International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

A Machine Learning Model for Predicting Social Insured Behaviors of Migrants in China

Jing An[1], Jinlong An[2]
[1]School of Management, Nanjing University of Posts and Telecommunications, Nanjing, China
[2]First People's Hospital of Changshu City, Hospital Affiliated to Soochow University, Changshu, China

Date of Submission: 17-01-2023 Date of Acceptance: 27-01-2023

ABSTRACT: The method of machine learning has been widely used in social sciences. However, to date few researches have focused on the application of machine learning to social insured behaviors of migrants. This study explores a machine learning model for predicting social insured behaviors of migrants in China. The participants were 8200 young and middle-aged migrant workers in China. A feature subset of work injury insurance and unemployment insurance is selected. A decision tree prediction model of migrants' participation in work injury insurance and unemployment insurance is constructed respectively and decision prediction mechanisms are generated correspondingly. The results indicate that enterprise ownership, working days per week, workplace and employment type are predictors of migrants' participation in both work injury insurance and unemployment insurance. Differently, occupation is further a significant predictor of migrants' participation in work injury insurance, while education and household type are predictors of migrants' participation in unemployment insurance.

KEYWORDS: Social Insured Behavior, Migrants, Machine Learning, Prediction Model, Decision Tree.

I. INTRODUCTION

The social security of migrants is getting more and more attention while they are creating economic values for their destination areas. Migrant workers have more dangerous jobs and have a higher risk of total and fatal injury than natives[1-3], and employment industry (economic sector), occupation, income, job tenure, gender, safety training, work environment (exposure to physical and chemical harmful substances), personal health and nationality may be the risk factors of work-related injuries and illnesses[4-6]. Migrants in China had disproportionately high incidences of work-related injuries and deaths (69.82%) [2][4]. Specifically,electricians, safeguards and construction workers were at the highest risk of getting injured [1]. Straining/spraining (56%) was the most common occurrence and lifting (21.5%) was the leading contributing factor [7]. However, only 2.8% of them applied for compensation [7].

Although all employers are required to pay insurance for their employees according to the law, however, in actual fact only 21.8% of the migrants in China had access to employee work injury compensation [8]. Hukou, job contract, employment industry, regions, monthly income, age, education, employment ownership, occupation, workplace were associated with their insurance coverage and there were also gender differences [8-12]. The participation rate of work injury insurance in the transportation sector was higher than that of the service sector. Due to the low participation rate of work injury insurance, after living a while in hospital in floating city, injured rural migrants choose to return to their hometown for further treatment and recovery [8]. Usually most of them seek compensation from employers in informal ways, such as bargaining, negotiating, threatening and violence. However, the amount of compensation they received through this informal approach was significantly less than that from formal insurance compensation [8].

Similarly, there was also a higher unemployment incidence of migrants than that of natives[13-17], and education, employment industry, income, nationality and ethnicity were contributing factors[15,17,18-22]. 20% of these migrants were unemployed during the economic recession [17]. In order to compromise with family life, these migrants had to choose to change work departments and types of work in addition to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

returning to their home country or circling between home and destination countries[14]. Additionally, youth unemployment became an arising social problem [13], approximately 50% of them had mental health problems and depression was the most common symptom[23].

Unemployment insurance guarantees the loss of income caused by involuntary unemployment. However, not every unemployed migrant worker could get the corresponding benefit. Income, education, employment industry, age and nationality were influencing factors[9,17,20,24]. It was difficult for low-income unemployed migrants to enjoy unemployment insurance benefits [24]. The participation rate of social insurance in individually-owned business was lower than that in other companies. Participation ratio of unemployment insurance and work injury insurance in foreign or joint ventures was relatively high. Migrants working in manufacturing had higher social insurance participation than those in other industries [9].

Former research on influencing factors of migrants' social insured behaviors, based on literature review, generated hypotheses and then constructed logistic regression models to test the hypotheses [10,17,20,24]. Or, based on the literature review, logistic regression models were constructed directly to search for factors associated with insurance coverage [9,11,12]. These researches analyzed the impact of each independent factor such as education, occupation or workplace on migrants' work injury insurance or unemployment insurance. However, there is no in-depth analysis of correlations between these influencing factors. Additionally, differences between influencing factors for participation in different insurances are not clearly explained. For example, how the factors influencing participation rate of work injury insurance differ from factors influencing that of unemployment insurance has not been discussed clearly. Moreover, the hypothesis or established influential factor model is susceptible to subjective understanding of researchers. It is the advantage of machine learning algorithm that it can accurately select the factors that have influence on participation in social insurance of migrants and construct corresponding prediction models.

