Research on Intellectual Dichotomiser 3 Decision Tree Algorithm Model for Financial Analysis of Colleges and Universities

Sujuan Guo
Smart Finance Industry College, Shandong Institute of Commerce and Technology, Jinan, 250103, China

***Abstract*—The rapid development of college information construction has promoted the processing and analysis of a large number of data in the college system. Decision tree algorithm is often used in the field of financial data analysis, but it has a bias in the selection of attributes. Aiming at the defects of the decision tree algorithm in attribute selection, ID3 algorithm in the decision tree algorithm is selected for weighted improvement, and it is optimized based on Synthetic Minority Oversampling Technique (SMOTE) algorithm and Bagging algorithm to balance the positive and negative data of its training samples, thus obtaining the DSB-ID3 financial analysis model. Using this model to analyze the financial data of a university, its G value and F value are both about 0.78, the recognition accuracy rate for normal samples is 0.7345, and the total recognition accuracy rate is 0.7893, which are the highest among the four models. Compared with other models, model designed in this study has significantly improved classification performance, and its distribution is the most centralized, showing superior stability. The experimental results show that the classification effect of model designed in this study is the best, and it shows superior accuracy and stability in the analysis of financial data. Its superior classification performance shows the potential of decision tree algorithm optimization and the feasibility and necessity of improving it. From the experimental data, it can be seen that the service life and parameters of the model designed in this study are obviously better than those commonly used in the financial analysis industry of colleges and universities. It can be seen from the overall analysis that this model provides a practical reference for the application of decision tree optimization in college financial analysis, and greatly improves the accuracy of financial system data analysis.**

Keywords*—Financial analysis; ID3; Decision tree; model; colleges and universities; machine learning

I. INTRODUCTION

With the gradual deepening of higher education reform and the introduction of higher education into the market as an industry, the financial management of colleges and universities has also changed. The financial analysis and management work in colleges and universities has gradually shifted from accounting to analytical, but the existing financial analysis and management system cannot meet the requirements of precision and real-time [1-2]. With the rapid development of information technology, the scale of financial informatization in colleges and universities is also expanding. The college information system can easily collect relevant financial data [3-4]. However, massive data brings management difficulties to data managers. Lack of data mining means and tools will result in a huge waste of data resources, and financial analysis cannot be carried out in a normal and orderly manner. Therefore, computer technology is integrated into financial analysis to effectively analyze and manage data, so as to improve the application level of financial data, optimize the financial structure of the school, and promote the healthy development of the financial management system [5-6]. Decision tree algorithm is often used in the field of data analysis. Research will also establish a university financial analysis model based on decision tree. However, the decision tree has the defect of attribute selection, so this research will select ID3 (intelligent dichotomy 3) algorithm in the decision tree to improve, and based on DSRR (Differential Sampling Rate Resampling, DSRR) technology, SMOTE (Synthetic Minor Over Sampling), so as to improve its noise resistance and balance its attribute selection, which is also the main contribution of this research.

II. RELATED WORK

Scholars such as Febriantono MA proposed a C5.0 deciMany scholars have done a lot of research on decision tree algorithm in financial analysis and other fields. Luo X et al. proposed an initialization method of dendrite neuron model (DNM) based on decision tree to trim those neurons that contribute less to the network output, aiming at the problem of insufficient calculation accuracy of decision tree model. This method can reduce the number of dendrites in DNM and improve the training efficiency, and can select appropriate initialization weight and neuron threshold, thus enhancing the classification accuracy of the algorithm. Experiments show that this method is significantly superior to the original DNM model with the lowest complexity and the fastest training speed, without loss of accuracy [7]. Podhorska I et al. believed that the non-linear data calculation ability of the decision tree model was insufficient, so they combined BP neural network with ID3 decision tree model to build an improved decision tree model, and applied it to predict financial distress. They also established a prediction model for enterprise financial distress, mined and analyzed enterprise financial data, and used it to predict enterprise financial development. The test results showed that the financial data prediction accuracy of the improved algorithm is significantly higher than that of the decision tree model before improvement [8]. Sion tree algorithm based on cost sensitivity, established a model, learned to use the Metacost Method (MM) to get the minimum cost model, and used it to solve multi-class imbalance problems. The performance of the algorithm is better than the C4.5 and ID3 algorithms [9]. Maulana MF et al. used two model classifications for the evaluation of students' learning time, namely Logistic Model Tree (LMT) and Decision Tree J48 Algorithm (DT J48), using the model to predict the effect of students' learning time and find out the influencing factors. The prediction accuracy of LMT is 71% higher than that of DT J48 [10]. Xia Y and other scholars proposed a SurvXGBoost model based on the Gradient Boosting Decision Tree (GBDT) method, and applied it to the consumer loan dataset. The performance is excellent [11].

Researchers such as Balta M proposed a three-level fuzzy decision tree model for Vehicle Ad-Hoc Networks (VANET) system based on Software Defined Networks (SDN). The physical structure of urban intersections often lead to traffic jams, delays, and accidents. Previous studies usually choose the VANET architecture, which can communicate between vehicles and road edge devices. SDN has also become a common research method in this domain, it can solve the performance, programming, scalability and security problems in traditional networks. The proposed model utilizes the architectural recommendations of both SDN and VANET network paradigms to optimize traffic management problems based on three-level fuzzy decision tree algorithm classification [12]. Qiu T et al. proposed a Directed Edge Weight Prediction (DEWP) model based on decision tree integration. It extends the local similarity index to the Directed Weighted Network (DWN), and extracts the similarity index to construct a mixed regression model [13]. In order to explain the underpricing event of Initial Public Offering (IPO) during the epidemic, Keuangan J and other scholars built a decision tree algorithm model and tested the latest 45 action data using more than 100 previous action data. The model can interpret IPO performance according to specific classification scope. [14]. Researchers such as L Xue proposed a Privacy-Preserving Decision Tree Classification (PDTC). First, customize an encryption primitive and a secret sharing technology to design a new secure bilateral comparison protocol, in which the digital inputs of each party are compared privately. Then, based on the protocol, PDTC is constructed using decision tree structure. The customer's input and the service provider's model parameters are hidden from the counterparty, and the classification results are only displayed to the customer. The results show that PDTC achieves ideal safety [15]. Ziweritin S and other scholars established the K-nearest neighbor and decision tree model to forecast the numeric value in the one pound table under a given interest rate and life. Cross validation was performed using the R-squared test to detect fit and summarize model function on the test data set. The model is robust and competitive, and has reference value for financial analysis and its application [16].

