v1: 13 August 2024

Research Article

Application of Data Mining Combined with K-means Clustering Algorithm in Enterprises' Risk Audit

Peer-approved: 13 August 2024

Qeios, Vol. 6 (2024)
ISSN: 2632-3834

Sharif Uddin Ahmed Rana[1]

1. Limkokwing University of Creative Technology, Putrajaya, Malaysia

The financial risk management mechanism of enterprises can be more complete through exploration of the application effect of data mining technology combined with the K-means clustering algorithm in enterprise risk audit. Hence, the K-means clustering algorithm is introduced to study the paperless status of electronic payment in the trading process of e-commerce enterprises. Additionally, a risk audit model of e-commerce enterprises is implemented based on the K-means algorithm combined with the Random Forest Light Gradient Boosting Machine (RF-LightGBM). In this model, the actual operation processes of data preparation, data preprocessing, model construction, model application, and evaluation are implemented to study the payment flow in the transaction process of e-commerce enterprises by using big data analysis technology. Eventually, the performance of the model is evaluated by simulation. The results show that, compared with the models and algorithms proposed by scholars in other related fields, the classification accuracy of the model proposed here reaches 95.46%. Simultaneously, the data message delivery rate of the model algorithm is basically stable at about 81.54%, and the data message leakage rate, packet loss rate, and average delay are lower than those of other models and algorithms. Therefore, under the premise of ensuring prediction accuracy, the audit model of e-commerce enterprises can also achieve high data transmission security performance, which can provide an experimental basis for the safety improvement and risk control of the audit process in e-commerce enterprises.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

I. Introduction

At present, along with the advancing progress of communication technologies such as wireless transmission and mobile equipment, e-commerce enterprises are also developing rapidly. Meanwhile, the emergence of computer technologies like cloud computing, artificial intelligence (AI), and big data (BD) has made e-commerce industries more intelligent. According to statistical analysis, the total volume of business in e-commerce industries has almost reached 3 billion yuan. The proportion of the online shopping market in the e-commerce market was only about 11% in 2011, and by 2018 it grew to nearly 30%. The rapid popularization of e-commerce in people's lives has also brought new business directions and certain risks to accounting and auditing work, such as paperless transaction processes and electronic payments [1,2]. Therefore, how to adopt data mining (DM) in the risk audit of e-commerce enterprises has raised a group of relevant scholars' interest.

Compared with the transaction process of traditional enterprises, e-commerce enterprises usually use the Internet, and electronic payments such as Alipay and WeChat payment are mainly taken in the transaction process online. Therefore, this process is also called an e-

 doi.org/10.32388/G9GoS3.2

commerce transaction. Online payment makes the transaction process of e-commerce basically paperless [3]. The electronic payment process also makes the audit of e-commerce enterprises different from traditional audits in many aspects, such as audit objectives, audit content, audit methods, and procedures. Therefore, the audit contents of e-commerce enterprises do not include just enterprises' annual financial reports, but also the analysis of whether the e-commerce system is still reliable. As the audit content of e-commerce enterprises is more extensive, the audit risk is further increased. Additionally, with the increasing number and scale of e-commerce enterprises, more and more e-commerce servers need certified public accountants to supervise the audit, which also makes the audit inevitably face some risks. DM technology can classify and predict sample data when auditing transaction data information in e-commerce enterprises [4]. The advantages of the e-commerce audit risk identification model based on DM technology are that it does not need the tedious conditional assumptions in the traditional statistical model, but is suitable for mass data identification without structure and form, and it uses a computer to effectively process information and to quickly obtain the identification results. It has the advantages of high identification efficiency and a good identification effect.

When auditing the transaction data information of massive e-commerce enterprises, the audit model has also changed from a model based on statistical technology to a model based on intelligent technology, including multivariate analysis, Logistic, AI, support vector machine (SVM), K-means clustering algorithm, decision tree (DT), neural network, random forest (RF), etc., which are all intelligent algorithms that can perform unsupervised learning of multi-level features of data from the original massive data in an unsupervised state. Among them, many studies have shown that the RF algorithm has incomparable advantages over other machine learning (ML) algorithms. It can not only deal with large data sets with high efficiency and high quality but also maintain excellent prediction accuracy under the condition of high-dimensional features. It sorts the importance of input variables and has strong adaptive ability and self-learning ability. It is very suitable for nonlinear modeling without the influence of multiple collinearities[5,6]. Moreover, the RF algorithm can overcome the over-fitting problem existing in other models and has been widely used in many fields, such as bioinformatics, medicine, and social science. The K-means algorithm is used to make iterative clustering analysis. When it is carried out, K objects are randomly selected as the clustering centers in the first stage. Then, the distance between each object and the distance of each seed clustering center is calculated. By these steps, each object is assigned to the nearest clustering center. Clustering centers and their assigned objects refer to a clustering. After one more sample is assigned, another calculation should be conducted on the cluster center, according to the cluster's existing objects. The process should be repeated until the model meets the termination condition [7]. Applying it to enterprise audit can effectively classify the transaction process and data, which is of great significance to enterprise risk audit.

To sum up, with the AI algorithms being continuously improved, the audit work is facing not only a rare opportunity but also a big challenge in the widespread popularity of e-commerce companies. It is innovative to introduce the K-means clustering algorithm into the e-commerce industry. Meanwhile, integrated with the improved RF algorithm in the ML algorithm, the Random Forest Light Gradient Boosting Machine (RF-LightGBM) fusion algorithm is designed. Construction is conducted on the risk audit model of e-commerce enterprises based on the K-means algorithm combined with RF-LightGBM. Ultimately, through simulation, its performance is evaluated to provide experimental reference values for later audit risk reduction and quality improvement.

II. Recent Related Work

A. Tendency of the Development of Enterprises' Risk Audit

With the advent of e-commerce, paperless transactions not only affect traditional manufacturing production but also bring new risks to accounting and auditing work. Many scholars have studied the risk audit of enterprises. Shad et al. (2019) combined the implementation of enterprise risk management with sustainable development reports to test the influence of risk audit on the enterprises' economic added value. Simultaneously, they proposed using ordinary least squares (OLS) analysis to obtain information about enterprise risk management practices and sustainability reports [8]. Hanggraeni et al. (2019) displayed significant results of risk management factors by using an offline questionnaire survey. Simultaneously, through the marketing and financial management risk audit assessment, they found that enterprise identification and management activities would have a vital influence on business performance [9]. Cheng et al. (2021) proposed a new q-rung orthopair fuzzy weighted averaging operator (q-ROFWAO) to rank and evaluate manufacturing small and middle-sized enterprises (SMEs). The results show that the method is effective in Sustainability Enterprise Risk Management (SERM) of SMEs [10]. Yang et al. (2021) introduced the time dimension to describe the dynamic, sudden, and timely evolution characteristics of enterprise risk events in view of the static mapping problem in the knowledge map of existing enterprises. A ResNet dynamic knowledge reasoning method was also proposed to improve the loss

doi.org/10.32388/G9GoS3.2

balance function of a multi-network model. Experiments show that the new model can effectively improve the accuracy of entity and relationship prediction [11].

