
		Total	*83*	*81*					
		Low readiness	27	14					
	Testing	High readiness	8	27	0.7105	0.6585	0.7714	0.6585	0.7105
		Total	*35*	*41*					
C5.0		Low readiness	67	11					
	Training	High readiness	10	76	0.8720	0.8736	0.8837	0.8590	0.8786
		Total	*77*	*87*					
		Low readiness	36	5					
	Testing	High readiness	3	32	0.8947	0.8649	0.9143	0.8780	0.8889
		Total	*39*	*37*					

4.2. Predictor importance of the input variables

Predictor importance is a sensitivity analysis technique. It is used to identify the more important variables and/or omit the least important variables in the forecasting model [76].

The important drivers of readiness to adopt BDA are presented in Table 7. In CHAID and Bayesian networks, the most important variable was the management support variable. Conversely, in the neural network, the cost variable was the most critical. The most critical variable in the C5.0 model was data quality. The predictors of four algorithms rank the predictors from most important to least essential based on the total value (total relative importance value for each attribute).

Table 7. The most important factors impacting the readiness for BDA.

Variable	Technique				Total value
	CHAID	Bayesian networks	Neural network	C5.0	
Management support	0.3566	0.3627	0.1179	0.2029	1.0401
Data quality	0.2079	0.2007	0.0973	0.2073	0.7132
Firm size	0.1485	0.0675	0.0000	0.1099	0.3259
Data security	0.0954	0.0455	0.1398	0.0000	0.2807
Cost	0.0000	0.0000	0.1613	0.0779	0.2392
Sector	0.0759	0.0293	0.0516	0.0685	0.2253
Competitive pressure	0.0028	0.0936	0.0610	0.0297	0.1871
Partner pressure	0.0000	0.0728	0.0775	0.0000	0.1503
Gender	0.0000	0.0000	0.0410	0.0958	0.1368
Government support	0.0254	0.0000	0.0720	0.0000	0.0974
Technical competence	0.0000	0.0507	0.0443	0.0000	0.0950
IT infrastructure	0.0000	0.0000	0.0000	0.0747	0.0747
Age	0.0028	0.0000	0.0000	0.0541	0.0569
Decision-making culture	0.0000	0.0349	0.0000	0.0000	0.0349
Education level	0.0343	0.0000	0.0000	0.0000	0.0343
Relative advantage	0.0000	0.0288	0.0000	0.0000	0.0288

To get an overview of the gauge result of the four models, we consolidated the values of the four models. The mix of these prescient models is known as aggregation-based sensitivity examination and is suggested in light of the fact that it produces hearty, exact models [76,81]. As a result, the sixteen input variables were categorized into four dimensions—technology dimension (relative advantage, IT infrastructure, data quality, data security, technical competence), organization dimension (top management support, cost, sector, firm size, decision making culture), environment dimension (competitive pressure, partner pressure, government support) and manager's characteristics dimension

(gender, age, education level)—that have an impact on the readiness of BDA adoption. The major predictor variables for BDA adoption among Vietnam SMEs were identified to be management support, data quality, firm size, data security and cost.

4.3. Predicting the effects of observed variables on the readiness to adopt big data in SMEs

Based on the results of the evaluation of the four forecasting models, the C5.0 is the model with the highest predictive accuracy. Therefore, the authors used the C5.0 model to evaluate in detail the observed variables affecting the readiness to use BDA in SMEs. The output variable was the readiness to apply BDA among SMEs (Low readiness and High readiness), and input variables were firm size, sector, gender, age and education level. The stream of the C5.0 model is presented in Figure 5.

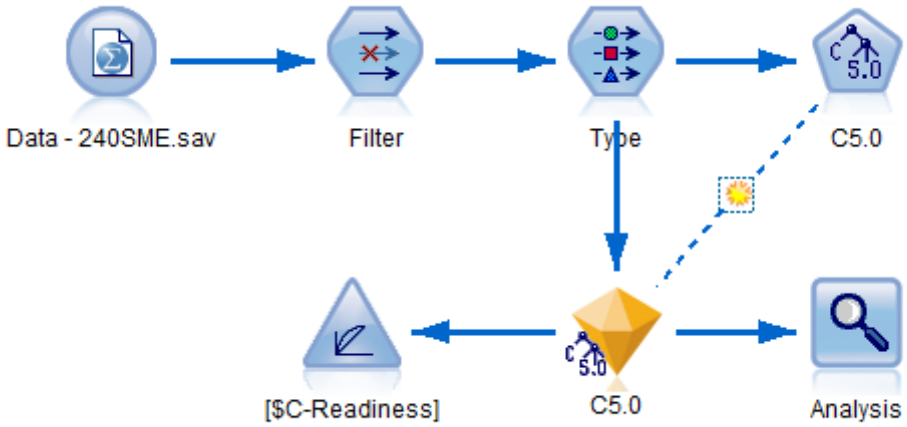

Figure 5. Stream of C5.0 model with five observed variables.

