Dean&Francis

ISSN 2959-6149

Analysis of physical training performance data based on decision tree

Yixiang Yu

Abstract:

Under the background of the current information age, the scientific, accurate and high efficiency of sports training effect evaluation is very important to improve the quality of physical exercise. However, the traditional manual evaluation method is limited by subjective and objective factors, which is often difficult to accurately reflect the training status of the students. This paper aims to solve the inaccurate and time-consuming problems in the evaluation of sports training performance, which has important practical significance and application value. Through the visualization technology to reveal the changes of the training performance, and the modeling and prediction combined with the decision tree algorithm, the accurate estimation of the training performance of the students can be realized, providing strong support for the training feedback and effect improvement.

Keywords: physical training, decision tree, data mining

1. INTRODUCTION

1.1 Background

Data mining technology can provide leaders in various fields with necessary knowledge information, which can produce huge economic or quality benefits. Therefore, although the current data mining technology is not mature, its demand in all walks of life is increasing, more enterprises and units gradually use data mining to analyze internal data, and then discover rules and assist decision-making. At present, data mining technology has been widely used in biomedical (Yao, 2019), industrial processes (Wu, 2017), military applications(Wang, 2020), aerospace(Li, 2020) and other important fields(Yu, 2021). However, the survey found that data mining technology has not been fully applied in physical exercise, and its application ability has not been paid attention to by physical exercise personnel. However, sports personnel do a lot of physical training and testing every day, which produces a large amount of data, but the intrinsic value of these data is often ignored by the people involved. They tend to judge the training performance of a period, that is, whether the training effect is achieved. On the one hand, this method will have a large deviation, and on the other hand, it is difficult to estimate for all the trainers. Therefore, how to use the data mining technology to mine the training effect information from the daily physical exercise performance and predict the training score will be the core of this study.

1.2 Research Objectives

The research objectives are to analyze and predict

the physical training data. The data mining techniques adopted in this paper are mainly two: data visualization technology to observe the effect of physical exercise, and the other is decision tree (DT) technology in machine learning. Using this technology to predict physical exercise performance, this can reduce the data mining burden of coaches and improve efficiency, and on the other hand to realize automatic sports performance prediction.

1.3 Reasons for choosing this topic

My plan for my future career tends to be focused on information technology, and at the same time, I am also a sports lover. Therefore, I expect to systematically learn data mining to improve personal performance in physical training. In addition, I have noticed that the PE teachers in our school face many challenges in analyzing the students' PE test scores, and I hope to help with this purpose. In fact, the analysis and prediction of student physical training data have significant practical significance in high school education. By making full use of the training data, teachers can better understand the students' needs and make more accurate predictions. At the same time, students can also use data analysis to clarify their own shortcomings in physical exercise, to strengthen the relevant training more targeted.

1.4 Outline of the research

The main content of this project is to make the application and analysis of practical problems by combining the knowledge and means of data mining. In this paper, DT technology is used to predict the comprehensive training performance to demolish the burden of coaches and improve the sports evaluation. Specifically speaking, this paper mainly includes the following aspects:

First, the background, significance of the research and the research status at home and abroad. Through the in-depth review of the relevant materials and literature at home and abroad and the analysis and combing of the research status of domestic and foreign papers, the significance of this paper and the practical problems that should be solved are expounded.

Second, the introduction of the relevant theoretical basis. The principle of the DT technology to be used in this paper is introduced, and the technical details to be used are combed.

Third, exploratory data analysis. The acquired data were cleaned and preprocessed. By using exploratory analysis, the performance of students, the changes of the training weeks, the difference between different grades, and the changes of the training weeks, to see the effect of physical exercise.

Fourthly, the DT algorithm is predicted based on the information entropy and Gini coefficient, and the best node classification rules are selected by comparison. At the same time, the grid optimization algorithm is used to find the optimal algorithm parameters and models.

Fifth, summarize. Summarize the work done by the institute and the shortcomings in the model process and propose prospects for future research.

2. LITERATURE REVIEW

2.1 Basic knowledge of classification and prediction

Data prediction is one of the mainstream technologies in data mining technology. This technology is mainly to mine the internal information of raw data through certain analysis and modeling means, and then establish the corresponding classification model, and finally realize the prediction of new achievements through this model. This process is usually divided into two steps:

In the first step, acquire the data and model it based on the known data. For the existing data, it needs to be normalized, and after processing, the data needs to meet the specific format requirements. Each row of the data represents a sample, and each row needs to correspond to a category, which is the attribute description of the data of the line (Hou, 2020). After obtaining the standard format, it can be led through machine learning algorithm. Considering that most of the machine learning algorithms are based on supervised learning models, this requires a part of the data to train the model, and then the model can be predicted for the specific data to be tested. Therefore, after standardizing the dataset, it is necessary to divide it into the training set and the test set in a certain proportion, which is usually set at 7:3 or 8:2. When the data set is large, it is usually set freely.

In the second step, the obtained model is loaded for classification prediction. Classification prediction usually needs to establish the corresponding model in the first step. The type of this model is not limited, and it can be a mathematical formula, a DT form, or a black-box classification based on a specific algorithm. At this point, the prediction and classification can be made. For example, after giving a customer data, it can estimate whether the customer is in good credit according to the established model, and then assist the bank to decide whether to lend to (She, 2010). Classification rules can also be used in other specific occasions, such as identifying and predicting the data situation in the future to realize the extension of functions.

