ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20200507022306
Original scientific paper

Comparison of Job Satisfaction Prediction Models for Construction Workers: *CART* vs. Neural Network

Tao CHEN, Zhonghong CAO*, Yuqing CAO

Abstract: To establish a suitable prediction model of construction workers' job satisfaction, this study chooses the widely used models *CART* (Classification and Regression Tree) and *NN* (Neural network) in the prediction model to make a comparison and finds out the main influencing factors of construction workers' job satisfaction in occupational health and safety training. Through the investigation and analysis of 280 cases of empirical data, it is found that the *CART* model based on Kappa value and Accuracy of categorical variables have a better prediction effect, and the main factors affecting job satisfaction are job categories, working days per week and the latest training time. The main innovation of this paper is to add the actual value set of empirical data on the basis of the usual training set, verification set, test set and prediction set, and draw a conclusion by comparing the predicted value with the actual value of kappa.

Keywords: *CART*; construction workers; job satisfaction; Kappa; neural network

1 INTRODUCTION

The job satisfaction of construction workers is one of the hot topics in occupational health and safety training and management in the construction industry. There are many prediction models, such as support vector machine, decision tree, neural network, logistic regression, and Bayesian Network, which could be combined with the stepwise regression method to improve the classification accuracy of almost all data mining technologies [1]. At the same time, there are corresponding prediction models in various fields and industries, such as different machine learning models (ML) in weather prediction models. Using different input combinations of meteorological variables, four different ML models of evaporation prediction are established: classification regression tree (*CART*), cascaded correlation neural network (CCNN), gene expression planning (GEP) and support vector machine (SVM). It is shown from the result table that all ML models can well predict the evaporation at the study site [2], and it can be seen that the comparative prediction of multiple models is often used, which is also worthy of reference in this study.

In the field of Biochemistry, principal component analysis (PCA), multiple regression analysis (MLR), and artificial neural network (ANN) are used to study the toxicity and risk assessment of chemical compounds. Based on this, a quantitative model is proposed. The results show that MLR is suitable for the prediction of toxicity, however, compared with its results, the prediction of the artificial neural network is better and more effective [3], so the ANN model has better performance in some specific backgrounds. Some scholars used the *CART* model to estimate the hospital death probability of acute myocardial infarction (AMI). They used *CART* induction model, calculating the area under ROC curve to evaluate its performance (AUC) (95% confidence interval (CI)), and found that *CART* model is easier to use and explain because the generated decision rules can be applied without mathematical calculation [4, 5]. In addition, the prediction analyses of the gene chip, artificial neural network, and classification, and regression tree were used for meta-analysis of gene expression profile, and the two-layer genetic screening method was used to reduce the number of variables, leading to a good accuracy rate (close to 100%) [6]. In particular, some scholars applied hidden Markov model to residential energy consumption prediction, used the energy consumption data collected from four multi-story buildings in Seoul, South Korea for model verification and result in analysis, and compared the model prediction results with three commonly used prediction algorithms, namely support vector machine (SVM), artificial neural network (ANN) and classification regression tree (*CART*) [7]. It can be seen that *CART* model and *NN* model are widely used, so this study attempts to apply these two models to the construction enterprise site and to make a breakthrough on the basis of previous research through the analysis and prediction of workers' job satisfaction.

CART is a decision tree intelligent discriminant analysis method [8, 9], which is widely used in social sciences. Research shows that the free statistical modelling of machine learning and equation artificial intelligence is a very promising comprehensive tool [10], and that *CART* analysis can also be applied to clinical-pathological monitoring of Medicine [11]. In terms of soil and water conservation, agricultural production, and biodiversity of ecological functions, the importance of analysis factors can be determined by *CART* [12]. At the same time, in the field of engineering, the *CART* is mainly used in equipment improvements and analysis of geotechnical engineering characteristics. In engineering construction, *CART* could be combined with multiple regression, monitoring and recording the operation process of TBM (tunnel boring machine) and evaluation system [13], which could evaluate important problems of dam operation. Design and safety are evaluated for dam structure in the estimation of dam storage capacity [14]. Moreover, the *CART* has the following advantages: (1) Large model capacity: the model will select independent variables according to the contribution in all independent variables for analysis, so it can automatically handle a large number of independent variables, without worrying about the interference of unrelated variables into the model effect and other issues. (2) Wide range of application: The target variable can be either a discrete variable or a continuous variable. It can also effectively deal with the problem of exact variables. (3) The model level is clear, readable, and understandable. Therefore, this model is selected as the main analysis method of empirical data.

NN was used in algorithm learning and optimization earlier formed an improved integrated learning algorithm and was applied to diagnosis and improve the quality of the track, providing an important guarantee for the safe operation of the track. At the same time, with the improvement of the resource utilization of rolling bearing, the operation cost was greatly reduced [15, 16]. Some scholars applied a neural network to the performance analysis of materials and established the tunnel risk evaluation model by combining fuzzy mathematics and BP neural network [17]. In addition, the response factors of the equipment were modelled to improve the accuracy of numerical simulation, improve the thermoforming process [18], as well as to measure and monitor research. The electromechanical impedance (EMI) technology and backpropagation neural network (BPNNs) are used to monitor the bolt looseness inside the bolt ball joint [19]. In addition, some scholars have gradually applied neural networks to the fields of biology [20], ecology, and medical chemistry [21]. The application of neural network in the field of engineering construction closely related to this study mainly includes the application of project risk assessment [17, 22], structural strength analysis, stability analysis [23], environmental safety and other factors [24] and early warning analysis of influencing factors [25]. BP (backpropagation) neural network is a concept put forward by the scientists led by Rumelhart and McClelland in 1986. It is a multilayer feedforward neural network trained according to the error backpropagation algorithm, which is the most widely used neural network at present. It is composed of 1 group of interconnected operation units, each of which has a corresponding weight. BP neural network consists of three parts: the input layer, middle hidden layer (one or more layers), and the output layer. This paper designs and models according to the principle of the BP neural network.