Machine learning can be described as the development of computational techniques on learning as well as the construction of systems capable of acquiring knowledge automatically [25]. As one of today's most rapidly growing technical fields, it lies at the core of artificial intelligence and data science. The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce, leading to more evidence-based decision-making across many walks of life, including health care, manufacturing, education, financial modeling, policing, marketing, energy andenvironment[26-37].

In addition to the above fields, data processing technology of machine learning has broader application in social sciences. Many methods such as decision trees, dimension reduction methods, nearest neighbor algorithms, support vector models and penalized regression demonstrated outstanding performance in solving social problems [38]. They could be used to select features of diversified census data [39], to classify users of social websites [40], to predict population fluctuation of the virtual world [41], to analyze cognitive level of teachers' classroom questions[42], and to construct prediction risk model preventing child abuse incidents [43], or they could be used for crisis management using social media data [44], for organization communication of social media [45]. As well as for research on social behaviors [46], and for coding of political problems[47], etc.

What we can learn from the above is that machine learning has been widely used in social sciences. However, till date limited research has focused on the application of machine learning to social insured behaviors of migrants. The sample data of this study includes 35 different attributes of basic information, insurance information, migratory condition, occupation and financial condition of migrants. Machine learning algorithm is a kind of algorithm that automatically analyzes and obtains rules from the data, and uses these rules to predict the unknown data. Therefore, this study tries to select accurately factors that have influence on migrants' participation in work injury insurance and unemployment insurance respectively from those 35 attributes, and to construct a corresponding prediction model respectively. This study will provide further evidence for the application of machine learning to social sciences. Social protection for migrant workers will have macroeconomic benefits in the long run because social protection has a positive impact on the production of human potential [12]. Furthermore, this study will provide reference for the government to formulate a social security system appropriate for migrants, to protect their rights and interests in social insurance, so as to promote economic development of floating cities and countries.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

II. METHODS

The dataset in this study was collected by the National Health and Family Planning Commission (NHFPC) of China in 2010. All participants had signed an informed consent form before filling in the survey. It included 8200 samples with a total of 35 attributes. These attributes covered basic information, migratory information, insurance information, occupation and financial information of migrants. This group mainly came from 32 provinces, cities and districts in China as their household registered places such as Beijing, Shanghai, Guangdong, Jiangsu, Liaoning, Xinjiang, Guizhou, Ningxia, Xizang, Taiwan, etc.

Firstly, attributes of work injury insurance and unemployment insurance are selected respectively as a feature subset, which requires strong correlation with the target attribute, while there is no strong correlation between each selected attribute. With the help of attribute evaluator and search algorithm, different subsets of the best attributes are selected, and then the same classifier and ten-fold cross validation are used to compare influence of the attribute subsets selected by different algorithms on classification accuracy. It is found that classification accuracy is improved after attribute selection.

Secondly, based on the attributes selected above, the decision tree prediction model of migrants' participation in work injury insurance and unemployment insurance is constructed respectively. Decision tree learning adopts a tree structure to establish a decision model based on the features of the data. A training set is used to train decision tree algorithm and a decision tree model is obtained. When this decision tree model is used to determine the classification of an unknown sample (category is unknown), it starts from the root node of decision tree and searches from top to bottom until it reaches a leaf node along a branch. Classification label of this leaf node is the category of this unknown sample, which predicts classification of unknown samples. Information entropy in the dataset is calculated as Entropy(S).

$$\text{Entropy}(S) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

S is a training dataset. M represents classification number. And Pi is the proportion of each category of the dataset in total samples. If characteristic A is chosen as judgement node of a decision tree, after its function on this tree, information entropy becomes as EntropyA(S).

$$Entropy_A(S) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} Entrop(S)_i$$

K indicates that sample S is divided into K parts. Gain(S,A) represents information gain of dataset S divided by feature A, which is calculated as the value of EntropyA(S) subtracted from Entropy(S).

$$\text{Gain}(S, A) = \text{Entropy}(S) - \text{Entropy}_A(S)$$

Information gain is adjusted by splitting information of introduced attributes SplitE(A).

$$\text{SplitE}(A) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

As criterion of attribute selection for a decision tree, information gain ratio GainRatio(A) takes splitting information as denominator.

