ISSN:1738-6764, eISSN:2093-7504

A Study on the Prediction of Community Smart Pension Intention Based on Decision Tree Algorithm

Lijuan Liu [1,2] and Byung-Won Min [3,*]

[1] Division of Information Technology Engineering, Mokwon University, Korea; Ph.D student; jkxllj1022@gmail.com
[2] School of Artificial Inteligence, Neijiang Normal University, China; Professor; jkxllj1022@gmail.com
[3] Division of Information and Communication Convergence Engineering, Mokwon University, Korea; Professor; minfam@mokwon.ac.kr
* Correspondence

https://doi.org/10.5392/IJoC.2021.17.4.079

Manuscript Received 07 October 2021; Received 15 December 2021; Accepted 27 December 2021

Abstract: *With the deepening of population aging, pension has become an urgent problem in most countries. Community smart pension can effectively resolve the problem of traditional pension, as well as meet the personalized and multi-level needs of the elderly. To predict the pension intention of the elderly in the community more accurately, this paper uses the decision tree classification method to classify the pension data. After missing value processing, normalization, discretization and data specification, the discretized sample data set is obtained. Then, by comparing the information gain and information gain rate of sample data features, the feature ranking is determined, and the C4.5 decision tree model is established. The model performs well in accuracy, precision, recall, AUC and other indicators under the condition of 10-fold cross-validation, and the precision was 89.5%, which can provide the certain basis for government decision-making.*

Keywords: Decision Tree; Prediction Model; C4.5 Algorithm; Community smart pension; 10-fold cross-validation

1. Introduction

China has been an aging society since 2000, and the proportion of the elderly in the total population continues to increase. According to Table 1, the age data of the national population in China's seventh census shows that. As of November 2020, the population of China over the age of 60 has reached 264 million, accounting for 18.7% of the total population. Among them, there are nearly 200 million people over 65 years old, accounting for 13.5%. It is predicted that by 2025, the elderly population will exceed 300 million [1].

The old-age dependency ratio in China is increasing year by year, and the social pension burden is becoming heavier and heavier. In order to better solve the pension problem, China began to vigorously develop the smart pension service mode in 2013. Community smart pension refers to the establishment of information management for the elderly in the community and the surrounding service institutions. The elderly and their children can apply for door-to-door services, health care, daily care, etc., through the mobile phone, to provide the elderly with health, convenience, home, medical, spiritual care and other professional services.

In China, the construction of community smart pension is an emerging big project of people's livelihood, and its construction and improvement need long-term and effective investment. Under the condition of limited government funds and shortage of community resources, the decision tree algorithm is used to predict the willingness of the elderly in the community. It can provide effective decision-making for the construction of community wisdom for the aged and avoid the unbalanced allocation and waste of resources.

Table 1. Age data of China's Seventh National Census Unit: person, %

age	population	proportion
total	1411778724	100.00
0 to 14 years old	253383938	17.95
15 to 59 years old	894376020	63.35
60 years old and above	264018766	18.70
Among them: 65 years old and above	190635280	13.50

Source: The Seventh National Population Census Bulletin (No. 5), National Bureau of Statistics,2021.05.11

2. Introduction to Decision Tree Algorithm

Decision tree is a basic classification method. It adopts a top-down recursive model, and is committed to deriving classification rules from irregular and disorderly data, and finally presenting a tree structure. It can be considered a collection of if-then rules. Every time the decision tree is split, the attribute value will be compared at the node to judge the direction of the next branch until the conclusion is reached at the leaf node. The final decision tree is a complete model and expression rules, one path corresponds to one rule [2].

Its main advantages are readable model and fast classification speed. It can generate the corresponding decision rule tree by learning the training data set, and get the importance of each feature by calculating the information gain, information gain rate and Gini coefficient, and then select the feature based on it to speed up the classification speed.The corresponding decision tree algorithm is ID3 algorithm, C4.5 algorithm and CART algorithm [3].

C4.5 algorithm is not only a classical algorithm for decision tree generation, but also an extension and optimization of ID3 algorithm.