Although *CART* and *NN* have many applications in the field of engineering construction, there are few kinds of literature used to predict and analyze the satisfaction of construction workers [26]. Besides, in the research process of the *CART* and *NN* model, the comparative analysis between the predicted value and the actual value of the model is added, which can better reflect the prediction effect of the established model.

2 MATERIAL AND METHODS

2.1 Data Sources

The data of this study comes from a one-to-one field questionnaire survey, which designs 22 questions related to occupational health and safety training and job satisfaction of construction site workers (including 21 multiple-choice questions and 1 open-ended suggestion question), and then preliminarily investigates 12 construction workers, modifies the answer options, and finally forms a formal questionnaire. The questionnaire was designed and investigated from December 2018 to May 2019. To ensure the representativeness of the questionnaire survey, four representative regions (provinces) in China, namely East China (Shandong Province), South China (Hainan Province), central China (Hubei Province), North China (Hebei Province), 10 construction projects, 299 workers were randomly interviewed face to face, forming 280 effective questionnaires. Based on the survey data, the effective questionnaire data is divided into two parts, the first part is 239 for the analysis of training set, verification set, and a test set of *CART* and *NN*, and the second part is 41 for the prediction set. The analysis tools are IBM SPSS statistics 23 [27] and IBM SPSS modeler 18.0 [28].

2.2 Molecular Descriptors

In this paper, 21 questions in the questionnaire are designed as 20 independent variables (*X*1 - *X*20) and 1 dependent variable (*Y*). See Tab. 1 for the variable table. For the convenience of analysis, the independent variables in the above variable table are divided into three parts, the first part is *X*1 - *X*7, the second part is *X*8 - *X*13 (*X*8 - *X*10, *X*11 - *X*13), and the third part is *X*14 - *X*20.

Table 1 Variable description table

Variable	Description
*X*1	When have you recently received occupational
*X*2	About the number of working days per week
*X*3	Your job type
*X*4	What do you think is the effect of recent occupational
*X*5	Have you received (OHS*) training
*X*6	How do you acquire your job skills and knowledge
*X*7	How many hours do you work per day
*X*8	other related training
*X*9	another related training is more appropriate
*X*10	Have you ever witnessed an accident
*X*11	Cumulative working life in the construction industry
*X*12	Have you ever experienced an accident
*X*13	Do you have a vocational skill certificate
*X*14	Your gender
*X*15	Your age
*X*16	Your level of education
*X*17	Your marital status
*X*18	Have you worked in other industries
*X*19	Do you often pay attention to information
*X*20	plays a role in job responsibility awareness
Y	Your job satisfaction

*OHS: Occupational Health and Safety

In the research of this paper, descriptive statistics of the survey population is crucial to the final conclusion. Therefore, the quantitative statistical description of the five variables (*X*14, *X*15, *X*16, *X*17, *X*3) of the survey population is shown in Tab. 2. In addition, 280 questionnaires are all collected and analyzed in this table.

2.3 Statistical Analysis

The statistical and modelling analysis process is shown in Fig. 1. The statistical analysis data comes from OHS (20190403).sav and OHS (0201test-PRE).sav. The modelling uses IBM SPSS modeler 18.0, which is divided into two parts: model establishment and model prediction application. The two models are divided into two parts: *CART* and *NN* for corresponding analysis. Finally, the *CART* and *NN* models are compared and optimized

Table 2 Frequency of the categories of Construction Site variables*

Variable/Category Codes	Variables/Category of Variables	Frequency / %
*X*14	Your gender	100.0
*X*14-A	Male	88.2
*X*14-B	Female	11.8
*X*15	Your age	100.0
*X*15-A	Under 18 years old	1.4
*X*15-B	18 - 29 years old	12.5
*X*15-C	30 - 39 years old	27.1
*X*15-D	40 - 49 years old	34.6
*X*15-E	50 - 59 years old	22.9
*X*15-F	60 years old and over	1.4
*X*16	Your level of education	100.0
*X*16-A	Primary school	17.9
*X*16-B	Junior high school	48.9
*X*16-C	Senior high school	22.1
*X*16-D	Secondary school	7.5
*X*16-E	Junior college	1.8
*X*16-F	College and above	1.8
*X*17	Your marital status	100.0
*X*17-A	Unmarried	8.2
*X*17-B	Married	86.1
*X*17-C	Divorced	5.0
*X*17-D	Widowed	.7
*X*3	Your job type	100.0
*X*3-A	Woodworking	30.0
*X*3-B	Reinforcing steel	17.5
*X*3-C	Concrete	9.6
*X*3-D	Masonry	7.5
*X*3-E	Painter or decorator	6.8
*X*3-F	Electrician	5.7
*X*3-G	Plumber	4.6
*X*3-H	Repairman	2.1
*X*3-I	Tower, crane(hoist), outdoor elevator or signaller	8.6
*X*3-J	Other types of work	7.5