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitE}(A)}$$

Attribute selection is implemented through software Weka 3.8. The same evaluator and search algorithm (BestFirst and CfsSubsetEval) is chosen to select the attribute subset for work injury insurance and unemployment insurance. Software SPSS 22.0 is used to construct the decision tree prediction model for work injury insurance and unemployment insurance of migrants respectively and for other statistical analysis as well. Figures are mainly conducted by R and Weka.

III. RESULTS

Table 1 shows a description of some of the 35 attributes. Among these migrants, male migrants accounted for 46.94% and females accounted for 53.06%. The average age was 33 years (SD=8.684). For education levels, 53.12% of this population completed secondary school and 3.62% held a degree of Bachelor or Master. In terms of household type, 85.99% held an agricultural household register. 62.24% of them worked in current city for less than 3 years. Their family monthly income was no more than 24002.4RMB (99.29%) and they on average have one child (44.70%, SD=0.658). 45.37% of these participants worked seven days per week and 41.17% worked 10-12 hours per day. In general, work injury insurance has a higher participation ratio (24.38%) than that of unemployment insurance (13.63%).

The result of attribute selection shows that five demographic attributes - occupation, enterprise ownership, working days per week, workplace and employment type - are closely related to migrants' participation in work-related injury insurance. For unemployment insurance, it indicates that

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

education, household type, enterprise ownership, working days per week, workplace and employment type, all six attributes have high correlation with migrants' participation.

Work Injury Insurance

The dataset of work injury insurance of migrants is classified into two categories (work injury insurance and no work injury insurance). Confusion matrix shows that 5557 instances whose real category are "No work injury insurance (No WII)" have been correctly predicted, and 1158 instances whose true category are "Work injury insurance (WII)" have been correctly predicted. Overall the classification accuracy is 81.90% and the value of ROC area is 0.8482 as shown in Figure 1-a, which indicates that this model has a very well classification result. Based on these selected five features, a decision tree prediction model of migrants' participation in WII is constructed, indicating correlations between these attributes and migrants'participation in this kind of insurance, as shown in Figure 2.

The decision tree has 33 leaf nodes (Node 0 - Node 32). The number of terminal nodes is 21. As shown in this decision tree classification prediction model, occupation (now_vocation), enterprise ownership (unitquality), working days per week (aver_days), workplace (workplace) and employment type (emly_ident) all have significant prediction ofmigrants' participation in WII (now_III), and decision prediction rules are generated correspondingly.

As the root node of this decision tree, enterprise ownership has a significant impact on the migrants' participation in WII ($p < .001$). Migrant workers employed in different ownership enterprises will go along different branches of this decision tree.

Only 7.2% of migrant workers employed in individually-owned business have access to WII. Their participation in WII is further significantly affected by occupation ($p < .001$) (see Figure 3). (1) If they are a service staff or a business staff, up to 95% of them do not have WII. Their detailed participation in WII is further significantly influenced by employment type ($p < .001$). If they work as an employee, the participation rate is 10.4%. If they are a self-employed worker, a family helper or an employer, up to 97.1% of them do not have WII. (2) If they are staff and related personnel, or production worker/transportation worker and related personnel, the participation rate is 32.2%. Their participation is further significantly affected by working days per week ($p < .001$). If they every week work one day or six days, 55.8% of them have WII. If they work two days, four days, five days or seven days per week, the proportion of migrants who have WII is 18.7%. (3) If they are a professional and technical staff, an irregular employed employee, or an agriculture and water conservancy industry production staff, the proportion of migrants who have WII is 11.0%. Their participation is further significantly influenced by workplace ($p < .001$). If they work in an indoor business place or at home, only 3.2% of them participate in WII. If their workplaces are office, outdoors, workshops or other places, 21.6% of them participate in WII.