However, it should be noted that for each test sample, the usual method is that the training and test set are complete-

ISSN 2959-6149

ly separated, that is, there is no crossover between the training and test set. Once the data set test set overlaps or crosses, the resulting model will approximate the original data indefinitely, which leads to the excessively optimistic results of the resulting model on the test set, and then leads to the generally low accuracy in the prediction of the location data, that is, poor generalization performance.

2.2 Classification and prediction step

Model classification and prediction usually have the following steps:

(1) Data collection and screening

Precise raw data is the basis for the model prediction. Building a predictive model requires sufficient historical data. Collect accurate data and analyze the influencing factors on the prediction, and accurately screen out the effective data required by the model training (Wang, 2020).

(2) Preprocessing of the data

The raw data was pre-processed. Normalization processes the data into a convenient standard for training and eliminates the irregular or interfering data in order to meet the training/testing requirements.

(3) Select the prediction model

According to the characteristics of the data quantity, choose the appropriate prediction method according to the situation of the data analysis, and then build the prediction model according to the corresponding method and set reasonable parameters for it.

(4) Verify the prediction results

The prediction model is trained, which is used to predict the unknown data samples, and compare the results with other prediction models to verify the validity of the results and simultaneously compare and analyze the error.

2.3 Modeling process of DT

2.3.1 DT generation algorithm

DT is a classic classification and prediction algorithm, which is mainly composed of many nodes. The nodes of the DT are roughly divided into two types, one internal and one external. The internal node is usually connected with the directed segment again, which means that is, the node can be separated, while the external node, also called the leaf node, is usually connected with the end of the tree network, and the leaf node indicates the final taxonomic attribution. Correspondingly, it is no longer connected to a directed segment. For the internal node, because he needs to continue the division, he needs certain rules of judgment.

For a DT, its classification and prediction process are usually started by the root node, then trained based on the properties of the training dataset, and then selected by the best branch based on the results of the training process, until the final leaf node is reached. At each leaf node, there is usually a corresponding classification result, that is, the decision result, which represents the final classification of the model training on each training data. When making prediction, it is necessary to establish the good tree structure network according to the realization and judge the attributes according to the trained node division rules. This process is the process of prediction.

As for how to determine which variable belongs to the root node or sub-node, it is necessary to introduce two concepts related to the classification rules of DT, respectively, information entropy and Gini coefficient.

Because the amount of information that each data can contain is completely determined by its uncertainty, so for the calculation of entropy, the larger the entropy of a number, the greater its uncertainty. Let there be a pending sample X, which has Xi (i=1,2,3,... n) possibility, using pi to indicate the probability that it belongs to i possibility. At this time, the definition formula of entropy can be written as:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2.1}$$

In the above formula, p represents the probability, and the larger the value, its entropy is closer to 0, while when the probability is 0.5, the entropy is 1 and the entropy is 0 when the probability is 1. Where entropy 1 means the data has the greatest uncertainty and 0 indicates no uncertainty (Sun, 2021). In the process of training, the value of entropy will constantly decrease until it drops to a minimum, which is essentially a process of reducing the classification error. If the entropy of a variable is found to be almost zero or reaches a minimum, the variable is set to a parent node and used to generate a DT. Thereafter, the parent node will continue to divide according to the above rules and produce many branches, and for each possible branch, the model will calculate the error rate during training until the minimum entropy value is achieved. Repeat this process reaches the final leaf node (Tang, 2021).

In the above process, to calculate the reduced level of uncertainty in the partition process, one usually needs to introduce a new concept, namely conditional entropy. Assuming the presence of variable variables (X, Y), the probability distribution is:

$$p(X = x_i, Y = y_i) = p_{ij}, i = 1,2,...n; j = 1,2,...,m \tag{2.2}$$

H (Y / X) represents the conditional entropy, which is the uncertainty of the model Y in the variable X condition, so in the X condition, the conditional entropy H of Y can be expressed as:

$$H(Y|X) = \sum_{i=1}^{n} p_i H(Y|X = x_i),\ p_i = P(X = x_i), i = 1, 2, ..., n \quad (2.3)$$

Accordingly, the gain of information can be easily obtained, that is, how much the model can reduce the classification uncertainty after obtaining the information expressed in X (Chen, 2021). Therefore, it can further represent the information gain of the data elements relative to the training set D (Zhang, 2021),

$$g(D, A) = H(D) - H(D|A) \quad (2.4)$$

Thus, for any given training set and features, empirical entropy can be employed to represent its uncertainty for classification. However, because the different element features in the sample often vary greatly in terms of information gain, features with large gain are usually used during training, which tend to have high classification accuracy. The greater the information gain, the greater the purity of the tree structure that can be obtained with this feature (Xue, 2019).

The above is the training and model tree construction process based on the entropy. The division method of Gini coefficient is like the method of information entropy.

Further, the workflow of the DT can be summarized, roughly as follows:

(1) Collect the data, normalize the obtained data, and add the corresponding labels, so that it can be loaded by the model.

(2) The data set is partitioned into the training and test set in a predetermined ratio, ensuring that there is no intersection between the two subsets.

(3) For the training sample, the information gain is calculated for each feature value of it, and the gains of different features obtained are sorted, and then the feature column corresponding to the largest gain is selected.

(4) For the feature column with the largest information gain obtained in step 3, take it as the root node, and then then divide the subsequent node into according to this node (Liu, 2019). Then repeat this step constantly, know that the data on the final leaf node belongs to a category or there is no way to continue the segmentation, to this time can get the complete model.

(5) The resulting DT model is loaded, and then the model is brought into the predicted sample, and then the parameters are optimized to obtain the accuracy of the prediction.