* Valid: *N* = 280

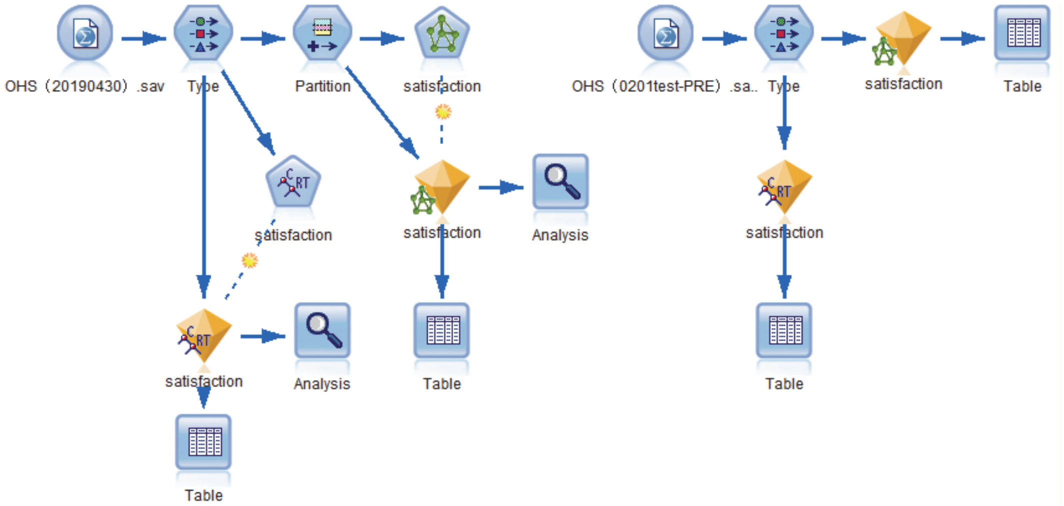

Figure 1 Modelling analysis process chart

3 RESULTS

3.1 Data Set for Analysis

First of all, the 20 independent variables of 239 survey data are preliminarily modelled by IBM SPSS modeler 18.0. The predictor importance parameter in the analysis results shows that (see Tab. 3), 7 of the 10 main variables of *CART* (*X*1 - *X*10) and *NN* (*X*1 - *X*7, *X*11 - *X*13) are consistent (*X*1 - *X*7), and the weight of the independent variables in *CART* and *NN* is also consistent from the table. Therefore, in the later comparative analysis, this paper will make comparative classification analysis with 20 independent variables, 13 independent variables, 10 independent variables, and 7 independent variables.

Table 3 Predictor Importance

CART	Importance	*NN*	Importance
*X*1	0.1745	*X*1	0.0871
*X*2	0.1644	*X*4	0.0754
*X*3	0.1231	*X*5	0.0673
*X*4	0.0978	*X*3	0.0669
*X*8	0.0904	*X*2	0.0645
*X*5	0.0622	*X*11	0.0644
*X*6	0.0516	*X*12	0.0643
*X*7	0.0468	*X*13	0.0628
*X*9	0.0238	*X*6	0.0565
*X*10	0.0207	*X*7	0.0507

Secondly, it can be seen from the relevant literature that the Kappa analysis method can be well used for the consistency analysis of comparative data. The focus of this paper is to compare the consistency between the predicted value and the actual value of two modelling analysis methods (*CART* and *NN*), and to judge the model prediction according to the size of Kappa value.

According to the number of independent variables, this paper classifies them into *K*20 (*X*1 - *X*20), *K*13 (*X*1 - *X*13), *K*10 and *K*7 (*X*1 - *X*7), and analyzes and compares the Kappa values of each category (see Tab. 4). In the analysis table, *R* stands for *CART* forecast value, *N* stands for *NN* forecast value, SA stands for actual value, and $R \cdot SA$ row stands for Kappa value (Value, Asymptotic Standardized Error^a, Approximate T^b, Approximate Significance).

Table 4 Symmetric Measures of Kappa

MAK		*Value*	*AST*^{a}	*AT*^{b}	*AS*
*K*20 (*X*1 - *X*20)	$R \cdot SA$	.570	.123	3.815	.000
	$N \cdot SA$	.230	.144	1.561	.118
	$R \cdot N$	.119	.151	.780	.435
*K*13 (*X*1 - *X*13)	$R \cdot SA$	.570	.123	3.815	.000
	$N \cdot SA$	.340	.113	2.841	.005
	$R \cdot N$	.250	.118	2.059	.040
*K*10	$R \cdot SA$	.570	.123	3.815	.000
	$N \cdot SA$	.037	.138	.267	.789
	$R \cdot N$	.225	.145	1.503	.133
*K*7 (*X*1 - *X*7)	$R \cdot SA$	.151	.120	1.223	.221
	$N \cdot SA$	−.034	.110	−.312	.755
	$R \cdot N$	−.101	−.125	−.692	.489
*K*13(5:5)	$N \cdot SA$	.010	.148	.065	.948
*K*13(7:3)		.340	.113	2.841	.005
*K*13(9:1)		.398	.146	2.533	.011
*K*13(8:1:1)		.302	.185	.203	.839
N of Valid Cases		41			

MAK: Measure of Agreement, Kappa, *AST*: Asymptotic Standardized Error, *AT* Approximate T, *AS* Approximate Significance.
a. Not assuming the null hypothesis. b. Using the asymptotic standard error assuming the null hypothesis.