In addition, in terms of college financial management, Natawibawa IWY team applied multiple linear regression method to analyze the financial management system of state universities, and the research results show that the availability of financial reports has a significant impact on the transparency of college financial management [17]. The Min Du team has carried out informatization reform on the traditional financial management model of colleges and universities, and the research results show that with the help of informatization, the financial management level of colleges and universities has been effectively improved [18]. The Na Sun team started from the financial position i and staffing of colleges and universities, analyzed the current situation of colleges and universities emphasizing accounting over management, and proposed strategic management and post balance strategy [19]. It can be seen that in the application of decision tree algorithm, there have been some precedents in applying decision tree algorithm to the economic and financial fields. At the same time, in the financial management of colleges and universities, there is still a lack of appropriate information technology to provide a technical basis for the development of financial management of colleges and universities. Therefore, in order to avoid the waste of financial data resources, the research provides data mining means and tools so that the financial analysis work of colleges and universities can be carried out normally and orderly. The research is oriented to college financial analysis, and the ID3 algorithm is weighted. Based on the DSRR technology, SMOTE algorithm and Bagging algorithm, ID3 is optimized, and the DSB-ID3 model is established to overcome the defect that the ID3 algorithm tends to select attributes with more values but not necessarily optimal. This technology can monitor the financial situation of colleges and universities in real time from a dynamic perspective, carry out meticulous financial management according to the financial characteristics of colleges and universities, and improve the efficiency and quality of financial management of colleges and universities.

III. ID3 Decision Tree Optimization Algorithm Model for University Financial Analysis

A. *Improved ID3 Algorithm based on Attribute Selection Weighting*

The ID3 algorithm is a decision tree algorithm based on information entropy. It introduces information theory, uses information entropy as a criterion for selecting test attributes, divides the training set, constructs a decision tree to predict according to the test attributes, and divides the instance space [20-21]. To build a decision tree, first measure the attributes of each node using the maximum information gain from top to bottom, and obtain the corresponding attribute values, which are used to segment the target object. This step is repeated until all objects in the node are judged to be of the same type according to the classification criteria. The features with more attribute values in the decision tree have a greater impact on the calculation of information gain, however, features with more values are uncertain optimal attributes, and the algorithm has poor noise resistance, making it difficult to control the positive and negative examples in data set, so the research improves the algorithm by weighting. The main principle of the information theory introduced by the algorithm is to regard the communication process as a process of transmitting information in an environment of random interference. In order to obtain mutual information, the posterior entropy is first calculated. When the channel receiver receives the output symbol $V_j = \{V_1, V_2, \cdots V_m\}$, the input symbol is calculated.

U The information measure of $P\left(U\left|V_j\right.\right)$ is as formula (1).

$$P\left(U\left|V_j\right.\right)=\sum_j P\left(V_j\right)\sum_j P\left(U\left|V_j\right.\right)\lg\frac{1}{P\left(U\left|V_j\right.\right)} \tag{1}$$

Before receiving the symbol set , V the average uncertainty of the $D\left(U\right)$ input symbol set is expressed as U , after receiving the symbol set , V the average uncertainty of the $D\left(U\left|V\right.\right)$ input symbol set is expressed as U , the relationship between the two is shown in Eq.

$$I\left(U,V\right)=D\left(U\right)-D\left(U\left|V\right.\right) \tag{2}$$

In formula (2), $I\left(U,V\right)$ is U the average mutual information between and , which represents V the amount of information V about the obtained after receiving the symbol set U . In order to avoid the ID3 algorithm to select attributes with many values but not necessarily optimal as test attributes, the research introduces risk weights. Given the user's risk weights for uncertain knowledge $\max\left(-a_1,-a_2,\cdots,-a_i,\cdots-a_n\right)\le b\le 0$, a_i it represents the conditional probability of attributes, which b is a dynamic variable whose value Depends on the conditional probability of the attribute obtained each loop. If there are several conditional probabilities in the calculation process $a_1,a_i,\cdots,a_n$, it is $-b$ equal to the minimum value among them. Introducing risk weights into the algorithm, the calculation of conditional entropy is as formula (3).

$$D_b\left(U\left|V\right.\right)=\sum_{j=1}^{n}\left(P\left(V_j\right)+b\right)\sum_{i=1}^{n}P\left(U_i\left|V_j\right.\right)\lg\frac{1}{P\left(U_i\left|V_j\right.\right)} \tag{3}$$

The corresponding mutual information calculation is shown in formula (4).

$$I_b\left(U,V\right)=D\left(U\right)-D_b\left(U\left|V\right.\right) \tag{4}$$

Equations (3) and (4) are used as criteria for selecting test attributes when constructing a decision tree. The research uses the AHP method to assign weights, compares and judges the importance of indicators at the same level, and constructs a judgment matrix to represent the judgment value of their relative importance. In the measurement process, the scale of the ninth is leaded into form a judgment matrix G . Each element in the g_{ij} matrix represents G the comparison value of the relative importance of the g_{ij} row index C_i to each column index, indicating the relative importance of C_j a certain index and another index, and its scale is shown in Table I.

TABLE I. RELATIVE IMPORTANCE BETWEEN TWO INDICATORS

Ratio of index 1 to index 2	extremely important	Important	Slightly important	Identical	Slightly unimportant	Unimportant	Extremely unimportant
Indicator 1 Evaluation	7	5	3	1	1/3	1/5	1/7
Intermediate value	Take 6, 4, 2, 1/2, 1/4, 1/6 as the intermediate value						

The principle of the AHP method used in the research is to judge the matrix, obtain the matrix sorting vector according to the sorting rules, and use the obtained results to calculate the weight coefficients of each index. Find the product of the elements of each row of the judgment matrix as in formula (5).

$$P_i=\prod_{j=1}^{n}c_{ij},i=1,2,\cdots,n \tag{5}$$

Calculate the square root value of each row P_i as n formula (6).

$$\overline{p}_i=\sqrt[n]{P_i},i=1,2,\cdots,n \tag{6}$$

Formula (6) is n the order of the matrix. The vector is normalized as in formula (7).

$$p_i=\frac{\overline{p}_i}{\sum_{j=1}^{n}\overline{p}_j} \tag{7}$$

p_j is the weight coefficient value of each indicator of demand. The maximum eigenroot of the judgment matrix is calculated as in Equation (8).

$$\lambda_{\max}=\sum_{i=1}^{n}\frac{\left(Gw\right)_i}{nw_i} \tag{8}$$

Compared with other methods for determining the weight coefficient of indicators, the AHP method can pass the consistency test and maintain the consistency of thinking logic. When judging the importance of indicators, when there are more than 3 indicators, the coordination among the judgments is the logical consistency of thinking. Let the maximum eigenroot of the $\lambda_{\max}$ judgment matrix be G , when the matrix G passes the consistency, $\lambda_{\max} = n$, otherwise $\lambda_{\max} \neq n$. To test G the consistency, the consistency index of the judgment matrix is established as formula (9).