The process of the C5.0 model consists of five input observed variables. This model used the whole dataset, with the result of a correct prediction percentage of 73.75% and an AUC value of 0.758. This proves that the model has high-performance measurements. The results of the model represented three descriptors splitting nodes (firm size, sector and age).

Figure 6 illustrates the results of the decision tree of the C5.0 model. The first splitting node of readiness to apply BDA in SMEs was firm size. In node 1, the proportion of small companies that are not ready to adopt BDA is 57.58%, while the number of small companies with high readiness is lower (42.42%). Next, node 1 diverged into nodes 2 and 3. In node 2, 69.83% of manufacturing companies were still not ready to adopt BDA, and only 30.17% of companies had high readiness. In node 3, the rate of the services companies' readiness to apply BDA is high, 59.76%, and the figure for low willingness companies was 40.24%. Next, node 3 diverged into nodes 4 and 5. In node 4, 70.97% of service companies with leaders under 46 have a high level of readiness to adopt BDA, whereas only 29.03% of service companies have low readiness. Otherwise, in node 5, with leaders aged 46 and over, the percentage of companies willing to adopt BDA (25.00%) was lower than the percentage of companies that were not ready to adopt BDA (75.00%). Finally, in node 6, the majority of medium companies have a high willingness to adopt BDA (88.09%), whereas only 11.91% of medium

enterprises have low readiness.

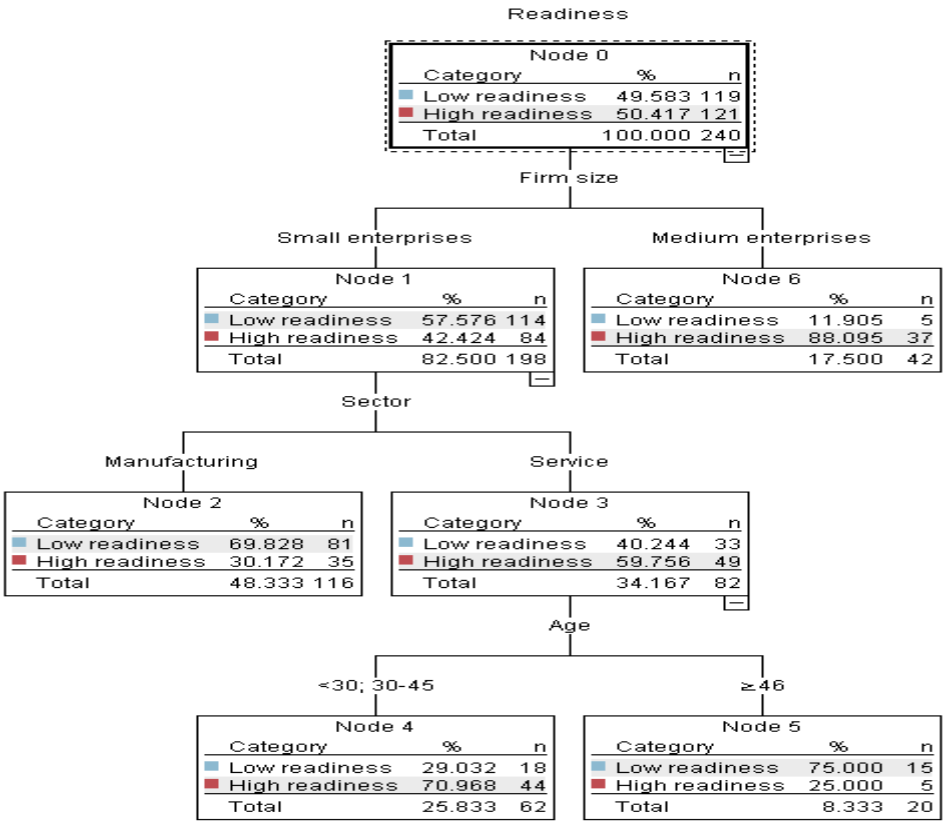

Figure 6. Prediction readiness to apply BDA by observed variables (C5.0 model).

5. Discussion

The findings of the current study demonstrated that sixteen factors of four dimensions (technology, organization, environment and manager's characteristics) have impacts on the readiness to adopt BDA. Furthermore, management support, data quality, firm size, data security and cost were revealed as major predictors of the readiness to apply BDA among Vietnamese SMEs. In addition, medium-sized companies in the service sector are assessed to have higher readiness to apply BDA than other SMEs. In addition, the results of the C5.0 model indicated that firm size, sector and age do have an impact on the BDA adoption readiness.

The results of the study show that management support is the strongest decisive factor in the readiness to apply BDA among Vietnamese SMEs. The result is similar to findings from previous studies such as Sun et al. [35], Maroufkhani et al. [9], Lai et al. [41], Asiaei and Rahim [82]. The support of managers will create favorable conditions for the company in maintaining and using technology [82]. Realizing the benefits of big data, management can allocate the resources needed for adoption and implementation. By contrast, if the management does not see the benefits of big data for businesses, they will oppose its adoption [47].