Besides, the *P*-value of $N \cdot SA$ is significant only in *K*13. This paper attempts to further optimize the value of *NN* modelling under *K*13. See the later part of Tab. 4 again. Compare and analyze the input value of the *NN* model according to different values of the training set, verification set, and test set. It can be seen that when the ratio of the training set and test set is 9:1, the Kappa value of the predicted value and actual value of *NN* model ($N \cdot SA$) was the highest (Value =.398, Approval Significance =. 011), and *P*-value was significant ($P < 0.05$).

Finally, it can be seen from the prediction accuracy of the two models (see Tab. 5.), in the *K*13 category (*NN* model input value is 9:1), the accuracy of *CART* is 76.15%, and the accuracy of *NN* is 71.70%. The accuracy of *CART* prediction is higher than that of *NN* prediction.

Table 5 Comparing $N-satisfaction with $R -satisfaction

Classification	*CART* / %	*NN* / %
*K*20	76.15	69.80
*K*13(*NN* = 9:1)	76.15	71.70
*K*10	76.15	70.60
*K*7	75.31	70.60

In conclusion, it is found that in the prediction model of job satisfaction of construction workers, the prediction consistency and accuracy of the *CART* model are higher than those of the *NN* model. Next, this paper will further analyze the two kinds of optimized prediction models.

It can be seen from Tab. 4 that the kappa value $R \cdot SA$ (.570), $N \cdot SA$ (.340) and $R \cdot N$ (.250) of *K*13 in the four types of data (*K*20, *K*13, *K*10, *K*7) are the largest, and their P values are significant ($P < 0.05$). At the same time, it is found that the results of $R \cdot SA$ in the first three types of data (*K*20, *K*13, *K*10) are consistent (Value =.570, Asymptotic Standardized Error^a =.123, Approximate T^b = 3.815, Approximate Significance =.000), which shows that the performance of *CART* modelling method is stable, and the Kappa value of *CART* in all categories is greater than that of *NN*, so the modelling analysis of *CART* in construction workers. The result of job satisfaction analysis is better than the *NN* method.

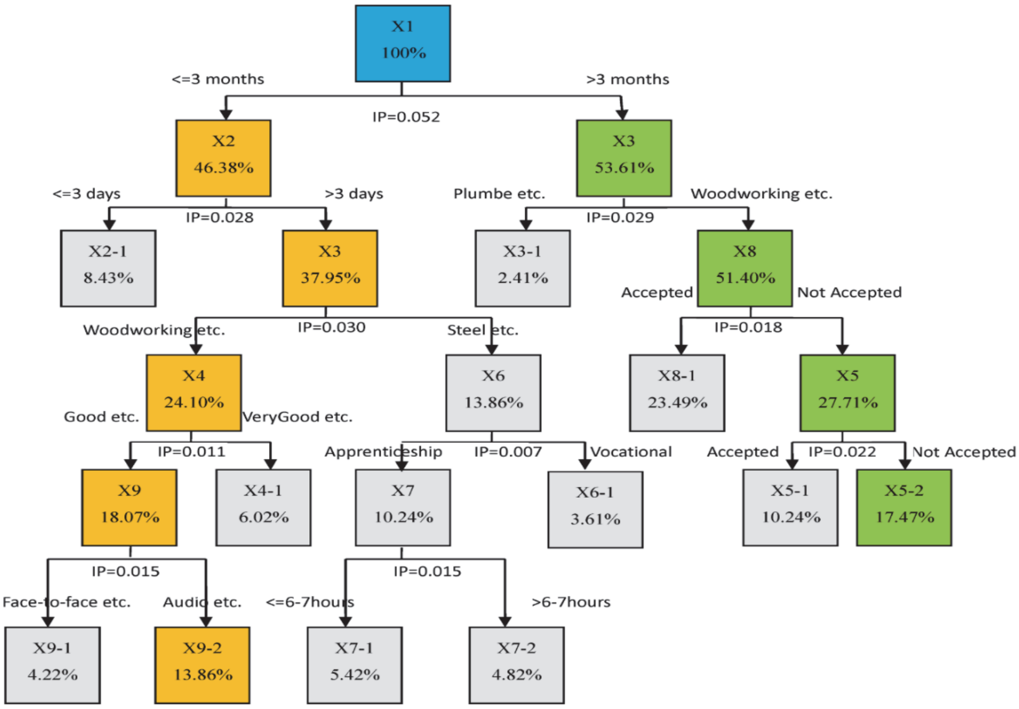

Figure 2 Chart of *CART* (IP: Improvement)

3.2 Classification and Regression Tree (*CART*) and Neural Networks (*NN*)

It can be seen from Fig. 2 that the probability of variable *X*1 is 100%, the probability of branch variable *X*3 in the lower two branches > 3 months is 53.61%, and the probability is 46.38% higher than that of the other branch *X*2, indicating that more workers have participated in training for more than 3 months recently. From the colour map of *CART* branch, it is obvious that the variables with a higher probability value of each layer of the two branches are *X*1 - *X*3 - *X*8 - *X*5 (*X*5-2) and *X*1 - *X*2 - *X*3 - *X*4 - *X*9 - (*X*9-2), and it can be seen that *X*3 appears in both branches, indicating that the variable *X*3 (work type) has a greater impact on the determination of job satisfaction, and the proportion of Woodworking workers in all work types is large. From the rightmost branch in the figure, it can be seen that the branch conditions of *X*3 - *X*8 and *X*5 - (*X*5-2) are not accepted, indicating that the workers with high probability have neither received relevant training nor OHS special training. From the left branch *X*2 - *X*3, it can be seen that the probability of working days > 3 days per week is high, and the probability of *X*4 - *X*9 is Good, indicating that the effectiveness of training is not very good, and the teaching methods of *X*9 - (*X*9-2) confirmatory video and apprenticeship are good.