B. Application Status of the DM Technology

With the Internet and information technology becoming increasingly advanced, the scale of data in all walks of life has been increasing. As one of the AI algorithms, ML has a wide range of applications, which provides a strong guarantee for DM of massive data information and is studied by many scientific researchers. Ping et al. (2019) introduced two ML methods for evaluating the fuel efficiency of driving behavior using natural driving data. Results indicate that this method can be adopted to make effective identification of the relationship between driving behavior and fuel consumption from the macro or micro levels and can effectively predict the driving behavior of vehicles [12]. Xu et al. (2019) proposed detailed methods rooted in remote sensing, ML, and computer vision, and made full use of existing data to combine convolutional neural networks (CNN) with subtle and scientific observation data of the earth [13]. In view of the current situation of flight delays, Gui et al. (2019) designed a normalized model using the LSTM algorithm and RF algorithm to classify and predict flight conditions. The results show that the proposed model based on a random forest can obtain higher prediction accuracy (binary classification is 90.2%) and overcome overfitting [14]. Lv et al. (2020) constructed a cognitive computing model by context-aware data flow by optimizing the decision tree algorithm in ML. The results show that the application of the model algorithm can ensure the accuracy and stability of behavior classification, which is of great significance for operators to analyze user behavior and develop personalized services [15]. Kanagaraj et al. (2021) put forward an enhanced multi-class normalized optimal clustering algorithm and applied it to data object grouping and classification. The results outline the needs of different regions in India in terms of energy consumption and show that the proposed method performs significantly better [16].

Through the analysis of the research of the above scholars, it is found that in the era of massive data generation, the application field of DM technology is becoming gradually more extensive. In the field of risk audit, most enterprises still adapt the traditional manufacturing audit work. Under the trend of rapid popularization in the field of e-commerce, there are not many studies related to risk audit. Therefore, aiming at the risks existing in the audit process of e-commerce enterprises in the field of e-commerce, the standard algorithm is optimized by the ML algorithm, and construction is conducted on the risk audit evaluation model of e-commerce enterprises, which is of great significance to the safety improvement and risk control of the audit process in e-commerce enterprises.

III. Construction and analysis of risk audit evaluation of e-commerce enterprises based on dm technology

A. Requirements Analysis of Data Sources and Risk Audit of E-Commerce Enterprises

In e-commerce enterprises, data sources are more extensive than financial transactions with the traditional supply chain. E-commerce enterprises can improve the real-time update frequency of data through multi-dimensional BD acquisition, thus improving the effectiveness of data audit. Fig. 1 illustrates the BD sources of e-commerce enterprises.

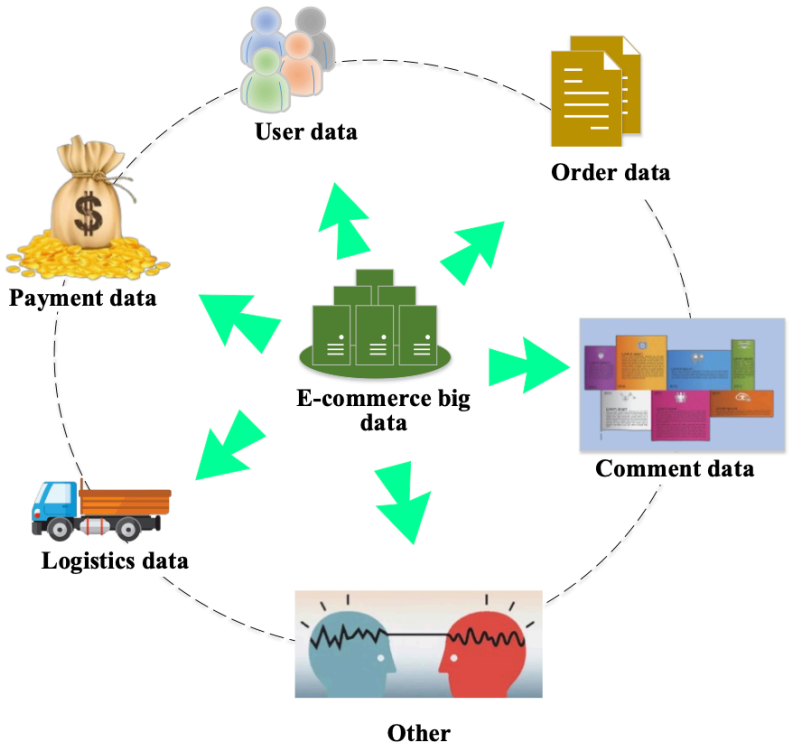

Fig. 1. BD resources of e-commerce enterprises

The collection of multi-dimensional data shown in Fig. 1 can effectively reduce information asymmetry and false information, which is conducive to the operation of the BD risk control model in the later stage to avoid errors in the analysis results due to insufficient user information, and to better prevent and control various audit risks. Moreover, through the analysis of the credit evaluation model constructed by diversified and deep-seated multidimensional big data, it is conducive to more accurate credit evaluation of users.

The audit mode and method of e-commerce business transactions are also constantly updated with the change of BD technology. The complexity of data types, the expansion of the data analysis scope, and the increase in audit projects put forward higher requirements for the professional quality of internal auditors. The practical application of BD audit mainly includes the configuration of personnel professional knowledge, the configuration of hardware and software equipment, and the sufficient and accurate data required [17,18]. Primarily, in terms of personnel professional quality, since the BD audit is still in the developing stage, some technical personnel in the audit department still adopt the traditional risk-oriented internal audit work method, which requires the company to increase the introduction of employees and personnel training. There are differences in the development of hardware and software equipment. Moreover, in terms of data preparation, if the business data collected in the audit operations are incomplete, the audit results will be affected. Data, as the basis of audit judgment, will influence the quality of internal audit to a certain degree. If there are information mis-records or missing records and cross-system extraction failures, the audit results will be misleading. Therefore, when auditing the risk of data information in e-commerce enterprises, DM is inevitable. Here, the combination of the ML K-means clustering algorithm and RD is used to audit the risk of BD of transactions in e-commerce enterprises, which is significantly practical for the transaction information security and accurate audit of e-commerce enterprises.