2.3.2 DT pruning

In the process of creating DT, because this paper intends to use physical exercise data set is large, directly training model will get a large model, the model because the model will consider a variety of different combination of segmentation, namely the model will be all features of the data set loaded into the model training, to design the DT algorithm , namely the model purity is high enough (Ma(2020)). However, if such a model is directly used for prediction, it will produce very large errors, and some features of extreme data in the training set will also be learned. At this time, the prediction model deviation will be very large, which is the so-called “overfitting” phenomenon.

For solving the overfitting problem, pruning is usually used to control the tree size and size. This method is to cut some branches at the end of the DT after the original model is built, and then use its leaf node as the final node, so that on the one hand, the network can greatly simplify the network, furthermore, it can solve the overfitting problem well and improve the generalization ability of the system. At this point, if a new data set is used to predict, its accuracy will be much higher than the original unpruned DT.

In the process of pruning, he mainly checks a group of nodes with the same parent node. If he merges them, he determines whether the increase in model entropy will be less than a certain threshold. If this condition is met, they can be merged, and if not, they cannot be merged.

2.3.3 Advantages of the DT algorithm

Using the DT algorithm for prediction has the following advantages, mainly including:

(1) The DT is easier to understand and explain. After building the DT, people can understand the meaning of the DT, which is also the key factor for modeling the DT model in this design.

(2) When using the DT method, it is often not a need to carry out complex preprocessing of the data. In other technologies, it is generally required to standardize the data before calculation and the blank and noise attributes are removed, while in the DT method, this step can usually be omitted.

(3) The DT method can make relatively good processing results for large data sets in a short time and can establish DT for data sets with many properties.

3. DATA PROCESSING AND ANALYSIS

3.1 Data sources

The data source is a record of the five general training achievements of sports, which are 1800 pieces, and each data includes the five general training achievements and the evaluation results of physical fitness. Among them, the five general training scores include 1000 meters, rope skipping, sit-ups, ball throwing and body shape, and the

ISSN 2959-6149

sports assessment results are scored by professionals, which are divided into four grades, namely excellent, good, passing and failing. The style of this dataset is shown in Table 1.

Table 1

part of the original data

No.	1000 Meters	rope skipping	sit-ups	ball throwing	body shape	sports assessment results
1	90	84	94	93	80	excellent
2	60	82	83	81	81	passing
3	84	83	88	86	83	good
4	86	82	83	83	86	good
5	84	83	85	80	78	good
6	87	82	88	77	85	good
7	78	82	82	57	86	failing
8	90	87	85	93	94	excellent
9	93	90	89	91	94	excellent
10	79	82	84	86	80	good
11	90	83	86	85	82	good

3.2 Data submission

What affects the final performance of physical exercise consists of 5 routine physical exercise scores. There are huge differences between different data. Therefore, to improve the prediction accuracy of the system, it requires to be normalized to keep their data scale consistent. The calculation formula is as follows:

$$X_t^{'} = \frac{X_t - X_{\min}}{X_{\max} - X_{min}} \tag{3.1}$$

In the output layer, use the formula to replace the actual value, and the formula is:

$$X_t = (X_{\max} - X_{\min})X_t^{'} + X_{\min} \tag{3.2}$$

In the formula, $X_{\max}$ and $X_{\min}$ represents the maximum and minimum values in the training sample set, respectively.

3.3 Visual analysis

3.3.1 Assessment distribution

Before classification, it is necessary to observe the distribution of data. Unbalanced data distribution will reduce the accuracy of classification, thus leading to inaccurate prediction results. Therefore, this paper needs to first observe the distribution of categorical variables before classification. The statistics of the different attribute values of this variable can get the distribution of candidates under the four levels of failing, passing, good and excellent. The quantity is plotted in Figure 1.

It can be seen that among all the candidates, 1238 candidates had ‘good’ grade, 325 candidates have ‘pass’ grade, 129 candidates has ‘excellent’ grade and 102 candidates has ‘failed’ grade.

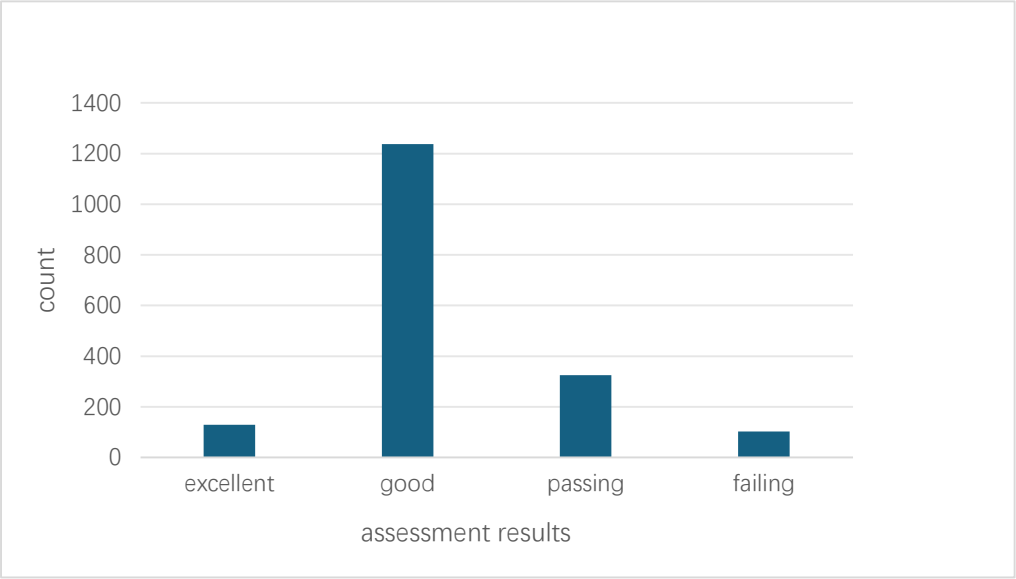


Figure 1 Distribution map of assessment

3.3.2 Score distribution of each training item

To observe the distribution of examinee scores in the five conventional tests, the current 1000 m, rope skipping, sit-ups, horizontal bar and body shape scores were calculated by the kernel density function, Figure 3.3 depicts the results.