Generally, data is supposed to be an important input when companies adopt BDA. To perform a successful BDA, data quality is extremely important. Firms have abundant data sources and have high accuracy that will contribute to applying big data readiness. In this study, data quality is a strong factor of BDA adoption, which is consistent with the findings of Park and Kim [23].

Not surprisingly, firm size affected the readiness of BDA adoption. This is consistent with the results of Sohaib et al. [83] and Alshamaila et al. [84]. To be more specific, medium enterprises have higher readiness to adopt BDA than small enterprises. This can be explained by medium-sized companies having larger revenue and more employees than small companies. Therefore, they have many advantages when investing in BDA applications.

Data security was also predicted as a strong influencing factor in this study. Big data includes a lot of personal information [14]; hence, it is of serious concern among firms when deciding to adopt BDA. The influence of data security in technology adoption was also found in many previous studies, such as in software-as-a-service adoption [85], cloud computing [51,83] and big data adoption [23,35,37].

Cost is one of the five factors that are predicted to have an important influence on the readiness of SMEs to adopt BDA. This finding is similar to Park and Kim [23] and Sun et al. [35], who found that cost is an important factor in maintaining and developing the analysis of big data in enterprises. In addition, costs for big data adoption can be a barrier for companies to implementing big data [17,86].

The classification results of the C5.0 model with five observed variables show that the service sector has a higher readiness to apply BDA than the manufacturing sector. This result is consistent with Gangwar [48], who identified factors influencing big data adoption in Indian companies. This is because service organizations like wholesalers, retailers and lodging providers have early access to information technology systems and high-quality human resources to analyze large amounts of data. Moreover, in the context of the complicated development of the COVID-19 pandemic, wholesale and retail companies in Vietnam have had a rapid shift from traditional shopping to online shopping. As a result, organizations must develop suggestion systems and find ways to respond to client information as quickly as possible. Hence, service SMEs are better prepared to adopt BDA. Manufacturing companies are stated to be encountering numerous obstacles, such as a lack of infrastructure and BDA tools, when it comes to using BDA to optimize supply chains [86].

The findings show that small service firms with managers under the age of 46 have a higher readiness to adopt BDA than those firms with older managers. This can be explained by young managers being bolder in adopting new technology, while older managers consider more carefully the necessary conditions when applying BDA, such as information technology, high-quality human resources and finance. In addition, in the implementation of new technologies, some of the older leaders have a lagging mindset, fear of risk and fear of change. This is consistent with the findings of Badri et al. [87], who mentioned that elderly teachers are thought to show less technology readiness than younger teachers.

6. Conclusions and implications

Applying BDA plays an important role in helping organizations improve competitiveness, enhance supply chains, optimize logistics and improve business performance. Based on the data mining technique, the findings of the study show that the C5.0 model is the best model to predict factors affecting BDA adoption readiness in SMEs. Five factors have the greatest influence on the

readiness to adopt BDA: management support, data quality, firm size, data security and cost. Moreover, an important finding of this study is that the age of managers also affects the readiness to adopt BDA.

This study is useful to managers of SMEs, providers and policymakers in developing better policies and strategies for the adoption of BDA. In terms of managers, the volume of data generated in organizations is growing exponentially. So, how to effectively analyze big data is a matter of great interest to organizations today. The proposed model can assist businesses in determining their readiness to adopt BDA. Furthermore, the findings of the study assist managers in increasing their awareness of the elements affecting the enterprise's readiness to use big data. For example, this research shows that management support is the most important factor influencing BDA adoption readiness. As a result, before deciding to embrace BDA, SME management should be proactive in studying to increase their knowledge of the technology and developing a clear strategy. In terms of service providers, the outcomes of this study reveal that SMEs should prioritize data quality, data security and cost factors when preparing to embrace BDA. SMEs, on the other hand, are having financial challenges. As a result, plans for developing BDA tools, hardware, software and other products that meet the needs of providers' clients in emerging and underdeveloped countries should be formed. In addition, when implementing BDA, suppliers must improve services to support SMEs. In terms of policymakers, the survey revealed that the service sector is more prepared to use big data than the manufacturing sector and that medium-sized businesses are more prepared to use big data than small businesses. As a result, the government should have policies in place to assist each sort of business.

Thanks to the great benefits that BDA contributes to business development, a huge number of businesses are interested in BDA. This study has made significant contributions that help practitioners and researchers understand the importance of influencing factors on the readiness to apply big data in SMEs. First, instead of using traditional analytical methods to perform information-based sensitivity analysis, as shown in previous studies, well-known data mining algorithms were used to develop predictive models in this study. Second, this study explored factors that have strong impacts on the readiness to adopt BDA among SMEs. From these findings, the research model is expected to be a useful reference for practitioners in developing countries and the scientific community for doing future related research.