B. Random Forest Algorithm Applied to Big Data Audit Analysis of E-commerce Enterprises

When auditing the transaction data of e-commerce enterprises, classifying the data is the primary work to be carried out. RF is an improved algorithm based on the common decision tree, which has more advanced advantages than common decision trees. It can generate training samples independently in each decision tree, and then form a forest. Ultimately, the results of multiple

doi.org/10.32388/G9GoS3.2

decision trees are combined by using some strategies. Based on the DT algorithm, a random forest is formed [19]. Fig.2 demonstrates the DT algorithm applied to BD in e-commerce enterprises.

doi.org/10.32388/G9GoS3.2

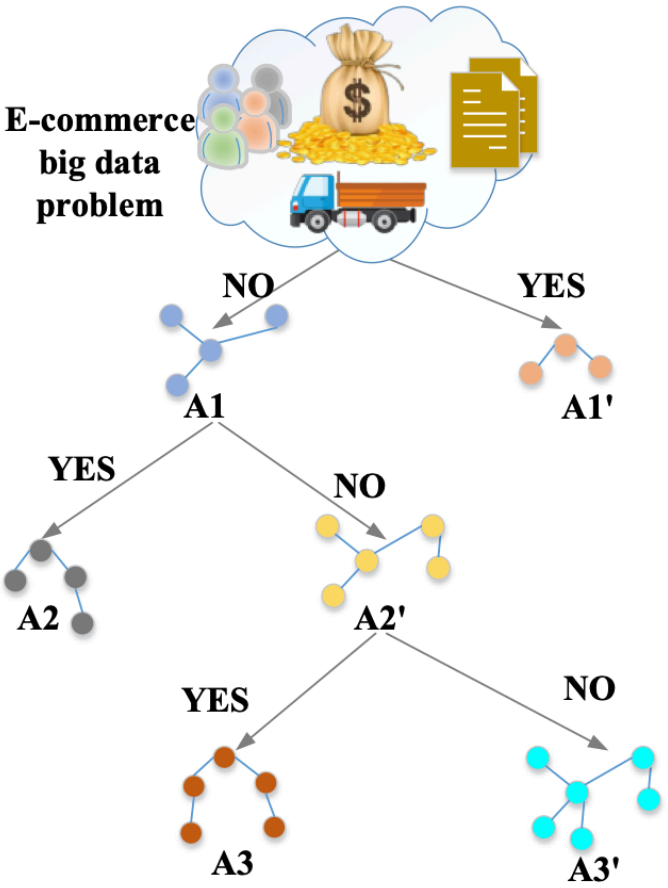

Fig. 2. DT algorithm applied to BD in e-commerce enterprises

As for the general DT algorithm, the segmented node usually selects an optimal feature attribute from all the sample feature attributes on the node as the basis of the segmented node [20]. However, RF randomly selects some feature attributes on the current node, and then selects an optimal feature attribute as the basis for dividing the node. In this way, RF further enhances the generalization ability of the model. Compared with the DT method, the RF algorithm is more effective in solving the problem of overfitting in DT. After the establishment of the RF, assuming that there is a new sample, it is put into the RF, and then each DT in the RF enters the sample attribute category for decision-making. Each tree has a vote, with a few subordinates to the majority method; the category with the largest number of DT votes is the final classification result of the sample [21]. The RF algorithm is applied to DT of e-commerce as shown in Fig. 3.

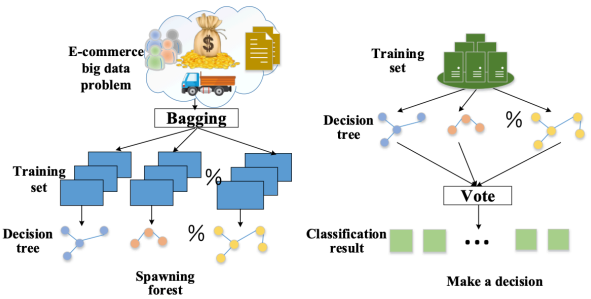

Fig. 3. RF algorithm applied to DT of e-commerce

As shown in the right part of Fig. 3, the construction of the RF algorithm includes three steps: training set generation, decision tree construction, and algorithm formation and implementation. Assuming that the scale of the random forest is *N*, the random forest algorithm needs *N* decision trees for training. Hence, a corresponding number of training sets need to be generated. To prevent the DT from producing a local optimal solution, the RF generates *N*

doi.org/10.32388/G9GoS3.2

training sample sets by using the bag random sampling technique with replacement. Besides, this operation will inevitably lead to repetition in the sampling of training samples.

Iterative Dichotomiser 3 (ID3 algorithm) (ID3 algorithm refers to a greedy algorithm used to construct decision trees) is one of the most basic algorithms in RF. The algorithm first calculates the information gain of each attribute, and then compares the information gain of each feature one by one, and selects the best attribute for node segmentation [22]. The so-called best attribute refers to the maximum information gain obtained by dividing the sample set according to the characteristics. Information entropy is a basic concept in algorithm operation, which is used to measure uncertainty. Equation (1) indicates the sample set of the decision tree at node *m*.

$$X = \{x_1, x_2, \cdots, x_n\} \tag{1}$$

The corresponding sample category can be expressed as:

$$\{\, c_i | \, i = 1, 2, \cdots, N\} \tag{2}$$

$p_i$represents the probability for each category, and *X* accords to the information gain obtained by dividing the sample by m corresponding to attribute *a*.

$$Gain(a) = Info(X) - Infoa(X) \tag{3}$$

In Equation (3),$Info(X)$means the information entropy of.

$$Info(X) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{4}$$

$Infoa(X)$ stands for the predicting information required by *X*.

$$Infoa(X) = \sum_{j=1}^{v} \left[\left(\frac{|X_j|}{|X|} \right) Infoa\left(X_j\right) \right] \tag{5}$$

The ID3 algorithm selects the maximum attribute as the test attribute. However, it cannot handle continuous variables and prefers to select properties with more values. Therefore, the ID3 algorithm usually leads to the DT solution being a local optimal solution rather than a global optimal solution. Scholars have conducted in-depth research and discussion on this issue and finally proposed the C4.5 algorithm. The C4.5 algorithm is based on the information gain rate. This algorithm uses the information gain rate to avoid the deviation of segmentation attributes, making it more equitable to select each attribute when dividing nodes [23]. Equation (6) illustrates the calculation of the information acquisition rate.

$$GainRatio(a) = \frac{Gain(a)}{splitInfoa(X)} \tag{6}$$

In Equation (6), $splitInfoa(X)$represents the information segmentation rate, and Equation (7) expresses it as a function.

$$splitInfoa(X) = \sum_{j=1}^{v} \left[\left(\frac{|X_j|}{|X|} \right) \log_2 \left(\frac{X_j}{X} \right) \right] \tag{7}$$

Although compared with the ID3 algorithm, the C4.5 algorithm can discretize the original continuous attribute variables, it can handle continuous numerical variables and is also suitable for missing data [24]. The classification rules generated by the C4.5 algorithm are easy to understand and have high precision; however, the algorithm is not dominant in execution time and storage space.