It can be seen from the Predictor Importance (weight) sorting of Tab. 6 that the weight variables of the top three in *CART* column are *X*3 - *X*2 - *X*1, the weight variables of the top three in *NN* column are *X*5 - *X*4 - *X*9, and the variables with high probability are *X*1, *X*2, *X*3, *X*8, *X*4, *X*5, *X*9 in Fig. 2. Because of this, this paper divides the above variables into two categories: *X*3 - *X*2 - *X*1 - *X*8, *X*5 - *X*4 - *X*9, to show that the main factors affecting job satisfaction are: job category (*X*3), working days per week (*X*2), the latest training time (*X*1), participating in relevant training (*X*8), and the secondary factors are receiving OHS special training (*X*5), training effectiveness (*X*4), and training method (*X*9).

Table 6 Predictor Importance

CART	Importance	*NN*	Importance
X3	0.2699	X5	0.1466
X2	0.2026	X4	0.1387
X1	0.146	X9	0.1323
X8	0.097	X2	0.109
X4	0.0923	X11	0.1005
X9	0.0473	X1	0.0857
X7	0.0393	X7	0.0674
X5	0.0386	X6	0.0606
X6	0.0218	X3	0.0479
X10	0.0151	X10	0.0395

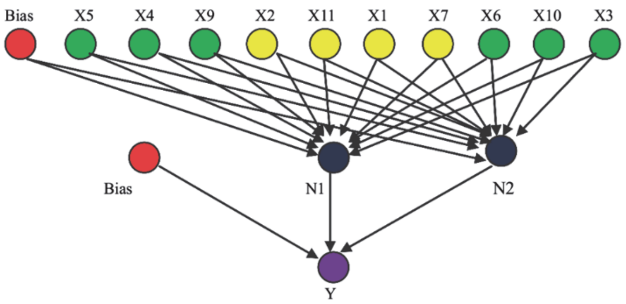

Figure 3 Chart of *NN* (*N*1: Neuron1, *N*2: Neuron2)

Besides, the *NN* model is shown in Fig. 3. It can be seen that the ratio of the training set and test set is 9:1, the network input layer has 10 independent variables (*X*5, *X*4, *X*9, *X*2, *X*11, *X*1, *X*7, *X*6, *X*10, *X*3) with larger weight. The middle hidden layer has two (*N*1, *N*2), and the Output layer is *Y*. The weight of each input variable to the hidden layer is shown in Tab. 6.

From the above analysis, we can see that there are many factors affecting job satisfaction (*X*3, *X*2, *X*1, *X*8, *X*5, *X*4, *X*9). SPSS software is used to test the independence of the Chi-Square test. The Chi-Square value and *P*-value of variable *Y* and other variables indicate that the relationship between variables is significant ($P < 0.05$), which can also be seen from Tab. 7.

Table 7 Chi-Square Tests

Variable	X^2	*P*-value	
$Y \cdot X3$	38.596	.003	Pearson Chi-Square
$Y \cdot X2$	96.889	.000	
$Y \cdot X1$	21.114	.002	
$Y \cdot X8$	15.703	.000	
$Y \cdot X5$	27.143	.000	
$Y \cdot X4$	24.294	.002	
$Y \cdot X9$	12.769	.047	

3.3 External Validation

Now, based on the above *CART* model and *NN* model established under *K*13, 41 pieces of data in the prediction set are selected to form a comparison chart (see Fig. 4, Fig. 5, and Fig. 6). It can be seen that the figures of $*N*, $*R* and *SA* on data point 12 - 21 from Fig. 4, the figures of $*R* and *SA* on data points 1 - 2, 5 - 9, 12 - 21, 27 - 29 from Fig. 5, and the figures of $*N* and *SA* on data points 12 - 21 from Fig. 6 are completely coincident, while the figures of other data points are not coincident significantly. Thus, it shows that the coincidence degree of the predicted value of *CART* ($*R*) and *SA* is fairly high, which means that the prediction effect of the *CART* model is better than that of the *NN* model.

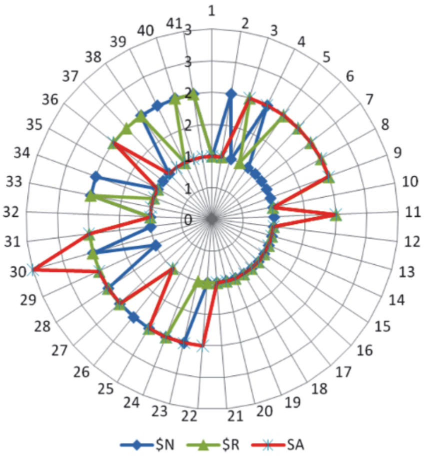

Figure 4 Comparison chart of predicted value ($*N*: *NN*, $*R*: *CART*) and the actual value (*SA*)