In addition to the study findings, this study also demonstrates some limitations. First is the limitation on the number of samples when using the data mining technique. Therefore, future studies should be conducted with larger sample sizes. Second, the numbers of input variables and prediction algorithms are limited. In future investigations, the number of input variables should increase, and different forecasting algorithms may be used to evaluate the predictive model's findings.

Acknowledgments

The authors would like to thank all respondents who spent valuable time answering questionnaires and the insightful comments of the reviewers.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, et al., Big data: from beginning to future, *Int. J. Inf. Manage.*, **36** (2016), 1231–1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009
2. S. S. Alrumiah, M. Hadwan, Implementing big data analytics in e-commerce: Vendor and customer view, *IEEE Access*, **9** (2021), 37281–37286. https://doi.org/10.1109/ACCESS.2021.3063615
3. T. M. Le, S. Y. Liaw, Effects of pros and cons of applying big data analytics to consumers' responses in an e-commerce context, *Sustainability*, **9** (2017), 1–19. https://doi.org/10.3390/su9050798
4. M. Janssen, H. van der Voort, A. Wahyudi, Factors influencing big data decision-making quality, *J. Bus. Res.*, **70** (2017), 338–345. https://doi.org/10.1016/j.jbusres.2016.08.007
5. G. Wang, A. Gunasekaran, E. W. T. Ngai, T. Papadopoulos, Big data analytics in logistics and supply chain management: Certain investigations for research and applications, *Int. J. Prod. Econ.*, **176** (2016), 98–110. https://doi.org/10.1016/j.ijpe.2016.03.014
6. S. Tiwari, H. M. Wee, Y. Daryanto, Big data analytics in supply chain management between 2010 and 2016: Insights to industries, *Comput. Ind. Eng.*, **115** (2018), 319–330. https://doi.org/10.1016/j.cie.2017.11.017
7. S. Akter, S. F. Wamba, A. Gunasekaran, R. Dubey, S. J. Childe, How to improve firm performance using big data analytics capability and business strategy alignment?, *Int. J. Prod. Econ.*, **182** (2016), 113–131. https://doi.org/10.1016/j.ijpe.2016.08.018
8. P. Mikalef, M. Boura, G. Lekakos, J. Krogstie, Big data analytics and firm performance: Findings from a mixed-method approach, *J. Bus. Res.*, **98** (2019), 261–276. https://doi.org/10.1016/j.jbusres.2019.01.044
9. P. Maroufkhani, M. L. Tseng, M. Iranmanesh, W. K. W. Ismail, H. Khalid, Big data analytics adoption: Determinants and performances among small to medium-sized enterprises, *Int. J. Inf. Manage.*, **54** (2020), 1–15. https://doi.org/10.1016/j.ijinfomgt.2020.102190
10. Z. Xu, G. L. Frankwick, E. Ramirez, Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective, *J. Bus. Res.*, **69** (2016), 1562–1566. https://doi.org/10.1016/j.jbusres.2015.10.017
11. S. Mandal, An examination of the importance of big data analytics in supply chain agility development, *Manag. Res. Rev.*, **41** (2018), 1201–1219. https://doi.org/10.1108/MRR-11-2017-0400
12. L. Wang, M. Yang, Z. H. Pathan, S. Salam, K. Shahzad, J. Zeng, Analysis of influencing ffactors of Big Data Adoption in Chinese enterprises using DANP technique, *Sustainability*, **10** (2018), 1–16. https://doi.org/10.3390/su10113956
13. B. Marr, *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*, Wiley, 2016. https://doi.org/10.1002/9781119278825
14. A. Alharthi, V. Krotov, M. Bowman, Addressing barriers to big data, *Bus. Horiz.*, **60** (2017), 285–292. https://doi.org/10.1016/j.bushor.2017.01.002
15. P. Tabesh, E. Mousavidin, S. Hasani, Implementing big data strategies: A managerial perspective, *Bus. Horiz.*, **62** (2019), 347–358. https://doi.org/10.1016/j.bushor.2019.02.001
16. S. Venkatraman, R. Venkatraman, Big data security challenges and strategies, *AIMS Math.*, **4** (2019), 860–879. https://doi.org/10.3934/math.2019.3.860