Meanwhile, the classification and regression tree (CART) algorithm is also very common in the RF algorithm. Different from the ID3 algorithm and the C4.5 algorithm, the CART algorithm uses the Gini minimum impurity criterion for node segmentation. Equation (8) displays the calculation process of the Gini minimum impurity criterion.

$$Gini(t) = 1 - \sum_{j=1}^{c} [p(\, j | \, t)\,]^{2} \quad j = 1, \cdots, c \tag{8}$$

In Equation (8), $p(\, j | \, t)$ refers to the probability of type *j* on node *t*. When the same category is composed of all the samples of node *t*, the minimum value is given to the Gini index, namely, 0, and the sample category is the purest. When the Gini index is the maximum 1, the purity of the sample category is the lowest, that is, categories are different. The sample set is divided into *m* branches, and Equation (9) expresses the Gini index used to split the current node.

$$Gini(X) = \sum_{i=1}^{m} \frac{n_i}{n} Gini(i) \tag{9}$$

In Equation (9), *m* refers to the number of sub-nodes, n_i accords to the number of samples at sub-node *i*, and *n* represents the number of samples at the upper node. Moreover, the application of the CART algorithm needs to calculate the Gini index of each attribute in the training process. After the variables with the smallest Gini index are selected to segment the current node, the decision tree needs to be recursively constructed until it reaches the stopping condition.

However, the Light Gradient Boosting Machine (LightGBM) algorithm, as an open-source and efficient distributed gradient boosting tree algorithm newly released in recent years, has the characteristics of fast operation, less memory consumption, and high accuracy, and is widely used in classification and regression. In the Gradient Boosting Decision Tree (GBDT) iteration, it is

assumed that the learner obtained in the previous round is defined as $Z_{t-1}(x)$, whose loss function accords with Equation (10).

$$L\left(Y, Z_{t-1}(x)\right) \tag{10}$$

Then, the goal of this training is to find a suitable weak learner to minimize the loss function. Equation (11) defines the loss function.

$$Z_t(x) = \underset{h \in H}{argmin} \sum L\left(y, Z_{t-1}(x) + h_t(x)\right) \tag{11}$$

Then, the negative gradient of the loss function is calculated to fit the approximate value of the current wheel loss function. Equation (12) demonstrates the approximate value of the loss function.

$$f_{ti} = -\frac{\partial\left(y, Z_{t-1}\left(x_i\right)\right)}{\partial Z_{t-1}\left(x_i\right)} \tag{12}$$

Square difference is usually used for approximation$h_t(x)$ as shown in Equation (13).

$$h_t(x) = \underset{h \in H}{argmin} \sum \left(r_{ti} - h(x)\right)^2 \tag{13}$$

In this round, the strong learner is defined as displayed in Equation (14).

$$F_t(x) = h_t(x) - F_{t-1}(x) \tag{14}$$

Therefore, the LightGBM algorithm is integrated with the RF, namely RF-LightGBM, to reduce the calculation cost of the audit process of e-commerce enterprises, improve the calculation efficiency of the model, and obtain high accuracy while maintaining high calculation efficiency.

C. Application of K-means Clustering Algorithm in Big Data Audit Analysis of E-commerce Enterprises

The K-means algorithm can be described as a centroid-based partition technology, that is, the centroid of the cluster$C_i$is used to represent the cluster. When the K-means algorithm is applied to the data analysis of e-commerce enterprises, the centroid of the cluster is defined as the mean value of the points in the cluster. In the clustering process, *n* objects are randomly selected with *k* as the parameter, and each object represents the initial mean value of a cluster. These objects are then divided into *k* clusters. The remaining objects are placed in the neighbor cluster based on their center distance from each cluster, so that the cluster has higher similarity [25,26]. This time, the mean value of each cluster changes, and the average values are recalculated, and the process is repeated until the resulting cluster is as independent as possible.

As Equation (15) indicates, a known set of *n* data samples is defined as Ω.

$$\Omega = \left\{x_i | x_i = \left(x_{i1}, x_{i2}, \cdots, x_{id}\right), i = 1, 2, \cdots, n\right\} \tag{15}$$

In Equation (15), $x_i = \left(x_{i1}, x_{i2}, \cdots, x_{id}\right)$refers to a *d*-dimensional vector, x_{id}refers to the d_{th} identical attributes of the i_{th}data, and *n* represents the sample size. Equation (16) illustrates the clustering center.

$$C = \left\{c_j | c_j = \left(c_{j1}, c_{j2}, \cdots, c_{jd}\right), c = 1, 2, \cdots, k\right\} \tag{16}$$

In Equation (16),$c_j = \left(c_{j1}, c_{j2}, \cdots, c_{jd}\right)$refers to the center point of the j_{th}cluster. There are d attributes in every c_j, and *k* represents the number of clusters.

Equation (17) expresses the Euclidean Distance $dis\left(x_i, c_j\right)$, which is the distance between$x_i$andc_j.

$$dis\left(x_i, c_j\right)=\sqrt{\sum_{l=1}^{d}\left(x_{il}-c_{jl}\right)^2}, i=1,2,\cdots,n; c=1,2,\cdots,k \tag{17}$$

In Equation (17), $x_i=(x_{i1},x_{i2},\cdots,x_{id})$ $c_j=(c_{j1},c_{j2},\cdots,c_{jd})$,$k$ refers to the number of clusters. Equation (18) corresponds to the calculation of the center of the same clustersc_j.

$$c_{jl} = \frac{1}{N(\phi_j)} \sum_{x_i \in \phi_j} x_{il}, l = 1, 2, \cdots, d; c = 1, 2, \cdots, k \tag{18}$$

In Equation (18), $N(\phi_j)$ represents the amount of data in the same cluster. The criterion function is generally defined by the sum of squared errors, which is expressed as Equation (19).

$$E = \sum_{j=1}^{k} \sum_{x_i \in \phi_j} dis(x_i, c_j) \tag{19}$$

In Equation (19), E refers to the total value of the squared error of all data objects in the data set of e-commerce enterprise audit, $x_i$corresponds to the point in the space, k represents the given e-commerce enterprise audit data object, and $c_j$stands for the average value of the center point of the j_{th} cluster class. Fig. 4 displays the algorithm flow of applying the K-means algorithm to the BD of e-commerce enterprises.