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{9}$$

When the judgment matrix passes the consistency check, $CI = 0$. The average randomness index of the judgment matrix is introduced RI to measure whether the matrix passes the consistency test. For the judgment matrix whose order is between 1 and 9, its RI values are shown in Table II.

TABLE II. AVERAGE RANDOM INDEX RI OF JUDGMENT MATRIX

Order	1	2	3	4	5	6	7
RI	0	0	0.65	0.98	1.15	1.34	1.47

The consistency ratio of the judgment matrix is CR shown in formula (10).

$$CR = \frac{CI}{RI} \tag{10}$$

$CR < 0.1$ At that time, the matrix is consistent, and the relative importance calculated using this matrix is reasonable. When this condition is not met, the matrix needs to be modified until it passes the consistency check. The comprehensive formula of multiple evaluation opinions is shown in formula (11).

$$Y_j = \frac{\sum_{i=1}^{N}\left(W_{ij} \times S_{ij}\right)}{\sum_{i=1}^{N} S_{ij}} \tag{11}$$

In formula (11), N represents the number of evaluation opinions, generally $N \geq 5$. Y_j Indicates N the calculation weight value of the W_{ij} evaluation opinion for the first indicator, indicates i the j calculation j weight value of the evaluation opinion for the first indicator, and indicates the familiarity coefficient of the S_{ij} evaluation opinion i to the first j indicator, generally a dynamic value between 0 and 4, and the evaluation The familiarity of the opinion i to the first j indicator is proportional, and the same evaluation opinion has different familiarity coefficients for different evaluation indicators. After the weights are set, the evaluation results are dynamically generated by expert evaluation, so that the results can change with the environmental conditions.

B. DSB-ID3 Model Optimized by Three Algorithms

When the ID3 algorithm is used alone, there is a defect in the classifier itself. When selecting the relevant classification attributes, it will not select the optimal attributes, but select attributes with more values [22-23]. On the basis of the weighted improvement of the algorithm, the random oversampling technology (Over Sample, OS) is used to optimize the algorithm, so that the number of samples in the two data sets is basically equal, so that the sample balance can be better obtained. The classification performance can reduce the deviation of sample attributes to a certain extent. The specific process of the OS-ID3 model is shown in Fig. 1.

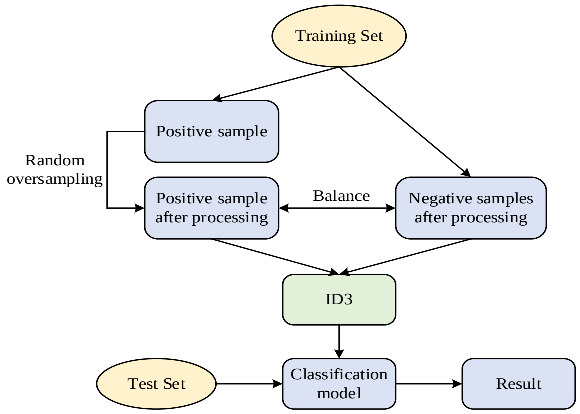


Fig. 1. Specific process of OS-ID3 model

The algorithm has a simple structure and fast calculation speed, which changes the state of imbalance between classes and makes up for the classification defects of the ID3 algorithm to a certain extent. Overfitting phenomenon. The algorithm is optimized using the Over-Under Sample (OUS) technique. When processing samples, the oversampling technique is used to add the quantity of positive samples, and the undersampling technique is used to reduce the quantity of negative samples, so as to balance the two types of samples. The specific process of the OUS-ID3 model is shown in Fig. 2.

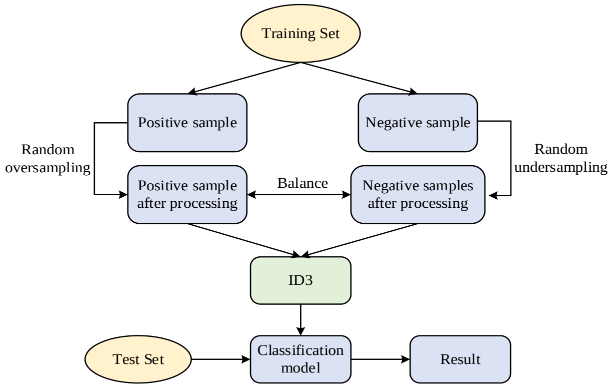


Fig. 2. Specific process of OUS-ID3 model

The structure and steps of OUS-ID3 algorithm are simple, and the calculation speed is fast. To some extent, it avoids the problem that too many positive samples lead to over fitting of the classifier. However, when the quantity of negative samples decreases, some important information will be ignored and the function of the classifier will be degraded. In view of the shortcomings of simple sampling technology, DSRR technology gives different sampling ratios to different classifiers according to the algorithm set internally in the sampling process, so that the number of sample data trained by each classifier is different. Statistically calculate the number of positive samples P_S and negative samples N_S, and calculate the difference between the two types of samples as shown in formula (12).

$$Value = N_S - P_S \tag{12}$$

Assuming that the number of classifiers is M , the sampling coefficient is calculated as shown in formula (13).

$$Coefficient = \frac{Value}{M} \tag{13}$$

Assuming i the sampling rate R of the the classifier, the number of positive samples of the the classifier is calculated i as shown in formula (14).

$$P_{S_i} = P_S + Coefficient \times R_i \tag{14}$$

In formula (14) $1 \le i \le M$. Aiming at the defects of random sampling, the SMOTE algorithm can effectively improve it. The principle of this algorithm is to assume that two positive samples close to each other are also positive samples between them. If the sampling rate is α , each positive class sample point is x_i, its positive class neighbor point is $y_{ij}\left(j=1,2,\cdots,l\right)$, $l = x * \alpha$ is the number of positive class neighbor points, and find all positive class neighbor points. A new negative class sample is generated q_j as shown in equation (15).

$$q_j = x_i + rand\left(0,1\right) \times \left(y_{ij} - x_i\right) \tag{15}$$

formula (15), it means that the $rand\left(0,1\right)$ value is randomly selected q_i in the interval , which will be $\left(0,1\right)$ mixed with the original positive sample to form a minority class sample. According to the difference in the quantity of samples, the balance rate of the two types of samples can be obtained. The sampling principle of the algorithm is shown in Fig. 3.