It can be clearly seen that the score distribution curve of the candidates in each subject is similar, which means that the performance of the five scores is highly consistent. Overall, the peak of these four scores belongs to the right, so the overall scores are better, and most of them are concentrated in the 80-90 score section.

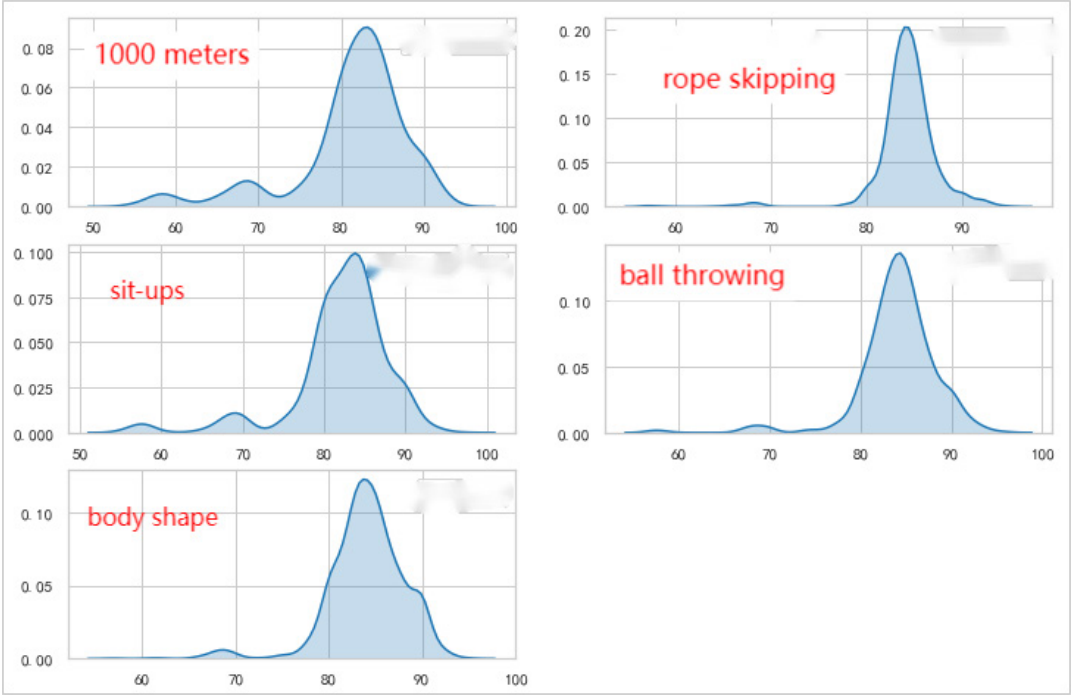


Figure 2 Distribution map of each training item

ISSN 2959-6149

4. PREDICTION BASED ON THE DT ALGORITHM

4.1 Evaluation index

For classification prediction problems, it is usually necessary to observe whether the prediction results are accurate, which needs not only from the macro evaluation indicators, but also from some micro indicators. Macro indicators are commonly used accuracy, F1 score, while micro indicators are generally for a certain category attributes, common recall rate, precision two, but also illustrated methods, such as confusion matrix.

For the calculation of several numerical indicators, we need to enter into several concepts, namely FP, TP, TN and FN. The interrelationship of these four concepts can be expressed in Figure 3.

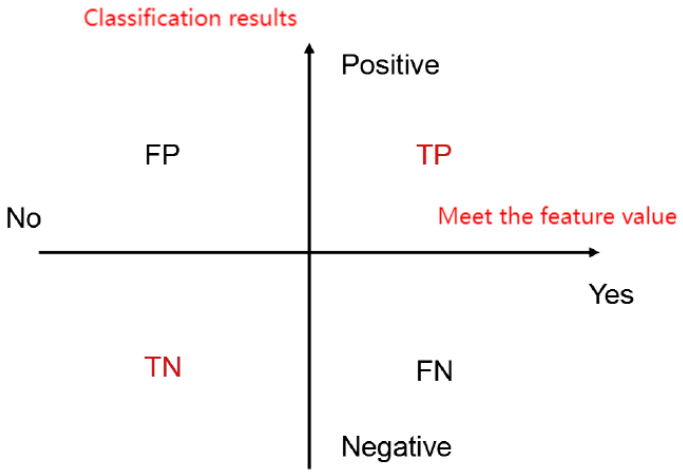


Figure 3 Confusion matrix

Based on the above four concepts, we can calculate the accuracy formula as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

The formula for the recall rate is given:

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{4.2}$$

The formula for the accuracy rate is:

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{4.3}$$

The formula for the F1 score is:

$$F_1 = \frac{2 * \mathrm{Re}call * \mathrm{Pr}ecision}{\mathrm{Re}call + \mathrm{Pr}ecision} \tag{4.4}$$

4.2 Performance prediction results based on the DT

In order to predict the comprehensive evaluation performance of the student according to the physical training performance, this paper takes the physical training performance as the influence factor variable and the comprehensive evaluation performance as the prediction amount. Considering that the predictor variables are object data, namely 'excellent', 'good', 'pass', and 'fail'. However, there is a size relationship between these discrete variables, so the numerical coding is according to a certain size order, that is, the coding is 4,3,2,1. At the same time, the ratio between training/test set was 4:1.