As for the migrants working in collective enterprises, private enterprises or government institutions, the participation rate of WII is 40.9%. Workplace has further significant influence on their detailed participation in WII ($p < .001$). (1) If they work in indoor business places or work at home, 29.6% of them participate in WII. Their detailed participation is further significantly affected by working days per week ($p < .001$). If they each week work five days, four days or two days, 44.9% of them have access to WII. If migrants work six days every week or three days, the proportion of migrants who have WII is 30.0%. If they each week work seven days, approximately 81.5% of them do not participate in WII. (2) Over half of the migrants (58.5%) working in offices have access to WII. (3) If migrants work outdoors or at other places, 39.7% of them participate in WII. Their detailed participation is further significantly affected by enterprise ownership ($p < .05$). If they work in collective enterprises or government institutions, 55.7% of them have WII. If they work in private enterprises, the proportion of migrants who have WII is 37.1%. (4) 46.5% of migrants working in workshops participate in WII. Their participation is further significantly influenced by weekly working days ($p < .001$). 57.2% of them who each week work between four to six days have WII. If they work one day, seven days or zero per week, 69.4% of them do not have WII.

For migrants working in sino-foreign joint ventures, state-owned and state holding enterprises or Japanese and Korean funded enterprises, their participation ratio of WII is 65.5%. Their detailed participation is further significantly affected by weekly working days ($p < .001$) (see Figure 4). (1) 71.4% of migrants working between four to six days per week participate in WII. Their participation is further significantly influenced by occupation ($p < .05$). If they are service staff, business staff or agriculture and water conservancy industry production staff, 54.6% of them have WII. If they are staff and related personnel, professional

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

and technicalpersonnel production workers/transportation workers and related personnel, irregular employed employees or state and social managers, 76.6% of them participate in WII. (2) The participation rate for the migrants weekly working one day, seven days, three days or zero is 37.9%.

The participation rate in WII of migrants working in other ownership enterprises is only 1.3% while up to 98.7% of them do not have WII. Detailed participation of these people is further significantly influenced by workplace ($p < .001$). If they work in indoor business places, offices, outdoors or at home, 13.1% of them have WII. If they work in other places, up to 100% of them participate in WII.

Considering migrants working in European and American enterprises or Hong Kong, Macao and Taiwan enterprises, 82.4% of them have WII. And 21.8% of the migrants working in a land contractor company have access to WII.

Unemployment Insurance

Based on the above selected six attributes, a decision tree prediction model of migrants' participation in insurance is constructed, indicating correlations between these attributes and migrants' participation in this kind of insurance, as shown in Figure 5. The dataset of unemployment insurance of migrants is classified into two categories (unemployment insurance and no unemployment insurance). Confusion matrix shows that 6856 instances whose real category are "No unemployment insurance (No UI)" have been correctly predicted, and 442 instances whose true category are "Unemployment insurance (UI)" have been correctly predicted. Overall classification accuracy is 89.0% and the value of ROC area is 0.8513 as shown in Figure 1-b, this indicates that this model has a very good classification result.

There are 38 leaf nodes of this decision tree namely Node 0 - Node 37. Thickness of this tree is 3 and number of terminal nodes is 24. As shown in this decision tree classification prediction model, education (edu_status_1), household type (acc_na_1) , enterprise ownership (unitquality), working days per week (aver_days), workplace (workplace) and employment type (emly_ident) all are significant predictors of the migrants' participation in UI (UI), and decision prediction mechanisms are generated correspondingly.

As root node of this decision tree, enterprise ownership has a significant influence on migrants' participation in UI ($p < .001$). Migrants employed in different ownership enterprises will go to different leaf nodes along branches of this decision tree.

3.2% of migrants working in an individually-owned business or a land contractor company participate in UI, while as high as 96.8% do not. Their participation is further significantly affected by weekly working days ($p < .001$). (1) If they every week work five or six days, 9.3% of them have UI. Their participation is further significantly impacted by employment type ($p < .05$). If they work as an employee or an employer, 11.7% of them participate in UI. If they are a self-employed worker or a family helper, 4.0% of them have access to UI. (2) For the migrants working between one to four days or seven days per week, up to 98.7% of them do not have UI. Their participation is further significantly influenced by workplace ($p < .001$). If they work in indoor business places, outdoors or at home, the participation rate is only 1.1%. If they work in offices, workshops or other places, approximately 93.7% of them do not have UI.