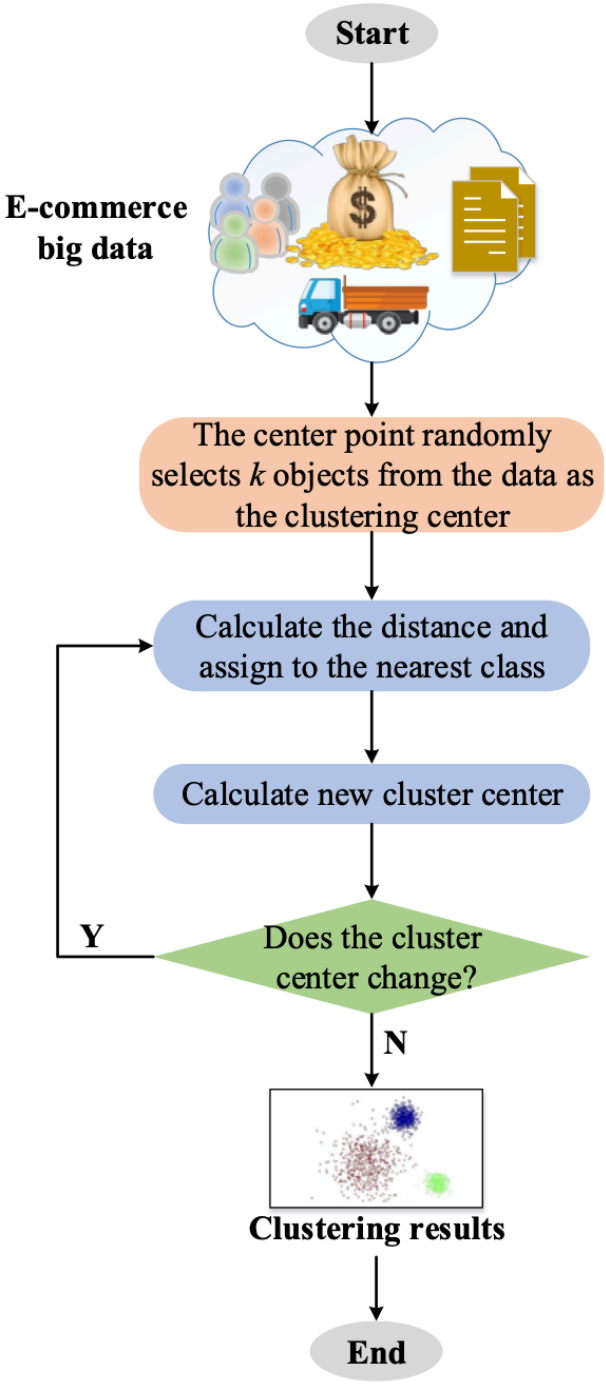

Fig. 4. K-means algorithm applied in the BD of e-commerce enterprises

Fig. 4 demonstrates the specific steps when the K-means algorithm is applied to the BD of e-commerce enterprises'

doi.org/10.32388/G9GoS3.2

audit. Firstly, *k* objects are randomly selected from the BD samples of *n* e-commerce enterprises as the initial clustering centers; secondly, the distance from each sample to each cluster centroid is calculated respectively, and the sample is assigned to the nearest cluster center category; thirdly, after all the samples are allocated, the centers of *k* clusters are recalculated; fourthly, compared with the previously calculated *k* cluster centers, if the cluster center changes, the process returns to the second step, otherwise it proceeds to the fifth step; fifthly, the process stops and the clustering results are output when the centroid does not change.

D. Construction and Analysis of Risk Audit Model for E-Commerce Enterprises Based on K-Means Algorithm Combined with Random Forest

In view of the diversity and complexity of traffic influencing factors in the road network area of smart city construction, the K-means clustering algorithm is introduced here. However, the results of the K-means clustering algorithm basically depend on initial values, and the difference in initial values is the direct reason for the appearance of diverse clustering results. Also, the number of clusters generated, k, must be given in advance. Therefore, the RF algorithm is introduced and improved to classify the BD of risk audit in e-commerce enterprises. Finally, construction is conducted on the risk audit model of e-commerce enterprises based on the K-means algorithm combined with RF-LightGBM. Fig. 5 demonstrates its overall architecture.

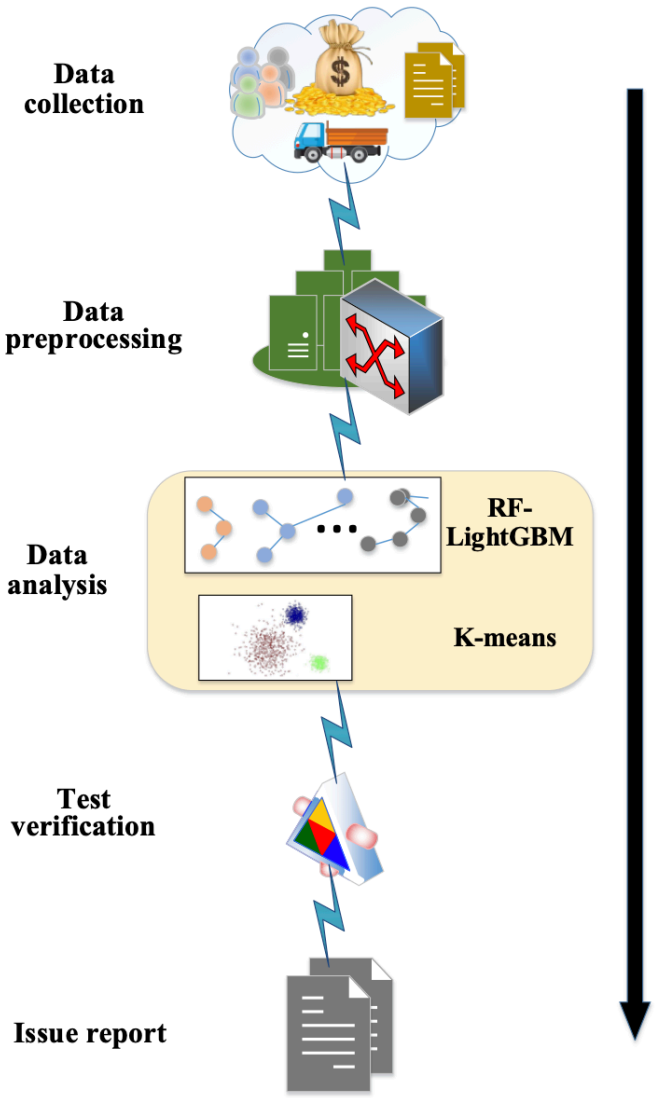

Fig. 5. Risk audit model of e-commerce enterprises based on K-means algorithm combined with RF-LightGBM

In the risk audit model of e-commerce enterprises, the first step is to collect the audit data required by e-commerce enterprises. The collected data does not only include financial data but also contains data that covers the business situation of the audited unit and the specific audit rules and regulations. After complete collection, the problems in the same specific direction are integrated and classified. Secondly, the audited data are extracted and cleaned, such as error data, invalid data, and abnormal data found in the audit process during data extraction. Assuming that data is not processed, it is equivalent to predicting future data with the wrong data, which makes the potential link between the data undetectable and leads to the wrong direction for the later development of the enterprise. The continuous accumulation of error data will make enterprises face a huge crisis, so data cleaning should not be ignored.