C4.5 algorithm mainly makes the following improvements to ID3 algorithm:

1. The information gain rate is used to select the split attribute, which overcomes the shortcoming of ID3 algorithm which tends to select the attribute with multiple attribute values as the split attribute.
2. It can handle both discrete and continuous attribute types.
3. After the decision tree is constructed, it can be pruned.
4. It can process training data with missing attribute values.

Because the decision tree generated by C4.5 algorithm is easier to be interpreted, and the scale is relatively small. So, in this research, the prediction model is constructed by C4.5 algorithm.

3. Data Collection and Processing

There are two data sources for this study. One is to obtain desensitization data from the existing community smart pension system. The other is to use a self-made questionnaire on the willingness to provide for the elderly through face-to-face or online filling by their children. After data collection, redundant attributes and abnormal data were screened by comparison and experience, and missing data were completed. The continuous data is discretized.

3.1 Data preprocessing

First, the original data in the community smart pension system are processed as follows: name, contact information, home address information, native place, nursing level, payment method and other attributes

irrelevant to data mining are deleted. Remove the redundant attribute birth date. Continuous data such as age and monthly income are converted into discrete data. Merge the original unit type and occupation into the status attribute.

Delete the records with too many missing values, and fill the attributes with fewer missing values with hot card filling method [4] (find an object most similar to it in the complete data set and fill it with this value) to complete the missing attribute values, so as to obtain dataset 1.

For the questionnaire data, the data with contradictory attributes were modified or deleted according to experience, and dataset 2 was obtained.

Specification and integration of dataset 1 and dataset 2. Finally, a total of 1342 data sets of pension intention prediction data were obtained, including 8 characteristic variables (gender, age, education, income, health, marital status, living situation and smart device proficiency) and 1 decision variable (smart pension).

Reassign the characteristic variables as Table 2. The data samples after the assignment as Table 3.

Table 2. Feature assignment table

Feature	Assignment
gender	Male: 0; Female: 1.
age	65-69:1; 70-74:2; 75-79:3; 80 and above: 4
education	Illiteracy: 1; Primary school: 2; Junior high school: 3; Bachelor (junior college) or above: 4
income	No income: 1; Below 1000:2; 1000-3000:3; 3001-5000:4; More than 5000:5
health	Illness, completely unable to take care of themselves:1; Illness, half a self-care:2; Illness, can provide for oneself:3; Health:4
marital status	Married: 1; Married passed away: 2; Divorced: 3; No marriage: 4
living situation	Living alone: 1; Living with a married couple: 2; Living with spouse or children or adult grandchildren: 3; Living alone with minor grandchildren: 4
smart device proficiency	Very skilled: 1; Proficient: 2; General: 3; Unskilled: 4; Not at all: 5
smart pension	YES:Y; NO:N

Source: Self-made form according to sample characteristics

Table 3. The first 15 lines of sample data

ID	gender	age	education	income	health	marital status	living situation	smart device proficiency	smart pension
0001	0	4	4	5	1	1	3	4	Y
0002	0	4	4	5	2	2	1	4	Y
0003	1	4	4	5	2	1	2	4	Y
0004	1	2	4	5	4	1	3	2	N

0005	0	4	4	5	1	2	2	4	Y
0006	0	1	4	5	4	1	3	1	Y
0007	0	1	3	5	4	1	2	3	N
0008	0	2	3	4	3	2	1	4	N
0009	1	4	1	1	2	2	2	5	N
0010	1	4	4	5	3	1	2	3	Y
0011	1	4	1	1	3	2	1	4	N
0012	1	2	4	5	4	1	3	2	N
0013	0	4	4	5	1	1	3	4	Y
0014	0	4	4	5	2	2	1	4	Y
0015	1	4	4	5	2	1	2	4	Y

Source: Self-made table based on sample data

3.2 Feature analysis

In order to deeply explore the static characteristics and business characteristics of the elderly with and without pension intention, the obtained data set was visualized by distribution map. The distribution of each feature in retirement intention as Figure 1-8.