The DT algorithm contains two node division methods, one based on the Gini coefficient and the other on information entropy. This paper first uses these two methods to test separately.

4.2.1 Design of DT model based on Gini coefficient

Gini coefficient is a common rule of DT node division to choose the best form of attribute division when making DT modeling. When choosing which characteristic attribute each node should divide according to the Gini coefficient, the smaller the value is better since large Gini coefficient means that there will be a great uncertainty in the sample data. Therefore, the goal is to ensure that the value is as small as possible. By calculating the Gini coefficient under different feature division, the feature corresponding under the minimum value is selected and the node is divided as a principle.

Let the training set used in this paper be D, and the number of customer types to be predicted be K, then the possibility that the sample should belong to class k is p_k, and the Gini coefficient can be calculated as:

$$Gini(p)=\sum_{k=1}^{K} p_k(1-p_k) \\ =1-\sum_{k=1}^{K} p_k^2 \quad (4.5)$$

Further, we can calculate that when selecting a feature as a split object, the Gini coefficient is:

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(|D_1|)+\frac{|D_2|}{|D|}Gini(|D_2|) \quad (4.6)$$

Furthermore, the DT modeling process when using the Gini coefficient as the division criterion is shown in Figure 4.

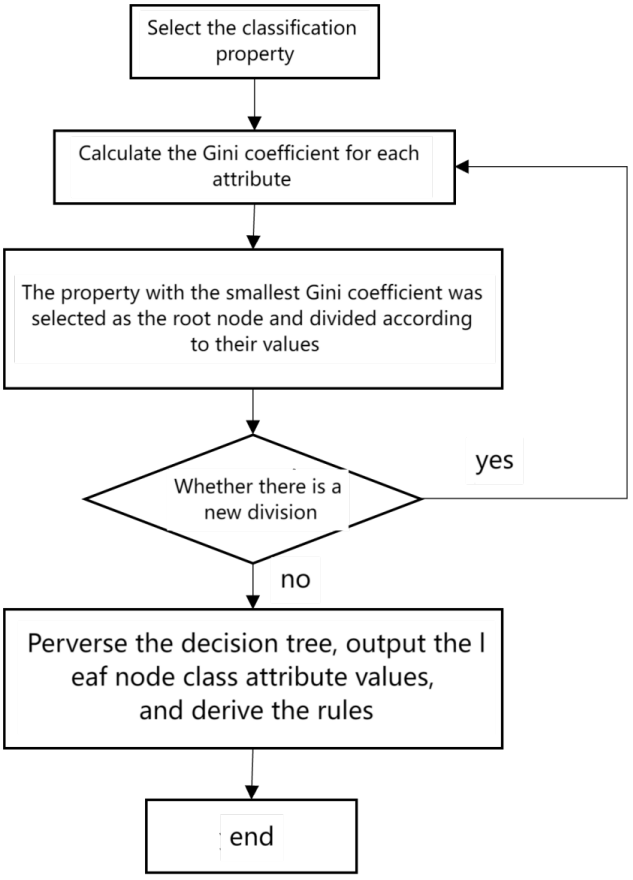


Figure 4 Flow chart of DT modeling based on the Gini coefficient

This method was used to predict the sports performance, Figure 5 shows the training process curve.

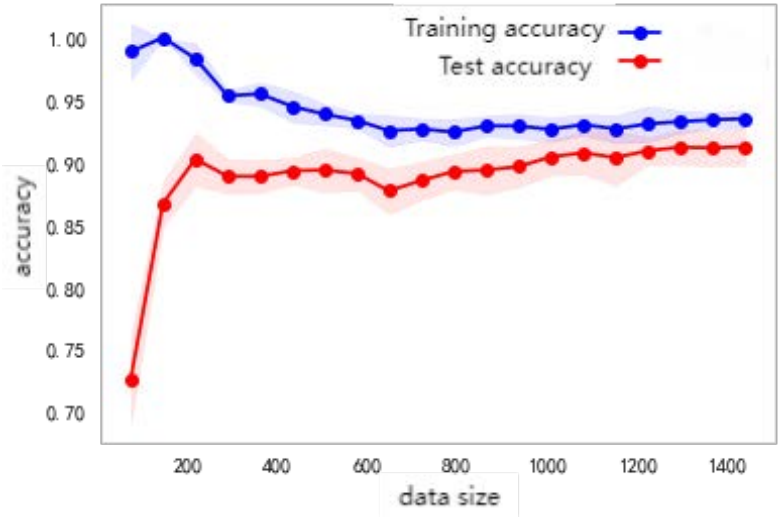


Figure 5 DT training process curve based on the Gini coefficient

4.2.2 Design of DT model based on Information entropy

Information entropy is another classification criterion for dividing DT nodes. In this way, it mainly obtains the attribute information with the maximum calculated value of the information gain. In this way to choose the best char-

ISSN 2959-6149

acteristic value, is usually choose the characteristics of the largest information gain, the reason is the largest information gain, representing the entropy value minus the current feature, the reduction is larger, this shows that the current entropy minimum, it also means that the current data set purity is higher, more stable. In this way, the calculation amount of the tree division is also very small. The formula for calculating the information gain can be expressed as follows:

$$igain(D,A) = Entroy(D) - \sum_{p=1}^{P} \frac{|D_p|}{|D|} Entroy(D_p) \quad (4.7)$$