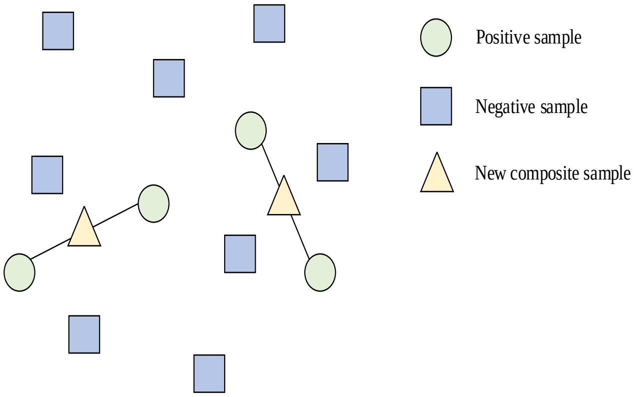


Fig. 3. Sampling principle of SMOTE algorithm

On the premise of retaining the original information and distribution of the minority class, the SMOTE algorithm constructs new samples without repetition, and can overcome the shortcomings of simple oversampling while completing the same task. The Bagging algorithm draws on the replacement sampling method in probability theory. The principle is that after each extraction, the extracted samples are put back into the original data set, so the sample sets extracted each time are the same size, the extracted samples are trained to get the same classifier with different parameters, and the prediction results are obtained. When building a model, small changes in the training samples will lead to significant changes in the classification model, resulting in large differences in the classification results. By training the ID3 algorithm through the Bagging algorithm, multiple different classifiers will be obtained. Different classifiers get different results because the built-in parameters of the ID3 algorithm have changed, resulting in changes in the internal operation of the model and changes in the results obtained. But for multiple classifiers, the majority weighted voting method does not affect the final result due to the result of a single classifier, which strengthens the generalization performance of the ensemble classifier. Assume the training set $T = \left\{\left(x_i, y_i\right), \left(x_2, y_2\right), \cdots, \left(x_n, y_n\right)\right\}$, use the single-classification algorithm to perform T operations on the training set z , extract different training subsets from it, and use it as a sample for each training to obtain T z a classifier with different results. The test set is classified, the z results are weighted and voted, and the class that appears multiple times in the classification result is used as the final result.

Randomly select T' a subset of training samples, put them into a given single-classification algorithm for training, and obtain a model h_t, and loop through the steps z to obtain a set $\{h_1, h_2, \cdots, h_t\}$. Each model h_t gets a classifier, and the unknown samples are classified to get a result, and the category to which they belong is judged according to the principle of getting the most votes. Bagging algorithm overcomes the instability of classifier classification prediction to a large extent, and can reduce a lot of time and cost in the operation process.

The ID3 algorithm model based on DSRR, SMOTE and Bagging uses DSRR technology, SMOTE sampling method and Bagging algorithm combined with ID3 decision tree algorithm to form an integrated classification model based on ID3 classifier. The DSRR technique and the SMOTE sampling method are effectively combined to generate minority class samples according to the sampling rate. The SMOTE sampling method has superior sampling characteristics, and no overfitting occurs during the sampling process, which increases the number of minority class samples and narrows the gap with the majority class samples. On the basis of retaining the original information of the samples, DSRR technology uses the changed sampling rate to establish a sub-classifier with the difference in the number of samples of the majority class and the minority class. The Bagging algorithm combines DSRR technology to generate majority class samples that match the minority class samples. The combination of the two highlights the technical superiority. After removing the weak classifiers, they are integrated into DSB-ID3 (DSRR, SMOTE and Bagging ID3) combined classifier to improve the accuracy of the algorithm based on unbalanced characteristics. The DSB-ID3 model is shown in Fig. 4.

In Fig. 4, the model determines the sampling rate of two samples through DSRR technology, and uses the SMOTE algorithm in the training set to generate several non-repetitive samples according to the determined sampling rate. From the generated samples, DSRR technology is used to extract n samples with different numbers $[X_1, X_2, \cdots X_n]$, and they are combined with the samples extracted by the Bagging algorithm according to the sampling rate X_f to form n new sample training sets. If the sampling rate is the same, the quantity of two samples is the same, and if the sampling rate is different, the number of two samples is different. The ID3 algorithm is used to train n sample training sets, and n ID3 sub-classifiers with different numbers are obtained $[K_1, K_2, \cdots, K_n]$. According to the training results of each sub-classifier, extract 20 classifiers with the best classification effect from the results, obtain $20n$ sub-classifiers, calculate the mean of the results obtained by the $20n$ sub-classifiers, and compare the results of the sub-classifiers with the obtained mean. Compare, if it is greater than the mean, keep the sub-classifier, otherwise remove the sub-classifier.