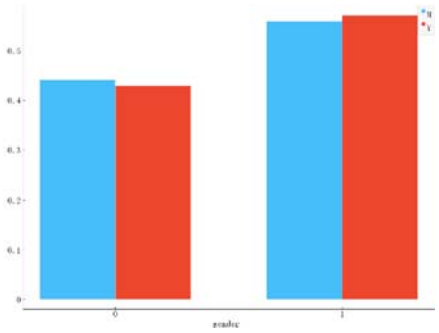

Figure 1. Distribution of 'gender' grouped by 'smart pension' (relative frequencies)

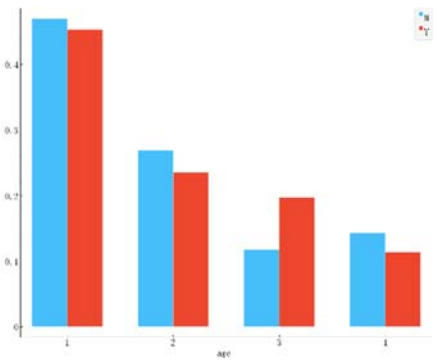

Figure 2. Distribution of 'age' grouped by 'smart pension' (relative frequencies)

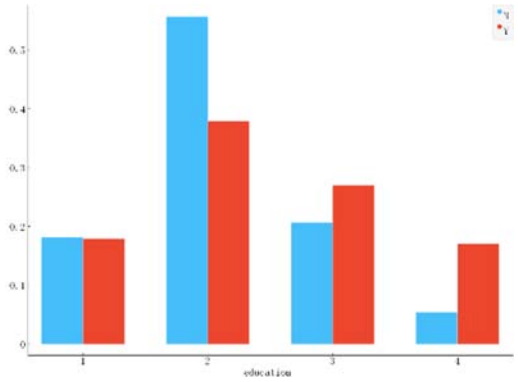

Figure 3. Distribution of 'education' grouped by 'smart pension' (relative frequencies)

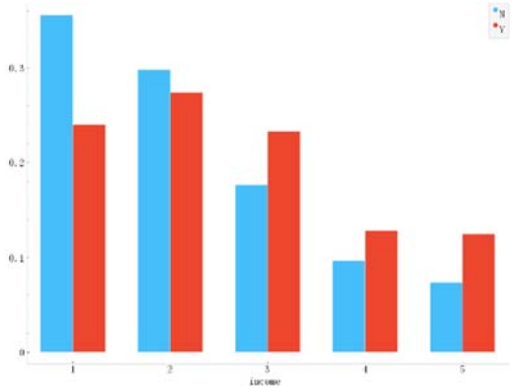

Figure 4. Distribution of 'income' grouped by 'smart pension' (relative frequencies)

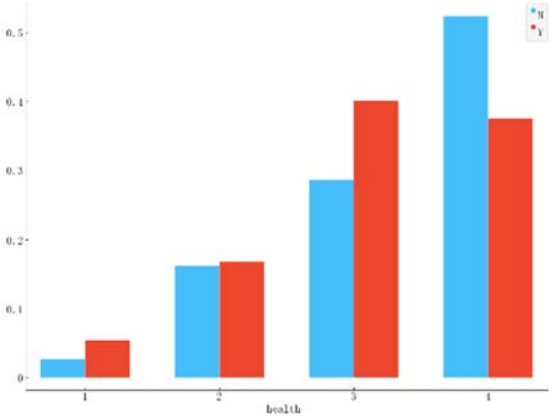

Figure 5. Distribution of 'health' grouped by 'smart pension' (relative frequencies)

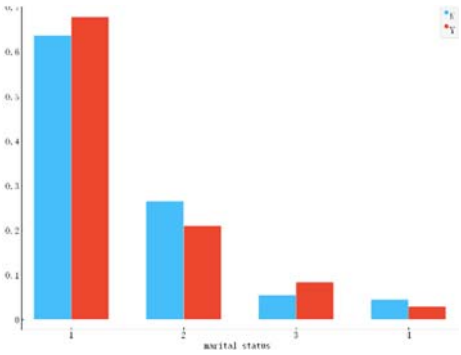

Figure 6. Distribution of 'marital status' grouped by 'smart pension' (relative frequencies)