17. S. Coleman, R. Göb, G. Manco, A. Pievatolo, X. Tort-Martorell, M. S. Reis, How can SMEs benefit from big data? Challenges and a path forward, *Qual. Reliab. Eng. Int.*, **32** (2016), 2151–2164. https://doi.org/10.1002/qre.2008
18. C. O'Connor, S. Kelly, Facilitating knowledge management through filtered big data: SME competitiveness in an agri-food sector, *J. Knowl. Manag.*, **21** (2017), 156–179. https://doi.org/10.1108/JKM-08-2016-0357
19. P. Del Vecchio, A. Di Minin, A. M. Petruzzelli, U. Panniello, S. Pirri, Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges, *Creat. Innov. Manag.*, **27** (2018), 6–22. https://doi.org/10.1111/caim.12224
20. W. Noonpakdee, A. Phothichai, T. Khunkornsiri, Big data implementation for small and medium enterprises, in *2018 27th Wireless and Optical Communication Conference (WOCC), Hualien, Taiwan*, (2018), 1–5. https://doi.org/10.1109/WOCC.2018.8372725
21. M. H. Chuah, R. Thurusamry, Challenges of big data adoption in Malaysia SMEs based on Lessig's modalities: A systematic review, *Cogent. Bus. Manage.*, **8** (2021), 81–91. https://doi.org/10.1080/23311975.2021.1968191
22. S. K. Mangla, R. Raut, V. S. Narwane, Z. Zhang, P. Priyadarshinee, Mediating effect of big data analytics on project performance of small and medium enterprises, *J. Enterp. Inf. Manag.*, **34** (2020), 168–198. https://doi.org/10.1108/JEIM-12-2019-0394
23. J. H. Park, Y. B. Kim, Factors activating big data adoption by Korean firms, *J. Comput. Inf. Syst.*, **61** (2021), 285–293. https://doi.org/10.1080/08874417.2019.1631133
24. L. G. Tornatzky, M. Fleischer, *The Processes of Technological Innovation*, Lexington Books, Massachusetts, 1990.
25. D. Grant, B. Yeo, A global perspective on tech investment, financing, and ICT on manufacturing and service industry performance, *Int. J. Inf. Manag.*, **43** (2018), 130–145. https://doi.org/10.1016/j.ijinfomgt.2018.06.007
26. R. Y. Zhong, S. T. Newman, G. Q. Huang, S. Lan, Big Data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives, *Comput. Ind. Eng.*, **101** (2016), 572–591. https://doi.org/10.1016/j.cie.2016.07.013
27. M. K. Saggi, S. Jain, A survey towards an integration of big data analytics to big insights for value-creation, *Inf. Proc. Manag.*, **54** (2018), 758–790. https://doi.org/10.1016/j.ipm.2018.01.010
28. S. F. Wamba, A. Gunasekaran, S. Akter, S. J. f. Ren, R. Dubey, S. J. Childe, Big data analytics and firm performance: Effects of dynamic capabilities, *J. Bus. Res.*, **70** (2017), 356–365. https://doi.org/10.1016/j.jbusres.2016.08.009
29. S. Dash, S. K. Shakyawar, M. Sharma, S. Kaushik, Big data in healthcare: management, analysis and future prospects, *J. Big Data*, **6** (2019), 1–25. https://doi.org/10.1186/s40537-019-0217-0
30. E. Yadegaridehkordi, M. Nilashi, L. Shuib, M. Hairul Nizam Bin Md Nasir, S. Asadi, S. Samad, et al., The impact of big data on firm performance in hotel industry, *Electron. Commer. Res. Appl.*, **40** (2020), 1–32. https://doi.org/10.1016/j.elerap.2019.100921
31. R. Dubey, A. Gunasekaran, S. J. Childe, Big data analytics capability in supply chain agility: The moderating effect of organizational flexibility, *Manag. Decis.*, **57** (2019), 2092–2112. https://doi.org/10.1108/MD-01-2018-0119
32. S. Wang, H. Wang, Big data for small and medium-sized enterprises (SME): a knowledge management model, *J. Knowl. Manag.*, **24** (2020), 881–897. https://doi.org/10.1108/JKM-02-2020-0081