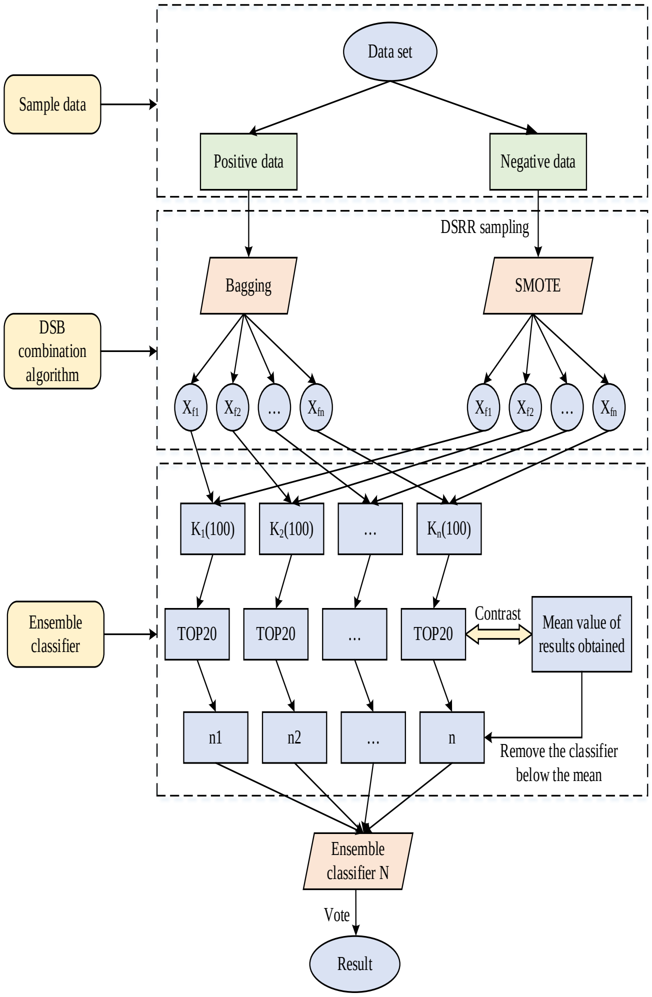
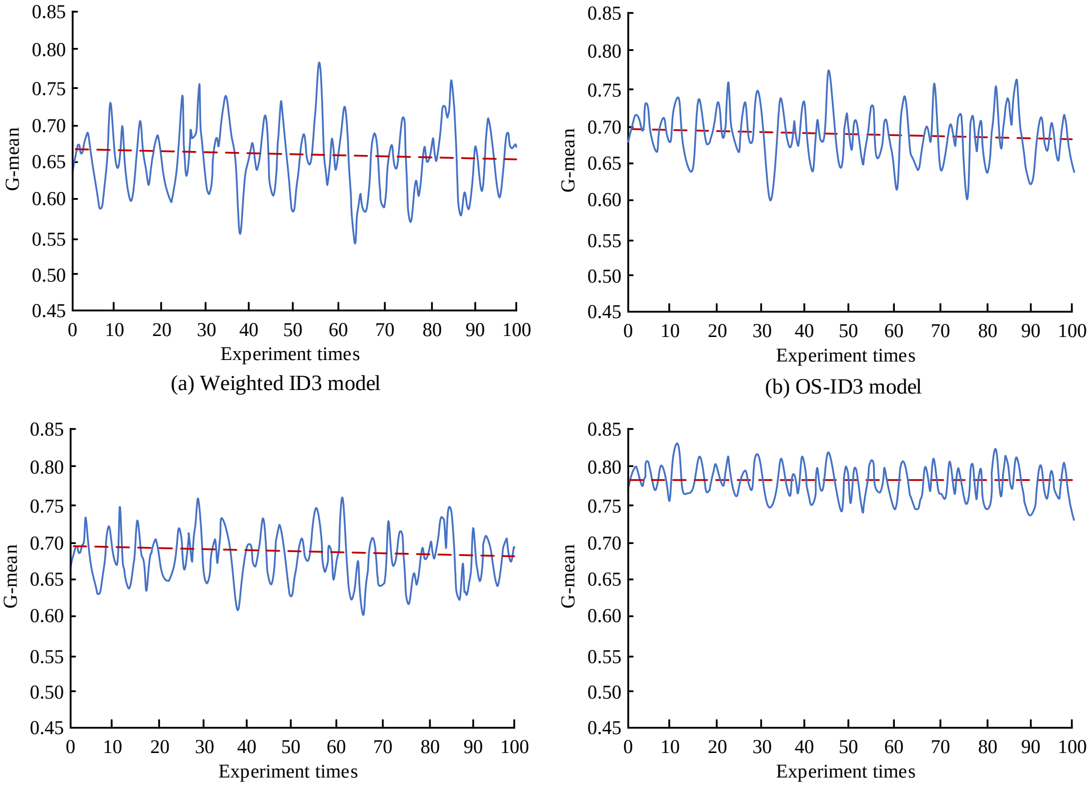


Fig. 4. DSB-ID3 model

The classifiers with different numbers are obtained according to different sampling rates $(n_1, n_2, \cdots, n)$, the remaining classifiers are integrated to obtain the combined classifier N, and different weights are given according to the quantity of remaining classifiers under different sampling rates, and finally the results are obtained by voting. The DSB-ID3 model overcomes the shortcomings of simple sampling technology, and assigns different sampling rates to different classifiers according to the internal algorithm, which makes the number of sample data of the classifiers different. The model also takes advantage of SMOTE sampling, increases the scale of sample data and avoids repeated results, and uses Bagging technology to significantly improve stability. The obtained negative samples are independent of each other and run fast.

IV. Experimental Analysis of DSB-ID3 Model for Financial Analysis of Colleges and Universities

The experimental data comes from the financial analysis report of 2021 of a university, which includes the daily operating expenses and income of the university, the current assets and liabilities of the university, and the change information of the cash flow of the university. This research mainly selects the student payment information in this report for experimental analysis. Student payment information includes student basic information data, student payment information data, receivables information data and fee reduction and exemption information data. The dimensions of defining information include the dimension of student information, the dimension of academic year and semester, the dimension of payable fees, and the dimension of reduced fees. The study used G-mean and F-value as the evaluation index of classification effect. The calculation of the G value is measured by the geometric mean of the classification accuracy of positive and negative samples. In order to reach the maximum value, the classification accuracy of positive and negative samples needs to be larger. The F-value is a combination of precision and recall, where a parameter is introduced to ω reflect the relative importance of precision and recall. The experiment is realized by Matlab and decision tree tools. The DSB-ID3 model and other models are run 100 times in the same environment, and each training set and test set are randomly divided and samples are randomly selected. The G value results of each model are shown in Fig. 5.