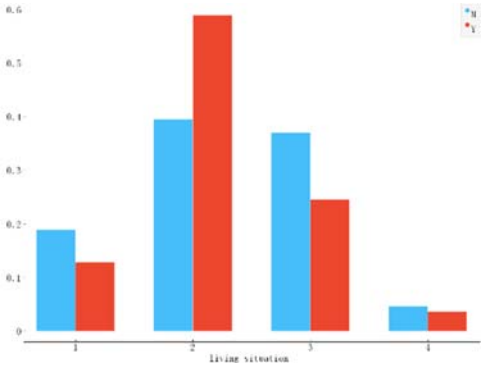

Figure 7. Distribution of 'living situation' grouped by 'smart pension' (relative frequencies)

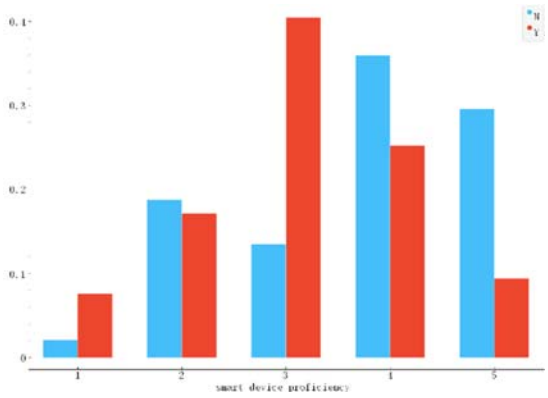

Figure 8. Distribution of ' smart device proficiency ' grouped by 'smart pension' (relative frequencies)

It can be seen from the distribution diagram that pension intention has a great degree of differentiation in five characteristics of smart device proficiency, education, living situation, health and income.

4. Establishment of Prediction Model

To construct prediction model with decision tree, feature selection is the first step. The C4.5 algorithm uses the info gain ratio to select features. The information gain ratio is defined on the basis of information entropy and information gain.

4.1. Information gain

The ID3 algorithm selects the feature with the highest gain as the test feature of the given data set by calculating the information gain g(D, A) of each feature. Create a node for the selected test feature and mark it with that feature. Create a branch for each value of the feature and divide the sample accordingly[5].

Information gain is the difference of information entropy before and after a feature partition data set.

g (D, A) is used to represent the information gain of feature A to sample set D. H(D) is used to represent information entropy. H(D|A) is used to represent conditional entropy.

$$H(D) = -\sum \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|} \quad (1)$$

$$H(D|A) = \sum_{i=1}^{n} p_i H(D|A = a_i) \quad (2)$$

$$g(D, A) = H(D) - H(D|A) \quad (3)$$

Among them, |Ck| means the number of samples of category K, |D| means the total number of samples.

According to Eq. (1)-(3), the information gain for each feature in the sample data can be calculated as Table 4. The top five features match the results of the histogram.

Table 4. Information gain

Feature	g(D,A)	Feature	g(D,A)
smart device proficiency	0.1027	income	0.0169
education	0.0355	age	0.0088
living situation	0.0262	marital status	0.0059
health	0.0182	gender	0.0001

Source: Calculated based on sample data

4.2. Info gain ratio

Info gain ratio gr(D,A)[6] takes number and size of branches into account when choosing an attribute, and corrects the information gain by taking the intrinsic information of a split into account (i.e. how much info do we need to tell which branch an instance belongs to).

$$g_r(D, A) = \frac{g(D,A)}{H_A(D)} \quad (4)$$

$H_A(D)$ indicates that for the sample set D, the entropy is obtained by taking the current feature A as a random variable (the value is the value of each feature A).

According to Eq. (4), the info gain ratio for each feature in the sample data can be calculated as Table 5.

Table 5. Info gain ration

Feature	$g_r(D,A)$	Feature	$g_r(D,A)$
smart device proficiency	0.0478	income	0.0077
education	0.0193	age	0.0048
living situation	0.0162	marital status	0.0045
health	0.0108	gender	0.0001

Source: Calculated based on sample data

The top five are also smart device professional, education, living situation, health and income.

4.3. Generate C4.5 decision tree

For intuitive representation of generated decision tree, the PMT tool was used in this research. The PMT tool is a big data analysis and mining tool for machine learning. It supports most data mining algorithms and has a user-friendly graphical interface.