For the migrants working in collective enterprises or private enterprises, 20.6% of them participate in UI. Their participation is further significantly impacted by workplace ($p < .001$). (1) If they work in indoor business places, the participation rate is 19.5%. Their participation is still further significantly influenced by weekly working days ($p < .001$). If they work five days, three days or two days per week, 36.6% of them have UI. If they each week work six days, the participation rate is 20.3%. If they work seven days or four days per week, 8.7% of them have UI. (2) For migrants working in offices, 51.8% of them have access to UI. Their participation is still further significantly influenced by household type ($p < .001$). If migrants hold a non-agricultural household register, 65.9% of them have UI. If they hold an agricultural one, the participation rate is 38.9%. (3) If they work outdoors, in workshops, at home or in other places, 13.3% of them have UI. Their participation is still further significantly influenced by weekly working days ($p < .001$). If they each week work five days, 25.3% of them participate in UI. If they work six days per week, the participation ratio is 15.2%. If they weekly work one day, seven days, three days, four days or zero, only 7.9% of them have UI while up to 92.1% do not.

The case for migrants working in sino-foreign joint ventures or government institutions is that 41.9% of them participate in UI. Their participation is still further significantly affected by education ($p < .001$) as shown in Figure 6. (1) If they have completed high school or technical

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

secondary school, 37.4% of them have UI. (2) For migrant workers who hold a degree of Bachelor or Master or have completed junior college, 66.7% of them participate in UI. (3) If the migrants have completed secondary school or primary school, 25.5% of them have access to UI. Their participation is still further significantly influenced by workplace ($p < .05$). If they work in indoor business places or offices, 41.1% of them have UI. If they work outdoors, in workshops, at home or in other places, the participation rate is 16.5% while 83.5% of them do not have access to this kind of insurance.

As for migrants employed in state-owned and state holding enterprises, their participation rate of UI is 29.3%. Detailed participation of these people is still further significantly influenced by household type ($p < .001$). (1) If migrants hold a non-agricultural household register, 70.2% of them have access to UI. (2) If they hold an agricultural household register, only 16.5% of them participate in UI. Their participation is still further significantly affected by education ($p < .05$). If they have completed high school, junior college, technical secondary school or hold a degree of Bachelor, 32.3% of them have access to UI. If they have completed primary school or secondary school or no schooling at all, up to 91.7% of them do not have UI.

In relation to migrants employed in other ownership enterprises, up to 99.3% of them do not have access to UI. Their participation is still further significantly impacted by workplace ($p < .001$). If they work in indoor business places, offices, outdoors or at home, 7.1% of them participate in UI. If they work in other places, up to 100% of them do not have access to this kind of insurance.

For the migrants working in European and American enterprises or Japanese and Korean funded enterprises, participation ratio is 77.6%. For migrants working in Hong Kong, Macao and Taiwan enterprises, 60.1% of them participate in UI. Their participation is still further significantly influenced by education ($p < .05$). If they are holding a degree of high school, secondary school, primary school or Bachelor, 48.0% of them do not have access to UI. If they have graduated from a junior college, technical secondary school or hold a degree of Master, approximately 86.5% of them have UI.

IV. CONCLUSION AND DISCUSSION

This study mainly investigated young and middle-aged migrants in China in 2010, employing the method of machine learning to analyze their participation in work-related injury insurance and unemployment insurance respectively. Based on the dataset of 8200 participants, this study constructs decision tree prediction models, classifying migrants' participation in work injury insurance and unemployment insurance, exploring influencing factors and their correlations contributing to it, and generating decision prediction mechanisms correspondingly. In general, it illustrates that the method of machine learning is an appropriate instrument to use for the purpose of predicting migrants' participation in social insurance in China.

The results of feature selection implemented by the method of machine learning indicate that occupation, enterprise ownership, working days per week, workplace and employment type are closely related to migrants' participation in work injury insurance, while education, household type, enterprise ownership, working days per week, workplace and employment type have high correlation with the migrants' participation in unemployment insurance.