The flow chart of the algorithm design for constructing the DT with information entropy is as follows:

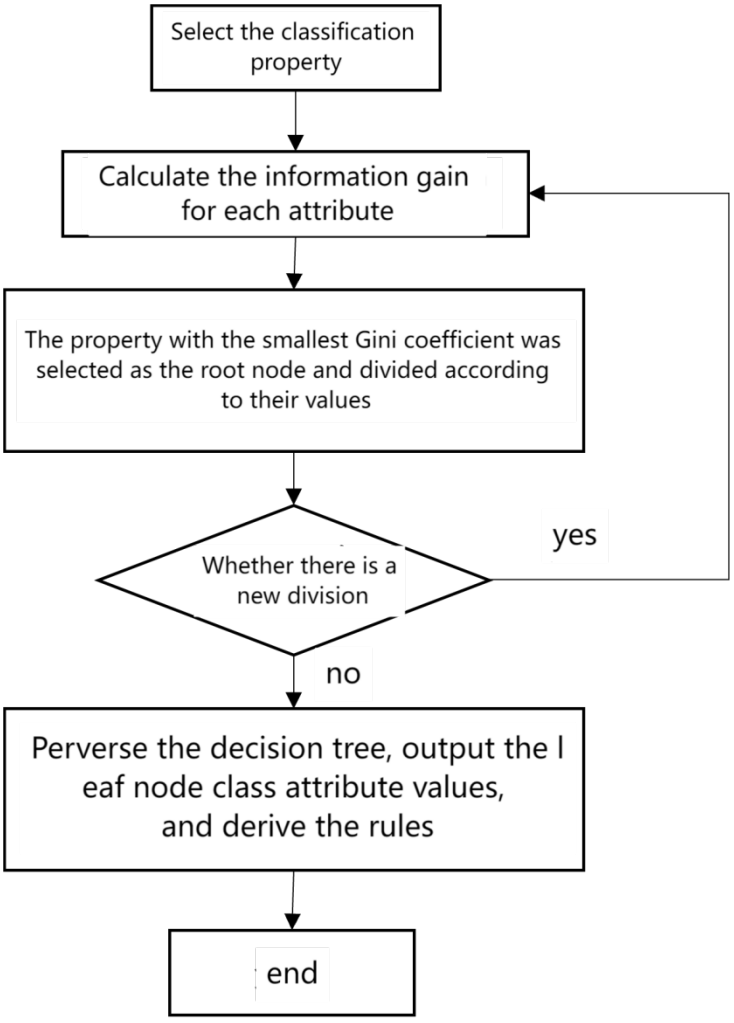


Figure 6 Flow chart of DT modeling based on Information Entropy

For the training set of this paper, Figure 7 shows the training process curve.

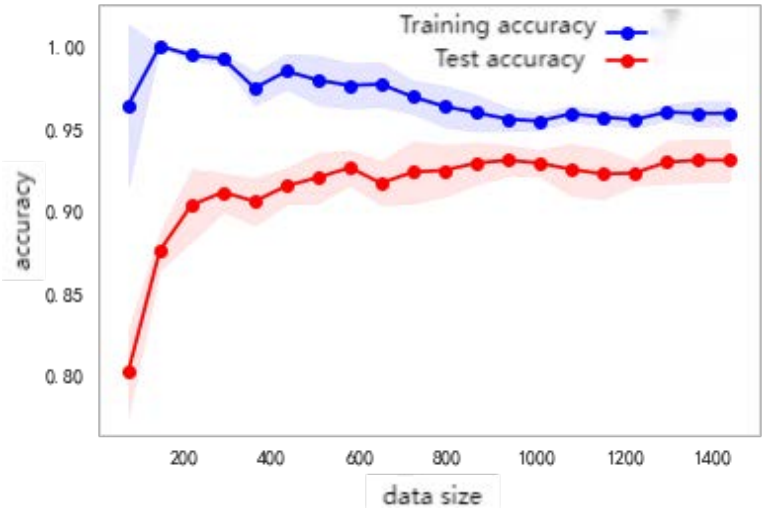
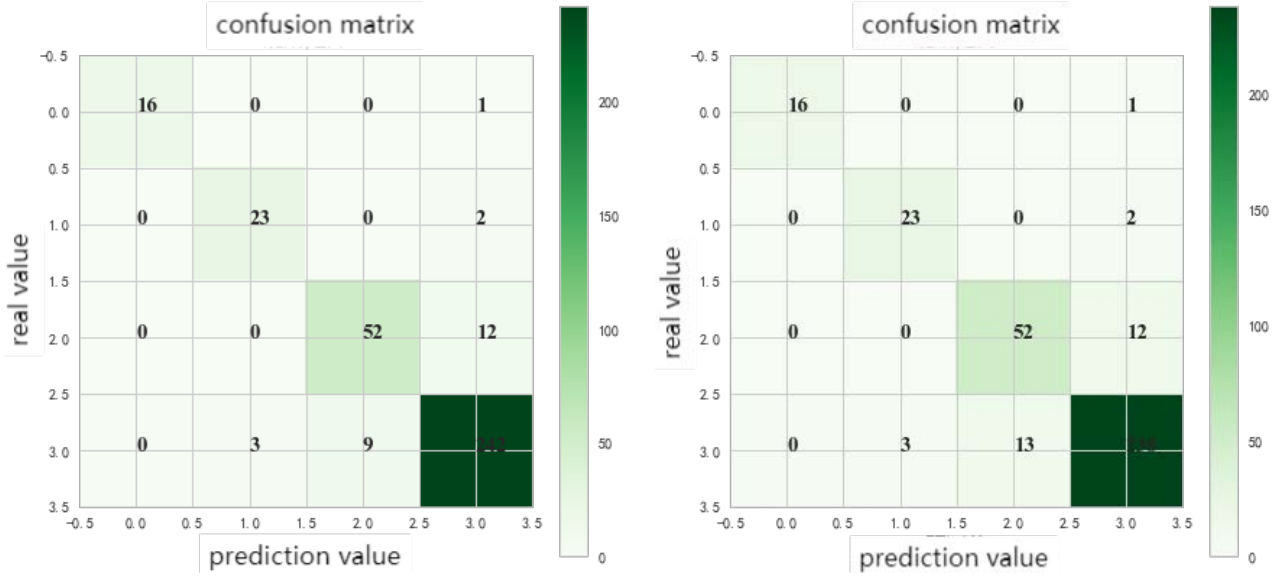