The results of C4.5 algorithm decision tree visualization as Figure 9. In Figure 9, each node is composed of pension intention (Y/N) of more than 50%, proportion, quantity and feature. Since the information gain rate of "smart device proficiency" is the highest, the root node is first generated according to the old person's proficiency in using smart devices. The root node is marked as "smart device proficiency". "61.5%, 825/1342" means that 825 skilled users of smart devices choose "Y" in 1342 data sets, accounting for 61.5%. So this node is labeled "Y" and forms two branches, left and right. The second layer node does the same according to the feature "income", which ranks second in information gain rate.

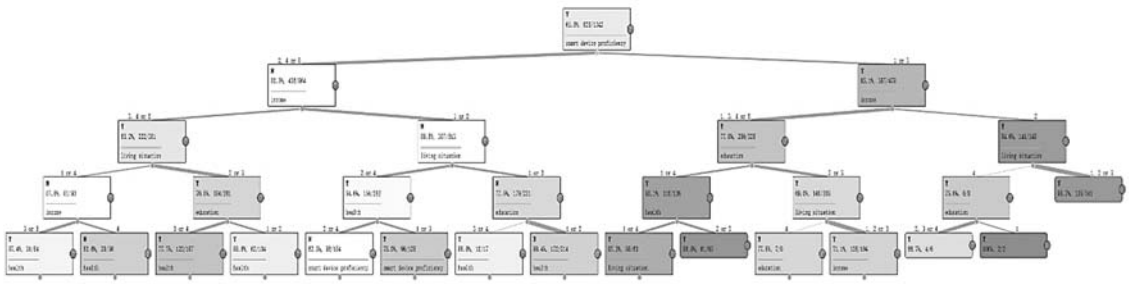

Figure 9. C4.5 decision tree visualization results

5. Model Training and Evaluation

5.1. Model training

To avoid overfitting problems, model training typically uses a cross-validation (CV) approach. CV is a statistical method used to verify the performance of taxonomies. The basic idea is to group the raw dataset, one part as a training set, and the other part as a validation set [7].

First, use the training set to train the classifier, and then use a set of validation set to test the trained model. It is used as an indicator to the performance of this classifier.

There are three common methods for cross-validation: hold-out cross-validation、k-fold cross-validation and leave-one-out cross-validation。

The hold-out cross-validation randomly divides the raw data into two sets, each as a training set and a test set, but the final results are closely related to the set segmentation ratio and the variation of the results may be large. This approach is difficult to achieve the goal of accurately evaluating the model when the total data set is not very large.

The leave-one-out cross-validation is based on the hypothesis that there are n objects in the total set, one object at a time is selected as a test set, and the rest as a training set. A total of n training sessions were conducted to take the mean as the final evaluation indicator. The leave-one-out cross-validation is relatively reliable, incorporating almost all objects into each model training, and taking longer to compute when the total set is large.

The k-fold cross-validation divides the raw data into k subset, each subset of which is tested, and the remaining k-1 subset is combined into a training set to perform k training. Take the average value for each evaluation indicator (sensitivity, specificity, AUC, etc.). This method can be predicted using all the samples in the dataset, and the average evaluation index reduces the impact of the odd training set and test set segmentation methods on the predicted results. The selection of k values also affects the final results, and studies have shown that larger k values increase the accuracy of evaluation, and that when k is 5 or 10, the overall performance is achieved under evaluation accuracy and computational complexity [8]. It's the best.

In this study, the model training was conducted using the 10-fold cross-validation.The basic steps for 10-fold cross-validation are as follows:

1.Divide the original data set into 10 samples as balanced as possible;

2.Use the first subset as a test set. The 2nd~9th subset is integrated as a training set.

3.Use the training set to train the model. Calculate the results of various evaluation indicators under the test set;

4.Repeat steps 2 through 3 to rotate the 2nd to 10th subset as a test set.

5.Calculate the average for each evaluation indicator as the final result.

The principle diagram of the 10-fold cross-validation as Figure 10.