A decision tree prediction model for migrants' participation in work injury insurance and unemployment insurance are constructed respectively based on these extracted features. What this study discovers is that occupation, enterprise ownership, working days per week, workplace and employment type all have significant prediction of migrants' participation in work injury insurance, while education, household type, enterprise ownership, working days per week, workplace and employment type are predictors of migrants' participation in unemployment insurance. The corresponding decision prediction rules are generated as well. The results show that classification accuracy of decision tree prediction models for two kinds of insurance is over 80%, which indicates that overall classification effect is satisfactory. Furthermore, it also illustrates that the method of machine learning has a very positive applicability in classification prediction of migrants' participation in work injury insurance and unemployment insurance in China.

Furthermore, what this study further discovers is that enterprise ownership, working days per week, workplace and employment type are common predictors of migrants' participation in both work injury insurance and unemployment insurance, with enterprise ownership as root node of two decision trees, which indicates that enterprise ownership is playing a crucial role in the migrants' participation in both kinds of insurances. Specifically, (1) in terms of work injury insurance, the participation rate of migrants working in European and American enterprises or Hong Kong,

Macao and Taiwan enterprises is the highest. 82.4% of them participate in work injury insurance and their participation is not influenced by other factors. There is the second highest participation rate (65.5%) of migrants employed in sino-foreign joint ventures, state-owned and state holding enterprises or Japanese and Korean funded enterprises. For migrants working in other six kinds of ownership enterprises, their participation rate is less than 50%, with the lowest participation rate (less than 8%) happening in individually-owned businesses or other ownership companies. (2) For unemployment insurance, the highest participation rate (77.6%) happens in European and American enterprises or Japanese and Korean funded enterprises and the participation of migrants working in these enterprises is not affected by other factors. There is the second highest participation rate (60.1%) of migrants employed in Hong Kong, Macao and Taiwan enterprises. For migrants employed in other eight kinds of ownership enterprises, their participation rate is less than 50%, with the lowest participation rate (less than 4%) happening in individually-owned businesses, land contractor companies or other ownership companies.

In addition to the above common prediction of migrants' participation in both work injury insurance and unemployment insurance, what is different, migrants' participation in work injury insurance is still further significantly influenced by occupation, while education and household type have significant prediction of migrants' participation in unemployment insurance. In detail, (1) if the migrants work is in an individually-owned business, or they work in sino-foreign joint ventures, state-owned and state holding enterprises or Japanese and Korean funded enterprises and weekly work between four to six days, their participation in work injury insurance is still further significantly influenced by occupation. (2) For migrants who work in Hong Kong, Macao and Taiwan enterprises, work in sino-foreign joint ventures or government institutions, or work in state-owned and state holding enterprises and hold an agricultural household register, their participation in unemployment insurance is still further significantly impacted by education. (3) For migrants who work in offices in collective enterprises or private enterprises, or migrants who are employed in state-owned and state holding enterprises, their participation in unemployment insurance is still further significantly influenced by household type.

There are also some limitations of this study. On the one hand, optimal feature subset is selected by attribute evaluator and search algorithm (BestFirst and CfsSubsetEval). Different feature selection algorithms can be explored in future studies to improve the classification accuracy of prediction model for migrants' participation in work injury insurance and unemployment insurance. On the other hand, although up to now the method of machine learning has been extensively used in social sciences, few researches have focused on the application of this method to social insured behaviors of migrant workers. This study explores two kinds of social insurances (work injury insurance and unemployment insurance) of the migrants in China with the application of machine learning method, which indicates a satisfactory result in classification prediction of the migrants' participation in work injury insurance and unemployment insurance in China. However, whether the machine learning method is suitable for prediction of migrants' participation in other social insurances such as endowment insurance, medical insurance, housing accumulation fund or maternity insurance still needs further exploration in the future.

Funding

This research was financially supported by the National Natural Science Foundation of China (NSFC) (71904019), Jiangsu Social Science Foundation (21GLB014), Suzhou Science and Technology Bureau (SYSD2019196) and a project of Nanjing University of Posts and Telecommunications (NYY221010).

Acknowledgments

The authors thank the National Health and Family Planning Commission (NHFPC) of China who contribute to data collection for this research.