In the data analysis stage, the K-means algorithm is combined with the RF. The construction of this model algorithm can not only avoid the sensitivity to the initial value when using the K-means clustering algorithm alone but also prevent different clustering results for different initial values. The number k of the generated clusters must be given in advance, which can also reduce the calculation cost of the audit process of e-commerce enterprises, improve the calculation efficiency of the model, and obtain high accuracy while maintaining high calculation efficiency. A number of n audit warning

doi.org/10.32388/G9GoS3.2

indicators are selected from the shared data center of e-commerce enterprises as the object of feature selection, as $X = \{X_1, X_2, \cdots, X_n\}$,$X_i = \{X_{i1}, X_{i2}, \cdots, X_{in}\}$. X_{in} indicates the n_{th}character of the i_{th}audit warning indicator. Now, m samples are randomly selected from N samples, and Equation (20) corresponds to the cumulative weight equation of audit warning features.

doi.org/10.32388/G9GoS3.2

$$W_j^{i+1} = W_j^i + \frac{diff(j, x, M(x))}{m} - \frac{diff(j, x, H(x))}{m} \tag{20}$$

In Equation (20), *j* refers to the audit warning features, which vary from 1 to *N*; *i* represents the randomly selected samples; *diff*$(\cdot)$means the distance;$M(x)$ stands for the heterogeneous nearest neighbor samples, and $H(x)$ denotes the similar nearest neighbor samples. Fig. 6 demonstrates the steps of the model based on the K-means algorithm combined with RF-LightGBM

doi.org/10.32388/G9GoS3.2

1	**start**
2	Input: dataset *D*, feature set K
3	Output: optimal classification of e-commerce enterprise data audit
4	Calculate the feature importance degree I_i of feature f_i using RF respectively
5	The features are sorted in descending order according to the obtained I_i
6	The LightGBM algorithm is used for evaluation. Backward selection is made for the sorted feature subset to calculate the accuracy a_{tmp} after deleting the feature
7	**For**(f_i) do
8	Calculation a_i// Calculates the accuracy of the current feature subset
9	$C \leftarrow C + f_i$ //delete feature f_i
10	Calculate a_{tmp} // calculate the accuracy after deleting the feature
11	**if** $a_{tmp} > a_i$then
12	$a_{best} = a_{tmp}, C_{best} = C$ // if the new accuracy rate is greater than the old, update feature subset
13	**Else** then
14	$C \leftarrow C + f_i$ // otherwise, the feature just deleted will be recycled
15	K-means calculates the distance between K cluster centers and samples
18	**end** if
19	**end** for
20	**end**

Fig. 6. Steps of the model based on K-means algorithm combined with RF-LightGBM

E. Stimulation

To verify the performance of the risk audit model of e-commerce enterprises based on the K-means algorithm combined with RF-LightGBM constructed here, Matlab software is used to conduct empirical research on the simulation generation of the constructed model. The data set used here comes from the post-desensitization transaction data provided by Jingdong Mall (a Chinese e-commerce platform). Analysis is mainly made on the accurate data and fuzzy data from January 17, 2020, to February 17, 2020. Accurate data consists of customer consumption records, return and exchange interest; fuzzy data includes user browsing records, commodity comparison records, and other information, and the number of all the enterprises' behaviors is more than 100 million. Initially, statistical and visual analysis is made on the original data, which is then pre-processed. Its pretreatment. Afterwards, duplicate values in the data and default values are removed. Finally, the data is divided into a training set and a test set in an 8:2 ratio.

In the simulation analysis, the risk audit model based on the K-means algorithm combined with the RF-LightGBM classification algorithm is compared with the models and algorithms proposed by other scholars in related fields, which mainly refer to RF-LightGBM [27], K-means [28], LightGBM [29], Support Vector Machines (SVM) [30], and Bayesian network (BN) [31], respectively, from the perspectives of classification accuracy, data message delivery rate, leakage rate, packet loss rate, and average delay of data transmission security. Among them, the model constructed here uses the cluster module of sklearn when designing the K-means clustering algorithm. The parameters are set as follows: *k* for the n_cluster classification cluster setting, valuing 2 ~ 6, and the maximum number of iterations defaults to 120 for max_iter. The specific simulation experiment configuration is mainly considered from both hardware and software. In the software, the operating system is Linux 64bit, the Python version is Python 3.6.1, and the development platform is PyCharm; in hardware, the CPU is Intel Core i7-7700 @ 4.2GHz 8-core, the memory is

Kingston DDR4 2400MHz 16G, and the GPU is Nvidia GeForce 1060 8G.

IV. Results and Discussion

A. Comparative Analysis of Classification Accuracy Performance of Each Model and Algorithm

To study the performance of the risk audit model of e-commerce enterprises based on the K-means algorithm combined with RF-LightGBM, the system model constructed here is compared in several aspects with the algorithms put forward by other relevant scholars. The classification accuracy is evaluated using Accuracy, Recall, Precision, and F1 value, and Fig. 7 displays the results. Fig. 8 presents further acceleration ratio analysis of its classification efficiency.

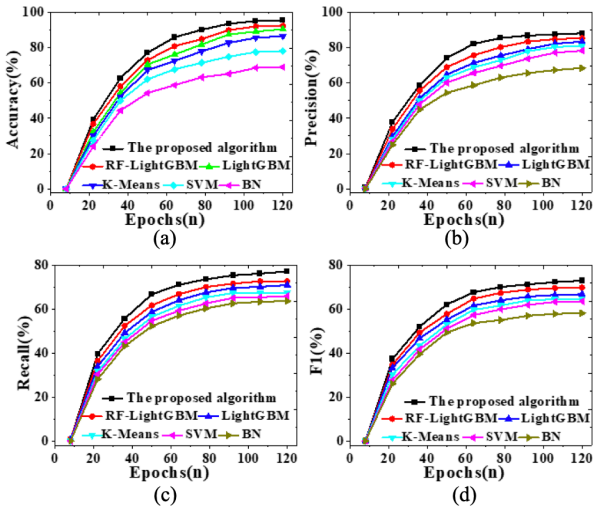

Fig. 7. Curves of the influence of iteration on classification accuracy of different algorithms (a. Accuracy; b. Precision; c. Recall; d. F1 value)

As Fig. 7 indicates, through the comparison of the system model proposed here with other DM algorithms based on Accuracy, Precision, Recall, and F1, results indicate that the recognition accuracy of the model proposed here reaches 95.46%, which is at least 3.06% higher than that of the models and algorithms proposed by other scholars. Further comparison from three angles of Precision, Recall, and F1 suggests that the Precision, Recall, and F1 of the model algorithm are 88.17%, 77.22%, and 73.05%, respectively. Through a comparison between the model and other algorithms, a conclusion can be drawn that the Precision, Recall, and F1 of the model algorithm are higher, at least 2.79% higher than those of other algorithms. Furthermore, through a comparison between the model algorithm proposed here and other algorithms proposed by other relevant scholars, the K-means algorithm combined with the RF-LightGBM algorithm used in the risk audit model of e-commerce enterprises constructed here has better classification accuracy of transaction data of e-commerce enterprises.