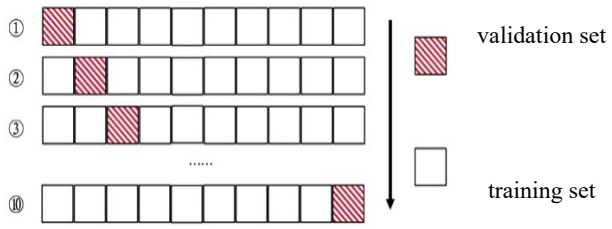

Figure 10. 10-fold cross-validation

5.2. Model evaluation

After using the classification model for sample prediction, it is necessary to use certain indicators to evaluate the performance of the classification model, so as to judge the authenticity, reliability and validity of the model.

The Confusion Matrix and ROC curve are usually used for model evaluation.

5.2.1. Confusion Matrix

Confusion matrix refers to the use of matrix form, the actual sample situation and model forecast situation for contingency table analysis. The rows of the matrix represent the real category of the sample, and the columns of the matrix represent the model prediction category [9]. Taking dichotomy as an example, the structure of confusion matrix as Figure 11.

	True (Predicted Class)	False (Predicted Class)
True (Actual Class)	True Positive (TP)	False Negative (FN)
False	False Positive	True Negative

(Actual Class)	(FP)	(TN)

Figure 11. Confusion Matrix

The sample data were classified under 10-fold cross-validation. According to the classification results of the prediction model and the real situation of the original data, the values of TP，FN，TN and FP were calculated [10]. Thus, the Confusion Matrix of the model is constructed, as Figure 12.

Confusion Matrix		Predicted class		
		Y	N	Σ
Actual class	Y	715	110	825
	N	84	433	517
	Σ	799	543	1342

Figure 12. Prediction Model Confusion Matrix

According to Table 7 and Eq. (5)-(8), Recall, Precision, Accuracy and F1 Score of the prediction model can be calculated. Based on the calculations, all four indicators performed well.

$$Recall = \frac{TP}{TP+FN} = \frac{715}{715+110} = 0.867 \quad (5)$$

$$Precision = \frac{TP}{TP+FP} = \frac{715}{715+84} = 0.895 \quad (6)$$

$$Accuracy = \frac{TP+TN}{P+N} = \frac{715+543}{825+517} = 0.855 \quad (7)$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} = 0.881 \quad (8)$$

5.2.2. ROC curve

The full name of ROC curve is Receiver Operating Characteristic curve, which is an evaluation indicator derived from Confusion Matrix. ROC curve is a coordinate graph composed of false positive rate (FPR) as the horizontal axis and true positive rate (TPR) as the vertical axis [9].

$$FPR = \frac{FP}{FP+TN} \quad (9)$$

$$PR = \frac{TP}{TP+FN} \quad (10)$$

The ROC curve of the prediction model as Figure 13. The larger the FPR value, the more the predicted positive classes are actually negative. The larger the TPR value, the more of the predicted positive classes are

actually positive. In other words, the horizontal axis means more wrong predictions, and the vertical axis means more right predictions.