REFERENCES

[1]. Xia, Q. H., Jiang, Y., Yin, N., Hu, J., and Niu, C. J. (2012) 'Injury among migrant workers in Changning district, Shanghai, China', International Journal of Injury Control and Safety Promotion, 19(1), pp. 81-85.

[2]. Fitzgerald, S., Chen, X., Qu, H., andSheff, M. G.(2013)'Occupational injury among migrant workers in China: A systematic review', Injury Prevention Journal of the International Society for Child and Adolescent Injury Prevention, 19(5), pp. 348-354.

[3]. You, S. F., and Wong, Y. F. (2015)'Explaining occupational injury rates between migrant and native workers

in Taiwan, 1998-2011', Asian and Pacific Migration Journal, 24(4), pp. 512-539.

[4]. Zhang, Q. (2012)'Occupational injury occurrence and related risk factors among Chinese migrant workers', Procedia Engineering, 43, pp. 76-81.

[5]. Lee, H., Chae, D., Yi, K. H., Im, S., and Cho, S. H. (2015)'Multiple risk factors for work-related injuries and illnesses in Korean-Chinese migrant workers', Workplace Health and Safety, 63(1), pp. 18-26.

[6]. Giraudo, M., Bena, A., and Costa, G. (2017)'Migrant workers in Italy: an analysis of injury risk taking into account occupational characteristics and job tenure', BMC Public Health, 17(1), pp. 1-9.

[7]. Scribani, M., Sherry Wyckoff, B. A., Jenkins, P., Bauer, H., and Earle-Richardson, G. (2013)'Migrant and seasonal crop worker injury and illness across the northeast', American Journal of Industrial Medicine,56(8), pp. 845-855.

[8]. Sun, L., and Liu, T. (2014)'Injured but not entitled to legal insurance compensation – ornamental institutions and migrant workers' informal channels in China', Social Policy and Administration, 48(7), pp. 905–922.

[9]. Gao, Q., Yang, S., and Li, S. (2012)'Labor contracts and social insurance participation among migrant workers in China', China Economic Review,23(4), pp. 1195-1205.

[10]. Cheng, Z., Nielsen, I., and Smyth, R. (2014) 'Access to social insurance in urban China: A comparative study of rural–urban and urban–urban migrants in Beijing', Habitat International, 41, pp. 243-252.

[11]. Zhao, Y., Kang, B., Liu, Y., Li, Y., Shi, G., and Shen, T., et al. (2014)'Health insurance coverage and its impact on medical cost: Observations from the floating population in China', Plos One, 9(11), pp. 1-9.

[12]. Yao, J., and Kim, B. (2015)'Social insurance participation of rural migrant workers based on gender dimension: Evidence from four Chinese cities', Asian Social Work and Policy Review, 9(1), pp. 57-69.

[13]. Helgesson, M., Johansson, B., Nordqvist, T., Lundberg, I., and Vingård, E. (2012)'Unemployment at a young age and later sickness absence, disability pension and death in native Swedes and immigrants', The European Journal of Public Health, 23(4), pp. 606-610.

[14]. Maroukis, T. (2013)'Economic crisis and migrants' employment: A view from Greece in comparative perspective', Policy Studies, 34(2), pp. 221-237.

[15]. Uhlendorff, A., and Zimmermann, K. F. (2014) 'Unemployment dynamics among migrants and natives', Economica, 81(322), pp. 348–367.

[16]. Viasu, I. (2014)'Migrant labor, unemployment and optimal growth', Computational Methods in Social Sciences, 2, pp. 30-38.

[17]. Laird, J. (2015)'Unemployment among Mexican immigrant men in the United States, 2003-2012', Social Science Research, 49, pp. 202-216.

[18]. Giulietti, C., Guzi, M., Kahanec, M., and Zimmermann, K. F. (2013)'Unemployment benefits and immigration: Evidence from the EU', International Journal of Manpower, 34(1), pp. 24-38.

[19]. Bratsberg, B., Raaum, O., and Røed, K. (2014) 'Immigrants, labour market performance and social insurance', Economic Journal, 124(580), pp. 644–683.

[20]. Cebollaboado, H., Miyarbusto, M., and Muñozcomet, J. (2015)'Is the Spanish recession increasing inequality? Male migrant-native differences in educational returns against unemployment', Journal of Ethnic and Migration Studies, 41(5), pp. 710-728.