33. N. Mahdi, R. Javaneh, S. Mahmoud, The impact of big data adoption on SMEs performance, *Res. Sq.*, **9** (2020), 1–12. https://doi.org/10.21203/rs.3.rs-66047/v1
34. A. Lutfi, A. Alsyouf, M. A. Almaiah, M. Alrawad, A. A. K. Abdo, A. L. Al-Khasawneh, et al., Factors influencing the adoption of big data analytics in the digital transformation era: Case study of Jordanian SMEs, *Sustainability*, **14** (2022), 1–17. https://doi.org/10.3390/su14031802
35. S. Sun, C. G. Cegielski, L. Jia, D. J. Hall, Understanding the factors affecting the organizational adoption of big data, *J. Comput. Inf. Syst.*, **58** (2016), 193–203. https://doi.org/10.1080/08874417.2016.1222891
36. P. Maroufkhani, R. Wagner, W. K. Wan Ismail, M. B. Baroto, M. Nourani, Big data analytics and firm performance: A systematic review, *Information*, **10** (2019), 1–21. https://doi.org/10.3390/info10070226
37. M. I. Baig, L. Shuib, E. Yadegaridehkordi, Big data adoption: State of the art and research challenges, *Inf. Process. Manag.*, **56** (2019), 1–18. https://doi.org/10.1016/j.ipm.2019.102095
38. S. Gupta, H. W. Kim, Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities, *Eur. J. Oper. Res.*, **190** (2008), 818–833. https://doi.org/10.1016/j.ejor.2007.05.054
39. P. F. Hsu, S. Ray, Y. Y. Li-Hsieh, Examining cloud computing adoption intention, pricing mechanism, and deployment model, *Int. J. Inf. Manage.*, **34** (2014), 474–488. https://doi.org/10.1016/j.ijinfomgt.2014.04.006
40. J. H. Park, M. K. Kim, J. H. Paik, The factors of technology, organization and environment influencing the adoption and usage of big data in korean firms, in *26th European Regional Conference of the Interational Telecommunications Society, Madrid, Spain, 24–27 June*, **3** (2015), 121–129.
41. Y. Lai, H. Sun, J. Ren, Understanding the determinants of big data analytics (BDA) adoption in logistics and supply chain management, *Int. J. Logis. Manag.*, **29** (2018), 676–703. https://doi.org/10.1108/IJLM-06-2017-0153
42. K. K. Kapoor, Y. K. Dwivedi, M. D. Williams, Empirical examination of the role of three sets of innovation attributes for determining adoption of IRCTC mobile ticketing service, *Inf. Sys. Manag.*, **32** (2015), 153–173. https://doi.org/10.1080/10580530.2015.1018776
43. E. M. Rogers, Lessons for guidelines from the diffusion of innovations, *Jt. Comm. J. Qual. Improv.*, **21** (1995), 324–328. https://doi.org/10.1016/S1070-3241(16)30155-9
44. M. Ghobakhloo, D. Arias-Aranda, J. Benitez-Amado, Adoption of e-commerce applications in SMEs, *Industrial Manag. Data Syst.*, **111** (2011), 1238–1269. https://doi.org/10.1108/02635571111170785
45. N. Kshetri, Big data's impact on privacy, security and consumer welfare, *Telecommun. Policy*, **38** (2014), 1134–1145. https://doi.org/10.1016/j.telpol.2014.10.002
46. A. A. Jahanshahi, A. Brem, Sustainability in SMEs: Top management teams behavioral integration as source of innovativeness, *Sustainability*, **9** (2017), 1–16. https://doi.org/10.3390/su9101899
47. S. Shamim, J. Zeng, S. M. Shariq, Z. Khan, Role of big data management in enhancing big data decision-making capability and quality among Chinese firms: A dynamic capabilities view, *Inf. Manag.*, **56** (2019), 1–16. https://doi.org/10.1016/j.im.2018.12.003
48. H. Gangwar, Understanding the determinants of big data adoption in India: An analysis of the manufacturing and services sectors, *Inf. Resour. Manag. J.*, **31** (2018), 1–22. https://doi.org/10.4018/IRMJ.2018100101

49. W. Xu, P. Ou, W. Fan, Antecedents of ERP assimilation and its impact on ERP value: A TOE-based model and empirical test, *Inf. Syst. Front.*, **19** (2017), 13–30. https://doi.org/10.1007/s10796-015-9583-0
50. C. Low, Y. Chen, M. Wu, Understanding the determinants of cloud computing adoption, *Ind. Mana. Data Syst.*, **111** (2011), 1006–1023. https://doi.org/10.1108/02635571111161262
51. J. W. Lian, D. C. Yen, Y. T. Wang, An exploratory study to understand the critical factors affecting the decision to adopt cloud computing in Taiwan hospital, *Int. J. Inf. Manag.*, **34** (2014), 28–36. https://doi.org/10.1016/j.ijinfomgt.2013.09.004
52. K. Zhu, K. L. Kraemer, S. Xu, J. Dedrick, Information technology payoff in e-business environments: an international perspective on value creation of e-business in the financial services industry, *J. Manag. Inf. Syst.*, **21** (2004), 17–54. https://doi.org/10.1080/07421222.2004.11045797
53. T. H. Kwon, J. H. Kwak, K. Kim, A study on the establishment of policies for the activation of a big data industry and prioritization of policies: Lessons from Korea, *Technol. Forecast. Soc. Change*, **96** (2015), 144–152. https://doi.org/10.1016/j.techfore.2015.03.017
54. J. I. Rojas-Méndez, A. Parasuraman, N. Papadopoulos, Demographics, attitudes, and technology readiness, *Mark. Intell. Plan.*, **35** (2017), 18–39. https://doi.org/10.1108/MIP-08-2015-0163
55. A. Parasuraman, C. L. Colby, An updated and streamlined technology readiness index: TRI 2.0, *J. Serv. Res.*, **18** (2014), 59–74. https://doi.org/10.1177/1094670514539730
56. T. Wendler, S. Gröttrup, *Data Mining with Spss Modeler: Theory, Exercises, and Solutions*, 1st edition, Springer Cham, Switzerland, 2016. https://doi.org/10.1007/978-3-319-28709-6
57. X. S. Yang, *Introduction to Algorithms for Data Mining and Machine Learning*, Elsevier Inc, 2019. https://doi.org/10.1016/C2018-0-02034-4
58. SPSS, *IBM SPSS Decision Tree 2*, SPSS: Chicago, IL, USA, 2012.
59. P. Cortez, A. M. G. Silva, Using data mining to predict secondary school student performance, in *Proceedings of 5th Annual Future Business Technology Conference, Porto*, (2008), 5–12.
60. E. Yukselturk, S. Ozekes, Y. K. Türel, Predicting dropout student: An application of data mining methods in an online education program, *Eur. J. Open, Distance E-learn.*, **17** (2014), 118–133. https://doi.org/10.2478/eurodl-2014-0008
61. C. M. Zhao, J. Luan, Data mining: Going beyond traditional statistics, *New Dir. Institutional Res.*, **131** (2006), 7–16. https://doi.org/10.1002/ir.184
62. K. D. Brouthers, L. E. Brouthers, Why service and manufacturing entry mode choices differ: The influence of transaction cost factors, risk and trust*, *J. Manag. Stud.*, **40** (2003), 1179–1204. https://doi.org/10.1111/1467-6486.00376
63. K. Ferdows, A. De Meyer, Lasting improvements in manufacturing performance: In search of a new theory, *J. Oper. Manag.*, **9** (1990), 168–184. https://doi.org/10.1016/0272-6963(90)90094-T
64. M. Ghasemaghaei, The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage, *Int. J. Inf. Manag.*, **50** (2020), 395–404. https://doi.org/10.1016/j.ijinfomgt.2018.12.011
65. K. K. Y. Kuan, P. Y. K. Chau, A perception-based model for EDI adoption in small businesses using a technology-organization-environment framework, *Inf. Manag.*, **38** (2001), 507–521. https://doi.org/10.1016/S0378-7206(01)00073-8
66. G. Premkumar, M. Roberts, Adoption of new information technologies in rural small business, *OMEGA-Int. J. Manag. Sci.*, **27** (1999), 467–484. https://doi.org/10.1016/S0305-0483(98)00071-1