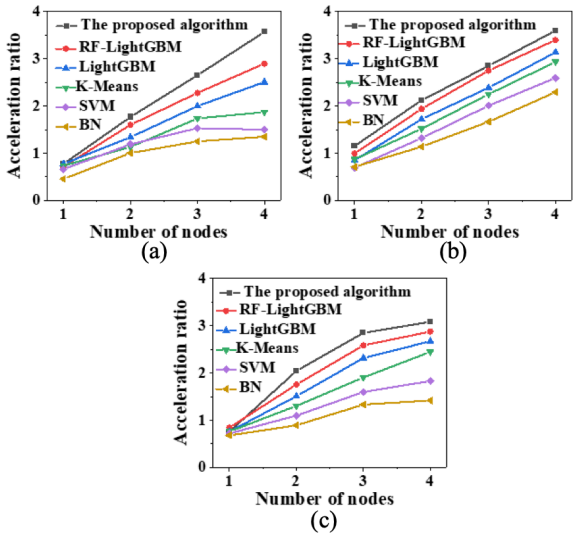

Fig. 8. Curves of comparison of different algorithms' acceleration ratio (a. pre-processing; b. training; c. test)

The acceleration performance of each algorithm is further compared and analyzed, and Fig. 8 illustrates the results. It is found that with the increase in nodes, acceleration is more effective than improving the classification of data blocks, and the degree of parallelism is improved. However, with the increase in nodes, the speedup increases more slowly because the communication between nodes takes up a certain amount of time. Furthermore, it is found that the acceleration ratio of the proposed algorithm is significantly superior to other algorithms, which indicates that the model and the algorithm constructed here can complete the classification of audit data in e-commerce enterprises more quickly.

B. Analysis of Models' Data Transmission Security Performance under Different Algorithms

To study the prediction performance of the model constructed here, analysis is made from the aspects of RF, LightGBM, K-Means, SVM, and BN based on accuracy, precision, recall, and F1 values, respectively. Fig. 9 demonstrates the results of the comparison.

doi.org/10.32388/G9GoS3.2

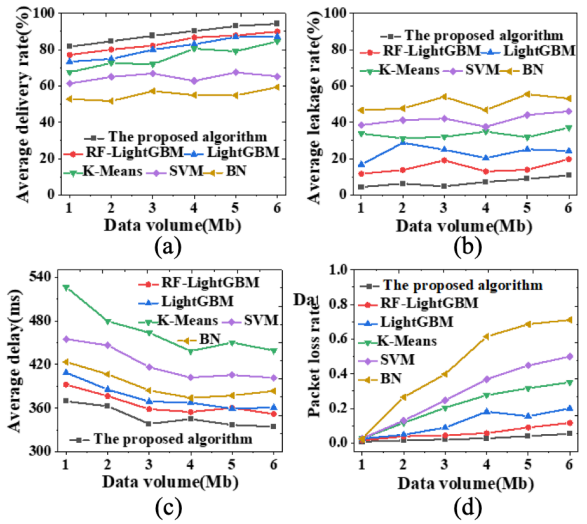

Fig. 9. Comparative analysis of data transmission security of audit data of e-commerce enterprises under different algorithms (a. average delivery rate; b. average leakage rate; c. average delay; d. average loss rate)

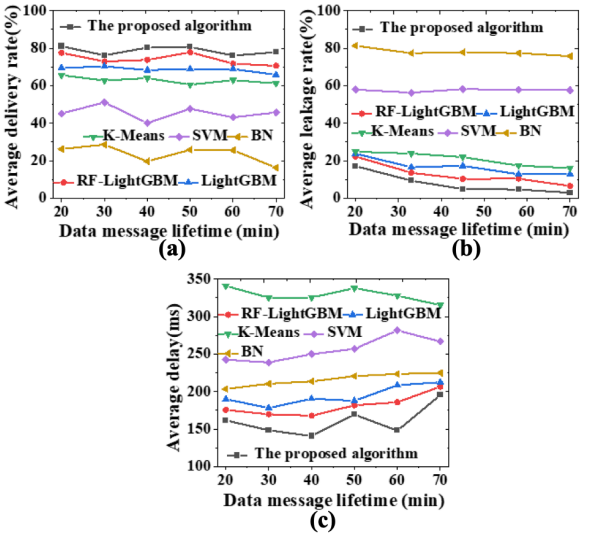

Fig. 10. Comparative analysis of data transmission security of each algorithm under different survival times of audit data messages of e-commerce enterprises (a. average delivery rate; b. average leakage rate; c. average delay)

After further analysis is carried out on each algorithm's data transmission performance, results show that as the amount of transmitted data increases, the mean delivery rate of network audit data shows an upward trend, and the data message delivery rate is not less than 81.54% (Fig. 9 (a)); the average leakage rate of network data has no obvious change, and the data message leakage rate in this study does not exceed 10.83% (Fig. 9b); in terms of average delay, when the transmission of audit data increases, the average delay decreases, and the mean value of the delay of the model algorithm in this study is basically stable at about 344.39 ms (Fig. 9 (c)); in the packet loss rate analysis, it is found that the BN algorithm has a higher packet loss rate, where there may be hidden terminal problems, namely, packet loss. The algorithm's packet loss rate is the lowest, less than 5.29%, which is due to the balanced processing of the transmitted data (Fig. 9 (d)). Therefore, judging by different transmission data, the risk audit model algorithm of e-commerce enterprises based on the K-means algorithm combined with RF-LightGBM constructed here has prominent features in a higher average delivery rate, lower delay, and the lowest average leakage rate. Therefore, it has fantastic performance in data security transmission on the Internet and a lower data transmission risk of the model.

By comparing the network data security transmission performance of each mechanism algorithm under different survival times of audit data messages, the results are shown in Figure 10. It can be found that with the increase in the survival time of data messages, the model algorithm in this study shows a better average delivery rate and a lower message leakage rate, which may be due to the fact that the model algorithm constructed in this study adopts the forwarding strategy of message fragmentation and trusted end users, while other algorithms adopt message fragmentation for the security protection of messages, lacking trust evaluation of forwarding nodes or only encrypting. In terms of the average delay, the algorithm in this study shows that with the increase of the survival time of the risk audit data message, the average delay gradually increases, and the average delay of the model algorithm in this study is basically stable at about 148.04 ms.

V. Conclusion

In the era of e-commerce, the traditional methods of risk audit for traditional enterprises cannot meet the needs in the face of paperless transactions in e-commerce enterprises. The desensitized transaction data of JD.com is taken as an example. In view of the paperless status of electronic payment in the transaction process of e-commerce enterprises, the K-means clustering algorithm is integrated with the RF algorithm in the ML algorithm to construct the risk audit model of e-commerce enterprises

doi.org/10.32388/G9GoS3.2

based on the K-means algorithm and RF-LightGBM. Finally, through simulation, it is found that the classification accuracy of the model algorithm constructed here reaches 95.46%, and the packet loss rate, data message leakage rate, and average delay are lower than those of other model algorithms, which provides experimental support for the safety improvement and risk control of the audit process in e-commerce enterprises. The business data of Jingdong Mall is adopted to establish a risk audit model of e-commerce enterprises, indicating that the model is applicable to e-commerce enterprises. However, there are some shortcomings in current research. For example, if the optimized risk audit classification model is directly applied to the risk audit data of insurance companies and other regions, whether it can ensure such high accuracy remains to be considered. The research is of great significance for the future application of DM in the risk audit of e-commerce enterprises.