Figure 7 Training process curve of DT based on Information Entropy

4.2.3 Comparison between prediction results

Finally, the Gini coefficient and the information entropy are 91.38% and 92.50%, respectively. Further, the confusion matrix of the two algorithms is shown in Figure 8.



(a) **(b)**

Figure8 Confusion matrix

(a) DT algorithm based on information entropy, (b)DT algorithm based on Gini coefficient

It can be learned that the DT algorithm based on information entropy can achieve good accuracy, so the DT algorithm based on information entropy is used for prediction. The experimental results of other specific index parameters of the algorithm are shown in Table 2:

Table 2 Experimental results of DT algorithm based on Information Entropy

assessment level	Precision	Recall	F1
fail	1.00	0.94	0.97
excellent	0.88	0.92	0.90
pass	0.85	0.81	0.83
good	0.94	0.95	0.95

ISSN 2959-6149

average	0.92	0.91	0.91

According to the above data results, it is shown that the prediction rate of the DT algorithm based on information entropy is very high in the classification, and the accuracy of all the classifications reaches 80%, which also shows that the DT algorithm is very suitable for this performance prediction task.

According to the results, the paper is summarized and obtains the identification accuracy data, as shown in Table 3.

Table 3 Comparison of prediction accuracy of DT algorithm in different stages

prediction algorithm	Gini coefficient	information entropy
F1	0.90	0.91
accuracy	91.38%	92.50%

It can be seen that after comparing the calculation methods of internal node division, the prediction accuracy is improved to 92.50%, and the F1 score changes from 0.90 to 0.91. This demonstrates that the DT algorithm adopted in this paper can better meet the needs of training performance prediction in the overall prediction accuracy.

4.3 Results analysis

The purpose of this paper is to automatically evaluate the comprehensive results of the five physical training scores of '1000 meters', 'rope skipping', 'sit-ups', 'horizontal bar " and' body shape '. In order to analyze the final results, this paper analyzes the results of different stages of performance prediction. Details are as follows:

According to the test data in this paper, there are a total of 17 + 25 + 64 + 254 groups according to different grades. According to Figure 4.6, The overall analysis of the final scoring results is as follows: a total of 17 samples in the 'failing' score, 16 correct but 1 scoring error, the correct percent is 94.12%. A total of 25 samples in the 'pass' score, 23 correct but 2 scoring errors. The correct percent is 92.00%; A total of 64 samples in the 'excellent' score, 55 correct but 5 scoring errors. The correct percent is 85.94%. A total of 254 samples in the 'good' score, 244 correct but 10 scoring errors. The correct percent is 96.06%. The final overall prediction accuracy was 93.89%. Further, some of the predicted results are presented here, resulting in Table 4.

Table 4 Part of the prediction results

No.	1000 Meters	rope skipping	sit-ups	ball throwing	body shape	Real results	Predict results
1	90	84	94	93	80	excellent	excellent
2	60	82	83	81	81	pass	good
3	84	83	88	86	83	good	good
4	86	82	83	83	86	good	good
5	84	83	85	80	78	good	good
6	87	82	88	77	85	good	good
7	78	82	82	57	86	fail	fail
8	90	87	85	93	94	excellent	excellent
9	93	90	89	91	94	excellent	excellent
10	79	82	84	86	80	good	good

According to the above results, among the 10 selected players, the results of students No.2 predicted incorrectly, and the other results were all correctly predicted, and the overall prediction result was good.

5. CONCLUTION

For physical exercise performance prediction, has always been the focus of attention, through the study of the influencing factors, build the decision index system, and use the DT model in the machine algorithm, with relevant data as training and test samples, to establish the accuracy

of the prediction model and generalization ability to test. From the final prediction results, the prediction accuracy of physical exercise performance reached 92.5%, with a high consistency between the prediction results and the actual data, and the model achieved a relatively ideal prediction effect.

In conclusion, the prediction model adopted in this paper can achieve better prediction of training performance. But training performance prediction is a very complex research work, the transmission mechanism, the influence factors involved complex, to its accurate prediction must use more advanced, complex computer methods, because the author's own research ability is limited, in some details on the shortcomings, hope the future can have more in-depth research.

6. REVIEW

I believe that my research on prediction of physical training score was successful because it shed light on the application of information technology in problem solving. I have used algorithm to analyze the training data. Specifically, I have collected about nearly two thousand record of candidates which is a large collection of data. Then, I present and analyze the data using data visualization techniques. I tried to use the DT algorithm to predict the score and it has achieved a pleasing accuracy.