(a) Weighted ID3 model

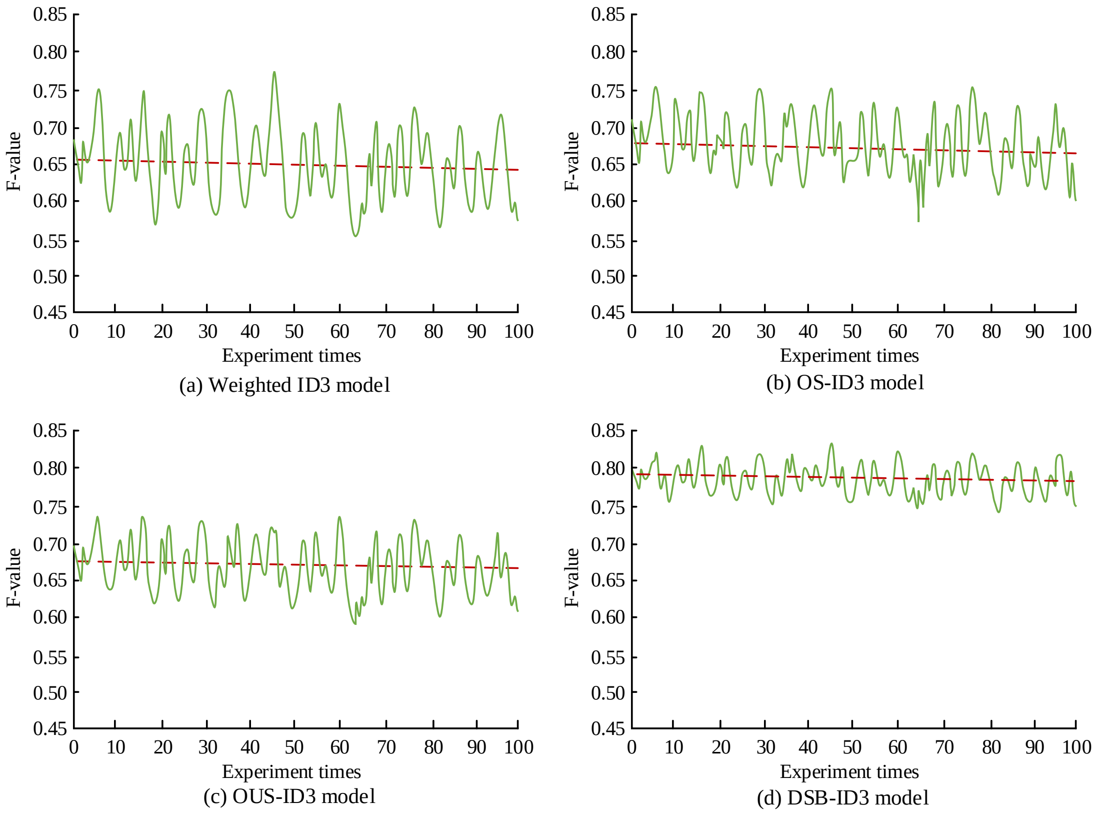
(b) OS-ID3 model

(c) OUS-ID3 model

(d) DSB-ID3 model

Fig. 5. G-mean results of each model

As shown in Fig. 5, with the increase of experimental time, the G value of each model fluctuates repeatedly, but the G value of Weighted ID3 model, OS-ID3 model, and OUS-ID3 model, except for DSB-ID3 model, shows a general downward trend, and their linear regression slopes are 1.74×10^{-4}、1.32×10^{-4} and 1.25×10^{-4} However, the G-value curve of DSB-ID3 model has no overall downward trend. In Fig. 5, the G value of the weighted ID3 model is around 0.66, which is the lowest among all models, and its distribution is also the most dispersed. The G value of the OS-ID3 model is about 0.69, which is improved compared to the weighted ID3 model, but its distribution is also relatively scattered, which is not much different from the weighted ID3 model. The line graph of the OUS-ID3 model is more concentrated than the previous two models, but its G value is basically not improved. The G value of the DSB-ID3 model is about 0.78, which is a significant improvement compared to other models, and the polyline distribution is concentrated. Among the four models, the weighted ID3 model has the worst classification effect, the OS-ID3 model and the OUS-ID3 model have similar classification effects, and the DSB-ID3 model has the best classification effect, and the classification results are accurate and stable. The F value results of each model are shown in Fig. 6.



(a) Weighted ID3 model

(b) OS-ID3 model

(c) OUS-ID3 model

(d) DSB-ID3 model

Fig. 6. F-value results of each model

As shown in Fig. 6, the F value of each model fluctuates repeatedly with the increase of experimental time, and the change rule of each model is generally similar to that in Fig. 5. Specifically, the F values of Weighted ID3 model, OS-ID3 model, OUS-ID3 model and DSB-ID3 model show a downward trend in general, and their linear regression slopes are 0.82 × 10^{-4}、0.94 × 10^{-4} and 0.71 × 10-^{4}、0.35 × 10^{-4}, it can be seen that the overall descending speed of DSB-ID3 model is the slowest. In Fig. 6, the F value of the weighted ID3 algorithm is about 0.65, and its distribution is as scattered as the G value. Due to the improvement of the balance of sampling rate, the OS-ID3 model has a slight improvement in the F value, which fluctuates around 0.67. Compared with the OS-ID3 model, the F value of the OUS-ID3 model is basically not improved, still fluctuating around 0.67, and the distribution is relatively scattered. The F value of the DBS-ID3 model is around 0.78, which is significantly improved compared to other models, and its distribution is also the most concentrated, showing excellent stability. Among the four models, the weighted ID3 model has the worst classification effect, the OS-ID3 model and the OUS-ID3 model have similar classification effects, and the DSB-ID3 model has the best classification effect. The classification effect of the combination on the data is significantly improved. The mean values of G and F values for each model are shown in Fig. 7.

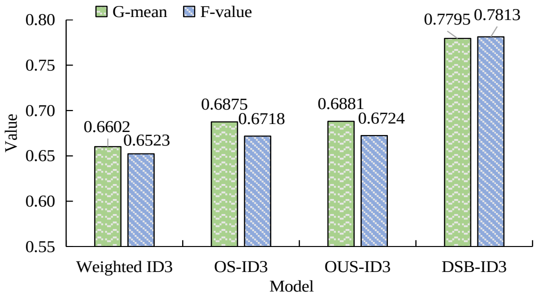


Fig. 7. Mean value of G-mean and F-value of each model

In Fig. 7, the mean G value of the weighted ID3 model is 0.6602, and the mean F value is 0.6523; the mean G value of the OS-ID3 model is 0.6875, and the mean F value is 0.6718; the mean G value of the OUS-ID3 model is 0.6881, and the mean F Value The mean value of the value is 0.6724; the mean value of the G value of the DSB-ID3 model is 0.7795, and the mean value of the F value is 0.7813. The G value and F value of each model are not much different. With the further improvement and optimization of ID3, the G value and F value of the model continue to increase, and its classification effect is also significantly improved. The G value and F value of the DSB-ID3 model largest, and its classification effect is also the best. The recognition accuracy of each model to the data is shown in Fig. 8.

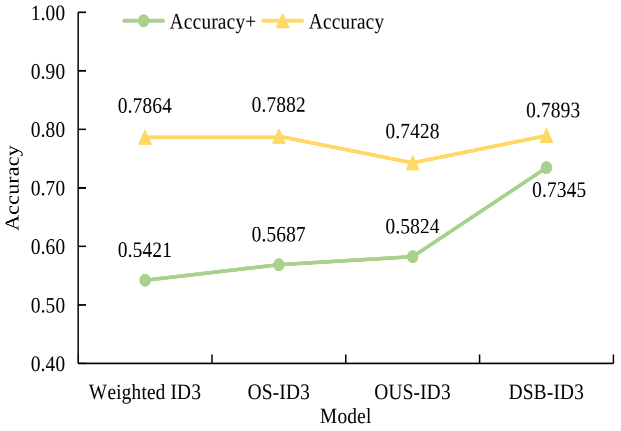


Fig. 8. Data recognition accuracy of each model

In Fig. 8, Accuracy+ represents the recognition accuracy of the model for positive data, and Accuracy represents the recognition accuracy of the model for all data. The accuracy of the weighted ID3 model for the positive data is 0.5421, and the total accuracy is 0.7864; the accuracy of the OS-ID3 model for the positive data is 0.5687, and the total accuracy is 0.7882; The accuracy of class data is 0.5824, and the total accuracy is 0.7428; the accuracy of DSB-ID3 model for positive class data is 0.7345, and the total accuracy is 0.7893. It can be seen that among the four models, DSB-ID3 model has the highest recognition accuracy and total recognition accuracy for regular data, indicating that its classification effect is the best, and it shows superior accuracy and stability in the analysis of financial data.The final student classification accuracy is shown in Fig. 9.