67. E. Yadegaridehkordi, M. Hourmand, M. Nilashi, L. Shuib, A. Ahani, O. Ibrahim, Influence of big data adoption on manufacturing companies' performance: An integrated DEMATEL-ANFIS approach, *Technol. Forecasting Social Change*, **137** (2018), 199–210. https://doi.org/10.1016/j.techfore.2018.07.043
68. M. Nasrollahi, J. Ramezani, A Model to evaluate the organizational readiness for big data adoption, *Int. J. Comput. Communi. Cont.*, **15** (2020), 1–11. https://doi.org/10.15837/ijccc.2020.3.3874
69. J. F. Hair, *Multivariate Data Analysis: A Global Perspective*, Pearson Education: Upper Saddle River, NJ, USA; London, UK, 2010.
70. J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, *Multivariate Data Analysis*, Peason, 2014.
71. B. M. Byrne, *Structural Equation Modeling with Amos: Basic Concepts, Applications, and Programming*, 3rd edition, Routledge: Abingdon, UK, 2016. https://doi.org/10.4324/9781315757421
72. C. Fornell, D. F. Larcker, Evaluating structural equation models with unobservable variables and measurement error, *J. Mar. Res*., **18** (1981), 39–50. https://doi.org/10.1177/002224378101800104
73. G. V. Kass, An exploratory technique for investigating large quantities of categorical data, *J. R. Stat. Soc.*, **29** (1980), 119–127. https://doi.org/10.2307/2986296
74. J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann Inc.: San Mateo, CA, USA, 1988. https://doi.org/10.1016/B978-0-08-051489-5.50008-4
75. J. R. Quinlan, Induction of decision trees, *Mach. Learn.*, **1** (1986), 81–106. https://doi.org/10.1007/BF00116251
76. D. Delen, C. Kuzey, A. Uyar, Measuring firm performance using financial ratios: A decision tree approach, *Expert Syst. Appl.*, **40** (2013), 3970–3983. https://doi.org/10.1016/j.eswa.2013.01.012
77. M. Taamneh, Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm, *J. Saf. Res.*, **66** (2018), 121–129. https://doi.org/10.1016/j.jsr.2018.06.002
78. Z. Chen, M. Yang, Y. Wen, S. Jiang, W. Liu, H. Huang, Prediction of atherosclerosis using machine learning based on operations research, *Math. Biosci. Eng.*, **19** (2022), 4892–4910. https://doi.org/10.3934/mbe.2022229
79. K. A. Tavakoli, R. Rabieyan, M. M. Besharati, A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers, *J. Safety Res.*, **51** (2014), 93–98. https://doi.org/10.1016/j.jsr.2014.09.004
80. D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, *Artif. intell. med.*, **34** (2005), 113–127. https://doi.org/10.1016/j.artmed.2004.07.002
81. R. Sann, P. C. Lai, S. Y. Liaw, C. T. Chen, Predicting online complaining behavior in the hospitality industry: Application of big data analytics to online reviews, *Sustainability*, **14** (2022), 1–22. https://doi.org/10.3390/su14031800
82. A. Asiaei, N. Z. A. Rahim, A multifaceted framework for adoption of cloud computing in Malaysian SMEs, *J. Sci. Technol. Policy Manag.*, **10** (2019), 708–750. https://doi.org/10.1108/JSTPM-05-2018-0053
83. O. Sohaib, M. Naderpour, W. Hussain, L. Martinez, Cloud computing model selection for e-commerce enterprises using a new 2-tuple fuzzy linguistic decision-making method, *Comput. Ind. Eng.*, **132** (2019), 47–58. https://doi.org/10.1016/j.cie.2019.04.020
84. Y. Alshamaila, S. Papagiannidis, F. Li, Cloud computing adoption by SMEs in the north east of England, *J. Enterp. Inf. Manag.*, **26** (2013), 250–275. https://doi.org/10.1108/17410391311325225