Through my research, I have a deeper understanding on data mining and processing, and these gains have stimulated my interest in learning information technology related major. In the research process, I also learned a lot of skills, such as how to plan time reasonably, how to use search engines and databases to search for literature, which made me more confident in learning how to write papers in the future.

However, since this is my first time that doing research individually, the study is far from perfect.

On the one hand, the research is only based on the overall performance, and the performance prediction of a single subject has not been realized. In the next step, I plan to conduct targeted analysis of each test subject. On the other hand, there are many improvements in the DT model, which can further improve the prediction accuracy. Next, I will further learn the improved model of the DT.

7. REFERENCES

[1] Chen, Y. T. (2021). Analysis and comparison of ensemble learning algorithms: Random forest and gradient boosting decision tree. Computer Knowledge and Technology, 17(15), 32-34.

[2] Fu, W. (2021). Sports team performance prediction based on data-driven and data envelopment analysis. Journal of Shandong Sport University, 37(04), 102-111.

[3] Han, Y. H., Cao, L. J., Wang, Y. Q., & Zhang, J. J. (2022). Analysis of pest prediction model based on decision tree CART algorithm. Modernizing Agriculture, (01), 45-47.

[4] Hou, J. (2020). Prediction of university students' academic performance based on educational data (Master's thesis). Dalian University of Technology.

[5] Li, L. Z. (2021). Research on grade prediction model based on online learning behavior analysis (Master's thesis). Guizhou University of Finance and Economics.

[6] Sun, J. (2021). Research on the application of Gaussian Naive Bayes algorithm in college students' performance prediction. Computer Knowledge and Technology, 17(20), 23-26.

[7] Sun, X. X., Zhong, H., & Chen, H. P. (2021). Statistical analysis system of student examination scores based on decision tree classification technology. Journal of Jilin University (Engineering and Technology Edition), 51(05), 1866-1872.

[8] Tang, L., & Li, F. (2021). Research on security situation prediction model of Internet of Vehicles based on decision tree. Computer Science, 48(S1), 514-517.

[9] Wang, J. Y., Zhang, Y. F., & Xu, Z. (2020). Student performance prediction based on feature selection optimization. Think Tank Times, (01), 124-125.

[10] Wang, X. X., & Tang, J. (2020). Comparative analysis of student performance prediction models. Computer Knowledge and Technology, 16(01), 199-202.

[11] She Yanda & Li Haichen. (2010). Research on credit evaluation system of credit customers based on data mining. Journal of Information, (07), 141-143+192.

[12] Xu, W., Liu, W. T., Zhan, X. Q., Xu, X. J., & Chen, X. L. (2021). Research on grade prediction method based on university classroom learning behavior. Modern Educational Technology, 31(06), 104-111.

[13] Xue, Y. N., & Yang, X. D. (2019). Application of decision tree algorithm in student performance. Science and Technology Information, 17(36), 83+85.

[14] Yao, M. H., Li, J. S., & Wang, N. (2021). Prediction of college students' academic performance based on BP neural network. Journal of Jilin University (Information Science Edition), 39(04), 451-455.

[15] Yu, J., Bai, S. Y., & Wu, D. X. (2021). Research on online teaching student performance prediction based on machine learning. Computer Programming Skills & Maintenance, (08), 118-119+154.

[16] Zheng, A. Q., Wang, Y. Q., & Hao, C. Y. (2021). Research on online learning performance prediction based on deep learning. Computer Era, (12), 69-72+75.

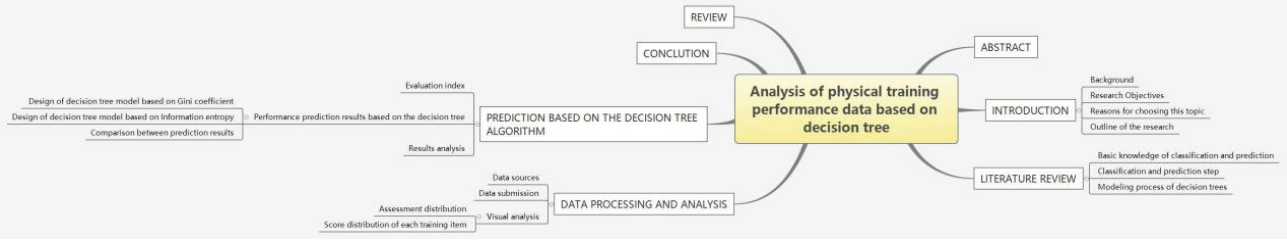
[17] Zhang, J. (2021). Analysis of error theory and survey adjustment course assessment based on decision tree. Surveying and Mapping Equipment, 23(01), 32-35+59.

[18] Song, J. (2020). Research on learning behavior analysis and academic performance prediction based on campus big data (Master's thesis).Central China Normal University.

ISSN 2959-6149

8. APPENDIX

Mind map:



Gantt Chart:

TASK		JUN				JUL				AUG				SEP			
		Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4
Project preparation																	
Determine the topic and list study objectives	Plan																
	Reality																
Research studies and surveys	Plan																
	Reality																
Project Execution																	
Complete the Proposal form	Plan																
	Reality																
Complete the Introduction	Plan																
	Reality																
Complete the Literature Review	Plan																
	Reality																
Complete the Method & Results and format the Dissertation	Plan																
	Reality																
Project Output																	
Complete everything elese and do the Presentation	Plan																
	Reality																