References

1. Zhu, L. (2020). Analysis and Research of Enterprise Information System Security Based on e-Commerce. Academic Journal of Computing & Information Science, 3(3),1-9.
2. Jiang, J., & Chen, J. (2021). Framework of Blockchain-Supported E-Commerce Platform for Small and Medium Enterprises. Sustainability, 13(15), 8158.
3. Huang, Y., Chai, Y., Liu, Y., & Shen, J. (2018). Architecture of next-generation e-commerce platform. Tsinghua Science and Technology, 24(1), 18-29.
4. Zhang, F., & Yang, Y. (2021). Trust model simulation of cross border e-commerce based on machine learning and Bayesian network. Journal of Ambient Intelligence and Humanized Computing, 1-11.
5. Abd Hamid, N., Ibrahim, N. A., Ibrahim, N. A., Ariffin, N., Taharin, R., & Jelani, F. A. (2019). Factors affecting tax compliance among Malaysian SMEs in e-commerce business. International Journal of Asian Social Science, 9(1), 74-85.
6. Zhao, J., Lu, Y., Ban, H., & Chen, Y. (2020). E-commerce satisfaction based on synthetic evaluation theory and neural networks. International Journal of Computational Science and Engineering, 22(4), 394-403.
7. Lakshmi, K. N., Neema, N., Muddasir, N. M., & Prashanth, M. V. (2020). Anomaly Detection Techniques in Data Mining—A Review. Inventive Communication and Computational Technologies, 799-804.
8. Shad, M. K., Lai, F. W., Fatt, C. L., Klemeš, J. J., & Bokhari, A. (2019). Integrating sustainability reporting into enterprise risk management and its relationship with business performance: A conceptual framework. Journal of Cleaner production, 208, 415-425.
9. Hanggraeni, D., Ślusarczyk, B., Sulung, L. A. K., & Subroto, A. (2019). The impact of internal, external and enterprise risk management on the performance of micro, small and medium enterprises. Sustainability, 11(7), 2172.
10. Cheng, S., Jianfu, S., Alrasheedi, M., Saeidi, P., Mishra, A. R., & Rani, P. (2021). A New Extended VIKOR Approach Using q-Rung Orthopair Fuzzy Sets for Sustainable Enterprise Risk Management Assessment in Manufacturing Small and Medium-Sized Enterprises. International Journal of Fuzzy Systems, 1-23.
11. Yang, B., & Liao, Y. M. (2021). Research on enterprise risk knowledge graph based on multi-source data fusion. Neural Computing and Applications, 1-14.
12. Ping, P., Qin, W., Xu, Y., Miyajima, C., & Takeda, K. (2019). Impact of driver behavior on fuel consumption: Classification, evaluation and prediction using machine learning. IEEE Access, 7, 78515-78532.
13. Xu, Y., Du, B., Zhang, L., Cerra, D., Pato, M., Carmona, E., ... & Le Saux, B. (2019). Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12(6), 1709-1724.
14. Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2019). Flight delay prediction based on aviation big data and machine learning. IEEE Transactions on Vehicular Technology, 69(1), 140-150.
15. Lv, Z., Qiao, L., & Singh, A. K. (2020). Advanced machine learning on cognitive computing for human behavior analysis. IEEE Transactions on Computational Social Systems,8(5), 1194 - 1202.
16. Kanagaraj, R., Rajkumar, N., & Srinivasan, K. (2021). Multiclass normalized clustering and classification model for electricity consumption data analysis in machine learning techniques. Journal of Ambient Intelligence and Humanized Computing, 12(5), 5093-5103.
17. Amani, F. A., & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. International Journal of Accounting Information Systems, 24, 32-58.
18. Nawaiseh, A. K., Abbod, M. F., & Itagaki, T. (2020). Financial Statement Audit using Support Vector Machines, Artificial Neural Networks and K-Nearest Neighbor: An Empirical Study of UK and Ireland. International Journal of Simulation--Systems, Science & Technology, 21(2),1-8.

19. Didimo, W., Grilli, L., Liotta, G., Menconi, L., Montecchiani, F., & Pagliuca, D. (2020). Combining network visualization and data mining for tax risk assessment. IEEE Access, 8, 16073-16086.
20. Ashraf, N., Ahmad, W., & Ashraf, R. (2018). A comparative study of data mining algorithms for high detection rate in intrusion detection system. Annals of Emerging Technologies in Computing (AETiC), Print ISSN, 2516-0281.
21. Eremeev, M. A., & Zakharchuk, I. I. (2020). A Procedure for Improving Information System Audit Quality by Enhancing Cyberthreat Simulation in Practice. Automatic Control and Computer Sciences, 54(8), 854-859.
22. Zuo, X., Chen, Z., Dong, L., Chang, J., & Hou, B. (2020). Power information network intrusion detection based on data mining algorithm. The Journal of Supercomputing, 76(7), 5521-5539.
23. Liou, J. J., Chang, M. H., Lo, H. W., & Hsu, M. H. (2021). Application of an MCDM model with data mining techniques for green supplier evaluation and selection. Applied Soft Computing, 107534.
24. Wu, M., & Moon, Y. (2018). DACDI (Define, Audit, Correlate, Disclose, and Improve) framework to address cyber-manufacturing attacks and intrusions. Manufacturing Letters, 15, 155-159.
25. Grover, D., Bauhoff, S., & Friedman, J. (2019). Using supervised learning to select audit targets in performance-based financing in health: An example from Zambia. PloS one, 14(1), e0211262.
26. Li, Y., Yan, C., Liu, W., & Li, M. (2018). A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification. Applied Soft Computing, 70, 1000-1009.
27. Pal, A. K., Rawal, P., Ruwala, R., & Patel, V. (2019). Generic disease prediction using symptoms with supervised machine learning. Int. J Sci. Res. Comput. Sci. Eng. Inf. Technol, 5(2), 1082-1086.
28. Yang, J. C., Chuang, H. C., & Kuan, C. M. (2020). Double machine learning with gradient boosting and its application to the Big N audit quality effect. Journal of Econometrics, 216(1), 268-283.
29. Rani, L. N., Defit, S., & Muhammad, L. J. (2021). Determination of Student Subjects in Higher Education Using Hybrid Data Mining Method with the K-Means Algorithm and FP Growth. International Journal of Artificial Intelligence Research, 5(1), 91-101.
30. Li, Z., & Liu, Y. (2017). A differential game-theoretic model of auditing for data storage in cloud computing. International Journal of Computational Science and Engineering, 14(4), 341-348.
31. Alekseeva, I. V., & Mosentseva, V. A. (2020). Methodological Approaches to Establishing the System of Internal Control in Agricultural Companies. Accounting. Analysis. Auditing, 7(5), 69-79.

1.

Declarations

Funding: No specific funding was received for this work.
Potential competing interests: No potential competing interests to declare.