85. Z. Yang, J. Sun, Y. Zhang, Y. Wang, Understanding SaaS adoption from the perspective of organizational users: A tripod readiness model, *Comput. Human. Behav.*, **45** (2015), 254–264. https://doi.org/10.1016/j.chb.2014.12.022
86. M. A. Moktadir, S. M. Ali, S. K. Paul, N. Shukla, Barriers to big data analytics in manufacturing supply chains: A case study from Bangladesh, *Comput. Ind. Eng.*, **128** (2019), 1063–1075. https://doi.org/10.1016/j.cie.2018.04.013
87. M. Badri, A. Al Rashedi, G. Yang, J. Mohaidat, A. Al Hammadi, Technology readiness of school teachers: An empirical study of measurement and segmentation, *J. Inf. Technol. Educ.: Res.*, **13** (2014), 257–275.https://doi.org/10.28945/2082

Appendix A

Table A1. Reliability and validity assessment.

Variable	Item number	Factor loadings	Cronbach α	CR	AVE
Relative advantage	4	0.530–0.865	0.807	0.805	0.519
IT structure	3	0.617–0.848	0.798	0.809	0.590
Data quality	3	0.569–0.728	0.691	0.714	0.457
Data security	3	0.705–0.756	0.716	0.764	0.520
Technical competence	4	0.744–0.816	0.867	0.867	0.620
Management support	3	0.520–0.781	0.707	0.722	0.471
Cost	3	0.755–0.849	0.798	0.843	0.643
Decision-making culture	3	0.615–0.846	0.746	0.768	0.528
Competitive pressure	3	0.787–0.842	0.856	0.856	0.664
Partner pressure	3	0.690–0.717	0.626	0.751	0.501
Government support	3	0.703–0.727	0.719	0.757	0.509
Readiness to adopt big data	9	0.684–0.810	0.773	0.909	0.526

*Note: CR: Composite Reliability, AVE: Average Variance Extracted

Table A2. The results of the 10-fold cross-validation for the four model types.

Fold No.	CHAID			Bayesian networks			Neural network			C5.0		
	Confusion matrix		Accuracy	Confusion matrix		Accuracy	Confusion matrix		Accuracy	Confusion matrix		Accuracy
1	12	5	0.649	13	4	0.784	12	5	0.703	14	3	0.892
	11	9		4	16		6	14		1	19	
2	21	11	0.729	28	4	0.847	26	6	0.780	27	5	0.881
	5	22		5	22		7	20		2	25	
3	28	13	0.684	37	4	0.855	27	14	0.711	36	5	0.895
	11	24		7	28		8	27		3	32	

Continued on next page

Fold No.	CHAID			Bayesian networks			Neural network			C5.0		
	Confusion matrix		Accuracy	Confusion matrix		Accuracy	Confusion matrix		Accuracy	Confusion matrix		Accuracy
4	32	21	0.615	49	4	0.846	36	17	0.692	45	8	0.875
	25	26		12	39		15	36		5	46	
5	53	10	0.736	58	5	0.840	46	17	0.720	55	8	0.888
	23	39		15	47		18	44		6	56	
6	46	25	0.697	64	7	0.828	51	20	0.752	62	9	0.890
	19	55		18	56		16	58		7	67	
7	66	13	0.669	71	20	0.828	65	14	0.761	69	9	0.883
	41	43		8	64		25	59		10	75	
8	84	12	0.728	87	23	0.836	62	34	0.692	83	13	0.887
	41	58		9	76		26	73		9	90	
9	93	14	0.657	96	11	0.833	86	21	0.681	93	14	0.884
	60	49		25	84		48	61		11	98	
10	62	55	0.627	105	12	0.835	44	18	0.732	101	16	0.877
	33	86		27	92		22	65		13	106	
Average			**0.679**			**0.833**			**0.722**			**0.885**

Confusion matrix illustrates the classification of the cases in the test dataset. In the confusion matrix, the columns represent the actual cases, and the rows represent the predicted. Accuracy = (TP + TN)/(TP + FP + TN + FN).