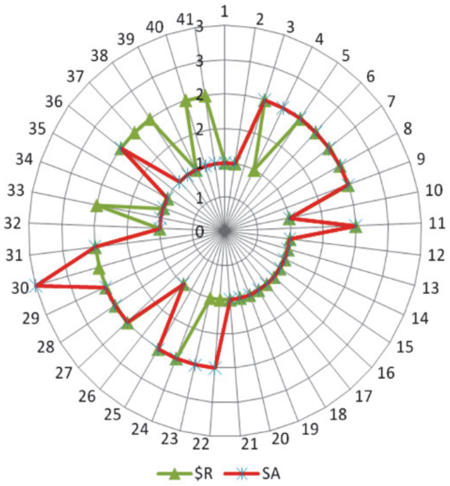

Figure 5 Comparison chart of the predicted value ($R: *CART*) and the actual value (*SA*)

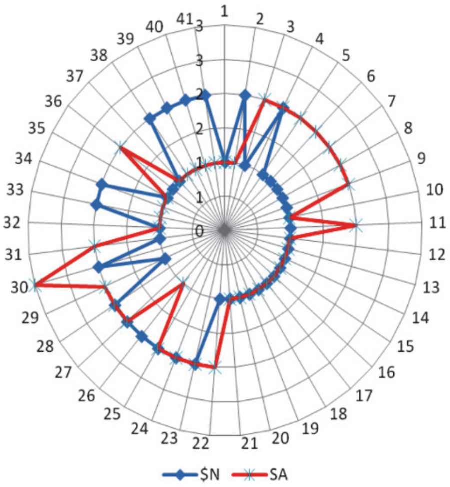

Figure 6 Comparison chart of predicted value ($N: *NN*) and the actual value (*SA*)

4 DISCUSSION

In order to establish a suitable prediction model of job satisfaction of construction workers, this paper makes a comparative analysis of *CART* and *NN* models. The research data comes from the face-to-face survey of field workers, and the reliability and validity of the data source are high. Most of the questions in the questionnaire are classified variables, and a few are continuous variables, which is different from the methods of continuous variable analysis in the past literature. For instance, there is a literature comparison between the *CART* and *NN* models, whose results show that the *NN* model has a better effect on the cost prediction of colorectal cancer patients [26]. However, in the prediction and analysis of this literature and other literature [29, 30], there is little comparative study of Kappa value from the actual value data source.

The main innovations of this study are as follows: first, based on the usual training set, verification set, test set and prediction set, the empirical data will be added to the actual value set, and the conclusion will be drawn through the comparative study of the kappa between the predicted value and the actual value. Therefore, the kappa value analysis of the predicted value and the actual value to evaluate the modelling effect is one of the main innovations of this paper. Secondly, the study found that the kappa value and accuracy of the *CART* model based on the classification variables are better, and the main factors affecting job satisfaction are job category, training time, and working days per week. Finally, use the *CART* model to find out the classification level relationship of the main influencing factors, and reveal the weight of sub-items, which is convenient to guide the management and practice of construction workers.

As for the main influencing factors of job satisfaction of construction workers, first of all, the job category has the greatest impact on satisfaction (the importance value is 0.2699, see Tab. 6). At the same time, according to Fig. 2, the proportion of woodworking, etc. is 51.40%. Therefore, it is recommended to use the classified management method to strengthen the post-management of woodworking, etc. in order to improve the overall job satisfaction and ensure the safety and efficiency of engineering production. Secondly, the importance value of working days per week on satisfaction is 0.2026 (see Tab. 6), and it can be seen from Fig. 2 that the workers whose working days per week are more than 3 days have the greatest impact on satisfaction. Through correlation analysis, it can be seen that the longer the working hours per week, the lower the satisfaction. Therefore, in combination with the actual situation of construction enterprises and posts, the arrangement of working days per week should be reasonable. Finally, recent training time has a great impact on job satisfaction. The occupational health and safety training in construction enterprises is not only important learning and training of daily management but also requires regular learning and training of four new technologies, namely, new technology, new materials, new processes, and new equipment, so as to improve the training frequency. It is suggested that the training interval should be 3 - 6 months.

In addition, in the process of this study, we also try to use stepwise regression dimension reduction method, but it is found that the classification accuracy of the model has not been improved. Et al. (2019) found that the stepwise regression method can reduce the dimension of variables and improve the classification accuracy of almost all data mining technologies. Obviously, the two conclusions are inconsistent [1], indicating that the dimension reduction method has limitations in the selection of variables. Therefore, the dimension reduction method is not generally applicable in *CART* and *NN* models.

In this study, *CART* and *NN* models are selected for prediction and comparative analysis. Because there are many prediction models, especially the support vector machine and Bayesian network model, which are widely used, a comparative analysis will be added in the follow-up study. In addition, performance indicators such as precision, recall, F1- score will also be added to the model analysis.

The *CART* model established in this paper can be used not only to predict and evaluate the job satisfaction of workers on-site but also to explore the main influencing variables and their hierarchical relationships. Certainly, the list of variable weight values output by the prediction model is directly related to the scale design and the number of input variables, that is, the subjective demand design of different researchers will affect the prediction effect of the model. Later, the author will explore and verify the prediction effect and applicability of the *CART* model in other industries.

Undoubtedly, this study is based on the specific background of the construction site, but the workers' scenes in the world are changeable, and none of the construction workers and industrial workers' scenes are the same. This is also an issue that needs to be paid attention to in this study, and it cannot be claimed that the *CART* model can always provide better results regardless of the research environment, especially for a specific study.