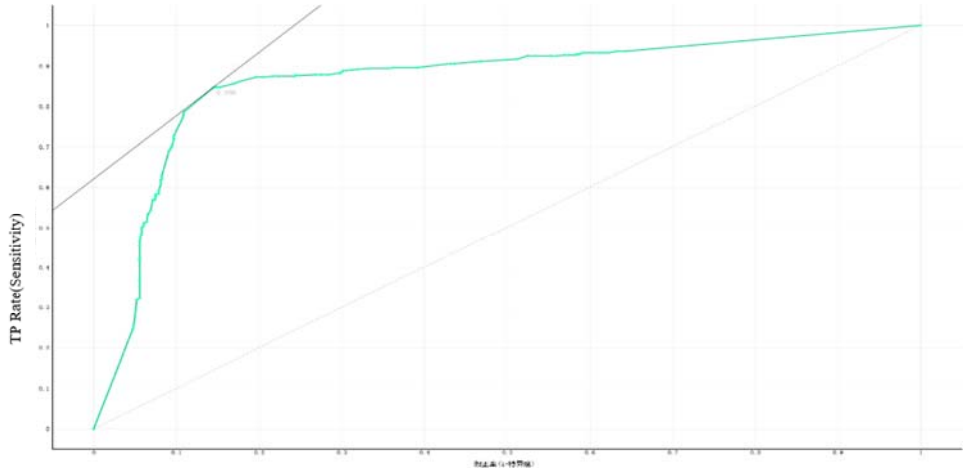

Figure 13. ROC curve

Therefore, the point in the upper left corner of the coordinate system (FPR = 0, TPR = 1) represents a perfect prediction and a non-categorical prediction error. The point above the diagonal (TPR > FPR) indicates that the classification prediction is generally correct. The point on the diagonal (TPR = FPR) indicates that each half of the predicted situation is incorrect. The point below the diagonal (TPR < FPR) indicates that the classification prediction is generally incorrect. The closer the ROC curve is to the upper left corner, the better the effectiveness of the separator.

At the same time, the model representation can be measured by calculating the area AUC value under ROC curve. If the AUC value is less than 0.5, the default judgment model is not useful.

Therefore, the AUC value is usually between [0.5,1]. The larger the AUC value, the higher the accuracy. Generally, AUC > 0.7 is recognized as having a good model effect [9].

5.2.3. AUC

The AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example (assuming 'positive' ranks higher than 'negative') [11].

$$AUC = \frac{\sum_{positives} k - \frac{n_{pos}(n_{pos}+1)}{2}}{n_{pos}n_{neg}} \quad (11)$$

According to Eq. (11), AUC = 0.868 > 0.7, So this model is a better model.

6. Conclusion

First of all, this paper introduces the problem of population aging in the world, and expounds the important significance of building smart pension community and pension intention prediction in China. Then, C4.5 algorithm was used to establish the prediction model, and the model was trained by the 10-fold cross-validation method. Finally, a prediction model of pension intention with high accuracy, precision, recall and AUC was obtained. Using this model can provide a certain decision for the government to establish the community smart pension, has a high application value.

Acknowledgements: We acknowledge the foundation of the 2021 Key Scientific Research Project of Neijiang Normal University (NO.2021ZD08): Application research of data mining technology in community smart pension system.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] J.Z. Ning, *The Seventh National Population Census Bulletin (No.5)*, National Bureau of Statistics, May 11,2021.

[2] Y. Qiang, J.Z. Lu, and B. Pei, "Cloud Scheduling Algorithm Based on the Decision Tree Classification," Joumal of Taiyuan University of Technology, Vol. 43, No. 6, pp.715-718, Nov. 2012, doi: http://cnki:sun:tygy.0. 2012-06-018.

[3] H. Li, *Statistical Learning Methods,* Tsinghua University Press, Beijing, 2012.

[4] W. Wang, *Semi-supervised Graph Learning with Missing Data*, MA thesis, South China University of Technology, Guangzhou, China, 2011.

[5] X.Y. Lin, "Comparative Study on Decision tree Algorithm in Data Mining," China Science and Technology Information, Vol. 17, No. 2, pp.94-95, Jan.1986, doi: http://cnki:sun:xxjk.0.2010-02-053.

[6] J.R. Quinlan, "Induction of Decision Trees," Machine Learning(1), Vol. 1, No. 2, pp.81-106,1986, doi: http://10.1007/BF00116251.

[7] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," 2016 IEEE 6th International Conference on Advanced Computing (IACC), pp. 78-83, 2016, doi: http://10.1109/IACC.2016.25.

[8] Z.C. Liang, etc., "Application of 10-fold cross-validation in the evaluation of generalization ability of prediction models and the realization in R," Chinese Joumal of Hospital Statistics, Vol.27, No.4, pp.289-292, Aug.2020, doi: http://CNKI:SUN:JTYY.0.2020-04-001.

[9] Y.M. Wang, *Study on Characteristic Genes of Pancreatic Cancer Classification Based on Multiple Data Sets*, MA thesis, Southwest University, Chongqing, China, 2020.

[10] Brett Lanta, *Machine learning with R*, China Machine Press, Beijing, 2016.

[11] Fawcett. Tom, "An introduction to ROC analysis," Pattern recognition letters, Vol.27, No.8, pp.861-874, Jun.2006, doi: http://10.1016/j.patrec.2005.10.010.