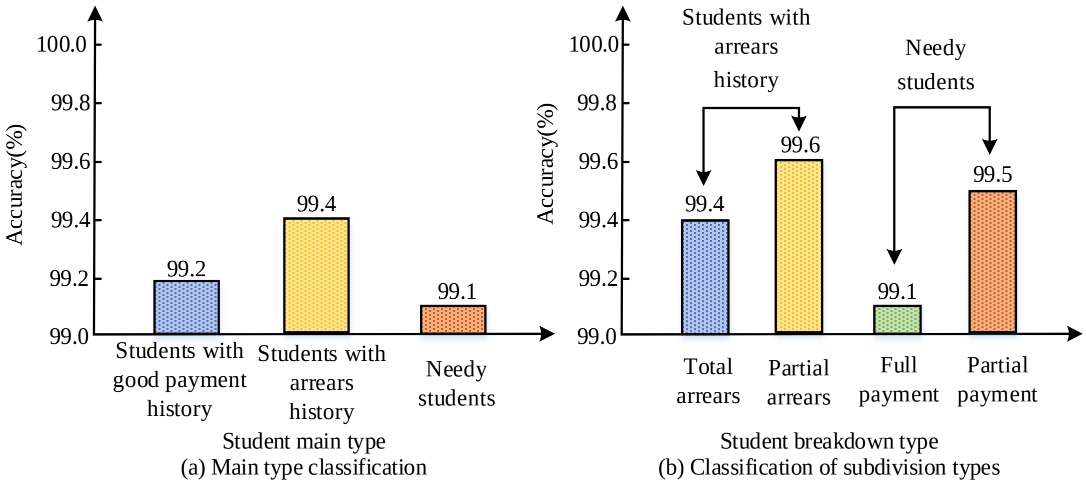


Fig. 9. Student classification accuracy

It can be seen that, for different types of students, the model has a classification accuracy of 99.2% for students with good payment history, 99.3% for students with arrears history, and 99.1% for poor students, all above 99.0%. At the same time, in terms of the classification of historical students in arrears, the classification accuracy reached 99.4% for all students in arrears and 99.6% for some students in arrears. In terms of poor students, the classification accuracy of all students who have paid all the fees has reached 99.1%, and the classification accuracy of students who have paid part of the fees has reached 99.5%. It can be seen that the model designed in this study can obtain good classification effect in practical application, has feasibility and practicality, and has high operational accuracy.

V. Discussion

With the development of management informatization of higher education structure, the demand for financial intelligent management in colleges and universities is growing day by day. The main goal of financial work has also changed from the original accounting type to the analytical type. ID3 algorithm itself has a significant disadvantage, that is, it tends to choose more values rather than the optimal attribute. Therefore, this research introduces risk weight to improve ID3 algorithm, and designs DSB-ID3 model based on SMOTE algorithm and Bagging algorithm. The analysis results show that with the increase of experimental time, the G value of each model fluctuates repeatedly, but except for DSB-ID3 model, the G value of weighted ID3 model, OS-ID3 model and OUS-ID3 model shows a general downward trend, and the linear regression slope is 1.74 respectively × 10^{-4}、1.32 × 10^{-4} and 1.25 × 10^{-4}。 The F value of each model fluctuates repeatedly with the increase of experimental time, and the change rule of each model is roughly similar to that in Fig. 5. Specifically, F values of weighted ID3 model, OS-ID3 model, OUS-ID3 model and DSB-ID3 model show a downward trend in general, and their linear regression slopes are 0.82 respectively × 10^{-4}、0.94 × 10^{-4} and 0.71 × 10^{-4}, 0.35 × 10^{-4}, it can be seen that the overall decline speed of DSB-ID3 model is the slowest. It can be seen that using DSB-ID3 model to analyze college financial data, its G value and F value are the highest among the four models, with the best classification effect. Compared with other models, it has significantly improved, and its distribution is the most centralized. The research results of Apalkova V and others also show that the classification ability of the decision tree model can be improved to a certain extent after using SMOTE algorithm to improve the decision tree model [24]. Among the four models, DSB-ID3 model has the highest accuracy, with a recognition accuracy of 0.7345 for positive data and a total recognition accuracy of 0.7893. In the final classification accuracy rate of students, the classification accuracy rate of this model for different types of students exceeds 99%. The classification effect of DSB-ID3 model is the best, showing superior accuracy and stability in financial data analysis. The research results of Podhorska I also show that the classification accuracy of the improved integrated algorithm using SMOTE algorithm is higher than that of the original algorithm [25].

VI. Conclusion

In order to overcome the disadvantage that ID3 algorithm tends to select more values but not necessarily the optimal attribute, this study introduces risk weight to improve ID3 algorithm and designs DSB-ID3 model. Using this model to analyze the financial data of colleges and universities, its G value and F value are the highest among the four models, about 0.78. The classification effect is the best. Compared with other models, it has significantly improved, and its distribution is the most centralized. For superior stability. Among the four models, DSB-ID3 model has the highest accuracy, with a recognition accuracy of 0.7345 for positive data and a total recognition accuracy of 0.7893. In the final classification accuracy rate of students, the classification accuracy rate of this model for different types of students exceeds 99%. The classification effect of DSB-ID3 model is the best, showing superior accuracy and stability in financial data analysis. The DSB-ID3 model of financial analysis in colleges and universities shows excellent classification performance in financial data analysis. The accuracy and stability of its classification effect provide practical reference for the application of financial analysis in colleges and universities. This research uses the annual financial report data of the real university to carry out the research, which ensures the rationality of the research results. However, the model still has many shortcomings in practical application. For example, this study only uses student payment data for experiments, and the application of school financial reports is relatively low.