5 CONCLUSION

In the prediction model of job satisfaction of construction workers, the prediction effect of kappa value based on classification variables and *CART* model based on accuracy is better than that of the *NN* model. At the same time, the main influencing factors of occupational health and safety of construction site workers on job satisfaction are job category (*X*3), working days per week (*X*2), and the latest training time (*X*1).

This paper suggests that in the aspect of occupational health and safety training for construction workers, managers should classify the training according to different post categories, increase the training frequency, and arrange the working days of each week reasonably, so as to improve the job satisfaction of construction workers.

Acknowledgments

The authors would like to acknowledge all the researchers who provided support, including Zhenggang Yao, Director of New Technology of New Eight Construction Group Co., Ltd., and Hui Yang, Project Manager of Wuhan Construction Engineering Third Construction Co., Ltd. for the support in the questionnaire survey.

Author Contributions

Conceptualization, methodology, writing of original draft preparation, supervision, and funding acquisition, T. Chen; formal analysis, data curation, visualization, and project administration, Z.H. Cao; validation, T. Chen and Z.H. Cao; software, Z.H. Cao and Y.Q. Cao; writing of review and editing, Z.H. Cao and Y.Q. Cao.

Funding

This study was funded by the Key Soft Science Project of Hubei Science and Technology Support Plan (2015BDF006), National Social Science Project of Education Science 2017 (BIA170192), and the Key Open Fund Project of Hubei Industrial Policy and Management Research Center 2016-17 (17CYZ01).

6 REFERENCES

[1] Yao, J., Pan, Y., Yang, S., Chen, Y., & Li, Y. (2019). Detecting Fraudulent Financial Statements for the Sustainable Development of the Socio-Economy in China: A Multi-Analytic Approach. *Sustainability, 11*(6). https://doi.org/10.3390/su11061579

[2] Yaseen, Z. M., Al-Juboori, A. M., Beyaztas, U., Al-Ansari, N., Chau, K. W., Qi, C., & Shahid, S. (2020). Prediction of evaporation in arid and semi-arid regions: a comparative study using different machine learning models. *Engineering Applications of Computational Fluid Mechanics, 14*(1), 70-89. https://doi.org/10.1080/19942060.2019.1680576

[3] Ghamali, M., Chtita, S., Ousaa, A., Elidrissi, B., Bouachrine, M., & Lakhlifi, T. (2017). QSAR analysis of the toxicity of phenols and thiophenolsusing MLR and ANN. *Journal of Taibah University for Science, 11*(1), 1-10. https://doi.org/10.1016/j.jtusci.2016.03.002

[4] Trujillano, J., Sarria-Santamera, A., Esquerda, A., Badia, M., Palma, M., & March, J. (2008). Approach to the methodology of classification and regression trees. *Gaceta Sanitaria, 22*(1), 65-72. https://doi.org/10.1157/13115113

[5] Kim, J. K., Rho, M. J., Lee, J. S., Park, Y. H., Lee, J. Y., & Choi, I. Y. (2017). Improved Prediction of the Pathologic Stage of Patient With Prostate Cancer Using the CART-PSO Optimization Analysis in the Korean Population. *Technology in Cancer Research & Treatment, 16*(6), 740-748. https://doi.org/10.1177/1533034616681396

[6] Chu, C. M., Yao, C. T., Chang, Y. T., Chou, H. L., Chou, Y. C., Chen, K. H., & Chang, C. W. (2014). Gene Expression Profiling of Colorectal Tumors and Normal Mucosa by Microarrays Meta-Analysis Using Prediction Analysis of Microarray, Artificial Neural Network, Classification, and Regression Trees. *Disease Markers*. https://doi.org/10.1155/2014/634123

[7] Ullahl, I., Ahmad, R., & Kim, D. (2018). A Prediction Mechanism of Energy Consumption in Residential Buildings Using Hidden Markov Model. *Energies, 11*(2). https://doi.org/10.3390/en11020358

[8] Chrysos, G., Dagritzikos, P., Papaefstathiou, I., & Dollas, A. (2013). HC-CART: A Parallel System Implementation of Data Mining Classification and Regression Tree (CART) Algorithm on a Multi-FPGA System. *Acm Transactions on Architecture and Code Optimization, 9*(4). https://doi.org/10.1145/2400682.2400706

[9] Stojadinovic, M. M., Stojadinovic, M. M., & Pantic, D. N. (2019). Decision tree analysis for prostate cancer prediction. *Srpski Arhiv Za Celokupno Lekarstvo, 147*(1-2), 52-58. https://doi.org/10.2298/SARH181127039S

[10] Ryo, M., Jeschke, J. M., Rillig, M. C., & Heger, T. (2019). Machine learning with the hierarchy-of-hypotheses (HoH) approach discovers novel pattern in studies on biological invasions. *Research Synthesis Methods*. https://doi.org/10.1002/jrsm.1363

[11] Guevara, R., Stothers, L., & Macnab, A. (2011). Algorithm construction methodology for diagnostic classification of near-infrared spectroscopy data. *Spectroscopy-an International Journal, 25*(1), 1-11. https://doi.org/10.1155/2011/752101

[12] Zhao, H., Fang, X., Ding, H., Strobl, J., Xiong, L., Na, J., & Tang, G. (2017). Extraction of Terraces on the Loess Plateau from High-Resolution DEMs and Imagery Utilizing Object-Based Image Analysis. *Isprs International Journal of Geo-Information, 6*(6). https://doi.org/10.3390/ijgi6060157

[13] Jakubowski, J., Stypulkowski, J. B., & Bernardeau, F. G. (2017). Multivariate Linear Regression And Cart Regression Analysis Of Tbm Performance At Abu Hamour Phase-I Tunnel. *Archives of Mining Sciences, 62*(4), 825-841. https://doi.org/10.1515/amsc-2017-0057

[14] Unes, F., Demirci, M., Tasar, B., Kaya, Y. Z., & Varcin, H. (2019). Modeling Of Dam Reservoir Volume Using Generalized Regression Neural Network, Support Vector Machines And M5 Decision Tree Models. *Applied Ecology and Environmental Research, 17*(3), 7043-7055. https://doi.org/10.15666/aeer/1703_70437055

[15] Han, L., Yu, C., Liu, C., Qin, Y., & Cui, S. (2019). Fault Diagnosis of Rolling Bearings in Rail Train Based on

Exponential Smoothing Predictive Segmentation and Improved Ensemble Learning Algorithm. *Applied Sciences-Basel, 9*(15). https://doi.org/10.3390/app9153143

[16] Quan, Z. & Zhang, Z. (2014). The Construction and Approximation of the Neural Network with Two Weights. *Journal of Applied Mathematics*. https://doi.org/10.1155/2014/892653

[17] Deng, X., Xu, T., & Wang, R. (2018). Risk Evaluation Model of Highway Tunnel Portal Construction Based on BP Fuzzy Neural Network. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2018/8547313

[18] Li, L. & Wang, L. Y. (2018). Artificial Neural Network-Based Three-dimensiona Continuous Response Relationship Construction of 3Cr20Ni10W2 Heat-Resisting Alloy and Its Application in Finite Element Simulation. *High Temperature Materials and Processes, 37*(5), 411-424. https://doi.org/10.1515/htmp-2016-0234

[19] Xu, J., Dong, J., Li, H., Zhang, C., & Ho, S. C. (2019). Looseness Monitoring of Bolted Spherical Joint Connection Using Electro-Mechanical Impedance Technique and BP Neural Networks. *Sensors, 19*(8). https://doi.org/10.3390/s19081906

[20] Hu, J., Xin, P., Zhang, S., Zhang, H., & He, D. (2019). Model for tomato photosynthetic rate based on neural network with genetic algorithm. *International Journal of Agricultural and Biological Engineering, 12*(1), 179-185. https://doi.org/10.25165/j.ijabe.20191201.3127

[21] Yang, R. Y. & Rai, R. (2019). Machine auscultation: enabling machine diagnostics using convolutional neural networks and large-scale machine audio data. *Advances in Manufacturing, 7*(2), 174-187. https://doi.org/10.1007/s40436-019-00254-5

[22] Liang, C., Qian, C., Chen, H., & Kang, W. (2018). Prediction of Compressive Strength of Concrete in Wet-Dry Environment by BP Artificial Neural Networks. *Advances in Materials Science and Engineering*. https://doi.org/10.1155/2018/6204942

[23] Dong, W., Liu, X., & Li, Y. (2014). Analysis of Stiffened Penstock External Pressure Stability Based on Immune Algorithm and Neural Network. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2014/823653

[24] Xu, L., Zhang, T., & Ren, Q. (2015). Intelligent Autofeedback and Safety Early-Warning for Underground Cavern Engineering during Construction Based on BP Neural Network and FEM. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2015/873823

[25] Kim, M. K., Cha, J., Lee, E., Van Huy, P., Lee, S., & Theera-Umpon, N. (2019). Simplified Neural Network Model Design with Sensitivity Analysis and Electricity Consumption Prediction in a Commercial Building. *Energies, 12*(7). https://doi.org/10.3390/en12071201

[26] Lee, S. M., Kang, J. O., & Suh, Y. M. (2004). Comparison of hospital charge prediction models for colorectal cancer patients: neural network vs. decision tree models. *Journal of Korean medical science, 19*(5), 677-681. https://doi.org/10.3346/jkms.2004.19.5.677

[27] See https://www.ibm.com/nl-en/products/spss-statistics.

[28] See https://www.ibm.com/products/spss-modeler.

[29] Peyk, E. & Shahbahrami, A. (2019). Development of a Data Mining System for Subscriber Classification (Case Study: Electricity Distribution Company). *Tehnicki Vjesnik-Technical Gazette, 26*(4), 947-952. https://doi.org/10.17559/TV-20171202222013

[30] Olfaz, M., Tirink, C., & Onder, H. (2019). Use of CART and CHAID Algorithms in Karayaka Sheep Breeding. *Kafkas Universitesi Veteriner Fakultesi Dergisi, 25*(1), 105-110. https://doi.org/10.9775/kvfd.2018.20388

Contact information:

Tao CHEN, Professor, PhD
Wuhan University of Science and Technology,
Wuhan 430081, Hubei, P. R. China
E-mail: 917735337@qq.com

Zhonghong CAO
(Corresponding author)
School of Management, Wuhan University of Science and Technology,
Wuhan 430081, Hubei, P. R. China
E-mail: 418402664@qq.com

Yuqing CAO
Huazhong University of Science and Technology,
Wuhan 430074, Hubei, P. R. China
E-mail: 1249053233@qq.com