Acknowledgment

The research is supported by the the 2022 Shandong Provincial Educational Science Planning and Innovation Literacy Special Project &apos; Research on the Integration Path of Innovation and Entrepreneurship Education and Professional Education for Finance and Economics major in Higher Vocational Colleges&apos; (Project No.: 2022CYB356, Project Leader: Guo Sujuan) and 2021 General subject of the China Vocational Technology Education Society &apos; Research on the Digital Course Development Model of the Smart Finance Professional Group in the New Professional Catalogue of Higher Vocational Education&apos;(Project No.: 2021B067, Project Leader: Guo Sujuan).

References

[1] A. Luthfiarta, J. Zeniarja, E. Faisal, W. Wicaksono, “Prediction on Deposit Subscription of Customer based on Bank Telemarketing using Decision Tree with Entropy Comparison,” Journal of Applied Intelligent System, Vol. 4, No. 2, pp. 57-66, 2020.

[2] J. Tian, Y. Wang, W. Cui, K. Zhao, “Simulation analysis of financial stock market based on machine learning and GARCH model,” Journal of Intelligent and Fuzzy Systems, Vol. 40, No. 2, pp. 2277-2287, 2021.

[3] T. Mahara, I. Naim, “Framework to identify a set of univariate time series forecasting techniques to aid in business decision making,” International Journal of Intelligent Enterprise, Vol. 7, No. 4, pp. 423-443, 2020.

[4] Z. Wu, “Using Machine Learning Approach to Evaluate the Excessive Financialization Risks of Trading Enterprises,” Computational Economics, Vol. 59, No. 4, pp. 1607-1625, 2022.

[5] Y. Li, Y. Pan, “A novel ensemble deep learning model for stock prediction based on stock prices and news,” International Journal of Data Science and Analytics, Vol. 13, No. 2, pp. 139-149, 2021.

[6] J. Wyrobek, “Application of machine learning models and artificial intelligence to analyze annual financial statements to identify companies with unfair corporate culture,” Procedia Computer Science Vol. 176, pp. 3037-3046, 2020.

[7] X. Luo, X. Wen, M. C. Zhou, A. Abusorrah, L. Huang, “Decision-Tree-Initialized Dendritic Neuron Model for Fast and Accurate Data Classification,” IEEE Transactions on Neural Networks and Learning Systems, Vol. 33, No. 9, 4173-4183, 2021.

[8] I. Podhorska, J. Vrbka, G. Lazaroiu, M. Kovacova, “Innovations in Financial Management: Recursive Prediction Model Based on Decision Trees,” Marketing and Management of Innovations Vol. 3, pp. 276-292, 2020.

[9] M. A. Febriantono, S. H. Pramono, R. Rahmadwati, G.Naghdy, “Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0.” Vol. 9, No. 1, pp. 65-72, 2020

[10] M. F. Maulana, M. Defriani, “Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period,” PIKSEL Penelitian Ilmu Komputer Sistem Embedded and Logic, Vol. 8, No. 1, pp. 39-48, 2020.

[11] Y. Xia, L. He, Y. Li, Y. Xu, “A Dynamic Credit Scoring Model Based On Survival Gradient Boosting Decision Tree Approach,” Technological and Economic Development of Economy, Vol. 27, No. 1, pp. 1-24, 2020.

[12] M. Balta, İbrahim Özçelik. “A 3-stage fuzzy-decision tree model for traffic signal optimization in urban city via a SDN based VANET architecture,” Future Generation Computer Systems Vol. 104, pp. 142-158, 2020.

[13] T. Qiu, M. Zhang, X. Liu, J. Liu, W. Zhao, “A Directed Edge Weight Prediction Model Using Decision Tree Ensembles in Industrial Internet of Things,” IEEE Transactions on Industrial Informatics, Vol. 17, No. 3, pp. 2160-2168, 2020.

[14] J. Keuangan, P. Dan, A. Muditomo, S. Broto, “IPO Performance Prediction during Covid-19 Pandemic in Indonesian Using Decision Tree Algorithm,” Jurnal Keuangan dan Perbankan, Vol. 25, No. 1, pp. 132-143, 2021.

[15] L. Xue, D. Liu, C. Huang, X. Lin, X. S.Shen, “Secure and Privacy-Preserving Decision Tree Classification with Lower Complexity,” Journal of Communications and Information Networks, Vol. 5, No. 1, pp. 16-25, 2020.

[16] S. Ziweritin, I. A. Ibiam, T. A. Oyeniran, G.E. Oko, “KNN and Decision Tree Model to Predict Values in Amount of One Pound Table,” International Journal on Computer Science and Engineering, Vol. 9, No. 7, pp. 17-21, 2021.

[17] Natawibawa I W Y, Mulya I M O, Yoh W H. Transparency and Accountability As Determinants in the Financial Management of Universities: a Study on State Universities in Malang City. Jurnal Tata Kelola Dan Akuntabilitas Keuangan Negara, 2019, 5(1): 57-72.

[18] Min Du. Management Research of College Finance under Informatization Background//.Proceedings of 2018 4th International Conference on Education & Training,Management and Humanities

Science(ETMHS 2018).Clausius Scientific Press,2018:70-74.

[19] Na Sun. Discussion on the Application of Management Accounting in University Finance//.Proceedings of 2019 5th International Workshop on Education,Development and Social Sciences(IWEDSS 2019).Francis Academic Press,2019:164-168.

[20] Atanasijevic J, Milosevic D. Upgrading the business intelligence system by implementing the decision tree model in the R software package. Studies in Informatics and Control, 2020, 29(2):243-254.

[21] Punitha S, Jeyakarthic M. Enhanced Particle Swarm Optimization with Decision Tree based Prediction Model for Stock Market Directions. Journal of Advanced Research in Dynamical and Control Systems, 2020, 12(5):1432-1442.

[22] Feofanov A, Egorov M. IMPROVING THE ENTERPRISE PERFORMANCE ON THE BASIS OF APPLYING DECISION TREE METHOD. Automation and Modeling in Design and Management of, 2021(1):29-34.

[23] Syamala M, Nalini N. A Filter Based Improved Decision Tree Sentiment Classification Model for RealTime Amazon Product Review Data. International Journal of Intelligent Engineering and Systems, 2020, 13(1):191-202.

[24] Apalkova V, Tsyganov S, Meshko N, Tsyganova N，Apalkov S. Application of decision tree model for prediction of immigration policy in different countries of the world. Problems and Perspectives in Management, 2021, 19(3):513-532.

[25] Podhorska I, Vrbka J, Lazaroiu G, Kovacova M. Innovations in Financial Management: Recursive Prediction Model Based on Decision Trees. Marketing and Management of Innovations, 2020(3):276-292.