[21]. Flick, U., Hans, B., Hirseland, A., Rasche, S., and Röhnsch, G. (2017) 'Migration, unemployment, and lifeworld: Challenges for a new critical qualitative inquiry in migration', Qualitative Inquiry, 23(1), pp. 77-88.

[22]. Madsen, J. B., and Andric, S. (2017)'The immigration–unemployment nexus: Do education and protestantism matter', Oxford Economic Papers, 69(1), pp. 1-24.

[23]. Chen, L., Li, W., He, J., Wu, L., Yan, Z., and Tang, W. (2012) 'Mental health, duration of unemployment, and coping strategy: a cross-sectional study of unemployed migrant workers in eastern

China during the economic crisis', BMC Public Health, 12(1), pp. 1-12.

[24]. Gould-Werth, A., and Shaefer, H. L. (2013)'Do alternative base periods increase unemployment insurance receipt among low-educated unemployed workers', Journal of Policy Analysis and Management, 32(4), pp. 835–852.

[25]. Mitchell, T. M. (1997)'Machine learning', Burr Ridge, IL: McGraw Hill, 45(37), pp. 870-877.

[26]. Chao, S., Cheng, C., and Liew, C. C. (2015)'Mining the dynamic genome: A method for identifying multiple disease signatures using quantitative rna expression analysis of a single blood sample', Microarrays, 4(4), pp. 671-689.

[27]. Jordan, M. I., and Mitchell, T. M. (2015)'Machine learning: trends, perspectives, and prospects', Science, 349(6245), pp. 255-260.

[28]. Luo, W., Nguyen, T., Nichols, M., Tran, T., Rana, S., andGupta, S., et al. (2015) 'Is demography destiny? Application of machine learning techniques to accurately predict population health outcomes from a minimal demographic dataset', PlosOne, 10(5), pp. 1-13.

[29]. Yadav, A. K., and Chandel, S. S. (2015)'Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of Weka in ANN based prediction model', Renewable Energy, 75, pp. 675-693.

[30]. Babichev, S. A., Kornelyuk, A. I., Lytvynenko, V. I., and Osypenko, V. V. (2016) 'Computational analysis of microarray gene expression profiles of lung cancer', Biopolymers and Cell, 32(1), pp. 70-79.

[31]. Irina Ioniță, and Liviu Ioniță. (2016) 'Prediction of thyroid disease using data mining techniques. BRAIN', Broad Research in Artificial Intelligence and Neuroscience, 7(3), pp. 115-124.

[32]. Fu, C. W., and Lin, T. H. (2017)'Predicting the metabolic sites by flavin-containing monooxygenase on drug molecules using svm classification on computed quantum mechanics and circular fingerprints molecular descriptors', Plos One, 12(1), pp. 1-20.

[33]. Lim, S., Tucker, C. S., and Kumara, S. (2017)'An unsupervised machine learning model for discovering latent infectious diseases using social media data', Journal of biomedical informatics, 66, pp. 82-94.

[34]. Mullainathan, S., and Spiess, J. (2017)'Machine learning: An applied econometric approach', Journal of Economic Perspectives, 31(2), pp. 87-106.

[35]. Pan, I., Nolan, L. B., Brown, R. R., Khan, R., van der Boor, P., andHarris, D. G., et al. (2017) 'Machine learning for social services: A study of prenatal case management in Illinois', American Journal of Public Health, 107(6), pp. 938-944.

[36]. Salvador, R., Radua, J., Canales-Rodríguez, E. J., Solanes, A., Sarró, S., and Goikolea, J. M., et al. (2017) 'Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis', PlosOne, 12(4), pp. 1-24.

[37]. Saxe, G. N., Ma, S., Ren, J., and Aliferis, C. (2017) 'Machine learning methods to predict child posttraumatic stress: A proof of concept study', BMC psychiatry, 17(1), pp. 1-13.

[38]. Hindman, M. (2015)'Building better models: Prediction, replication, and machine learning in the social sciences', The ANNALS of the American Academy of Political and Social Science, 659(1), pp. 48-62.

[39]. Shanmuganathan, S., and Li, Y. (2016)'An AI based approach to multiple census data analysis for feature selection', Journal of Intelligent and Fuzzy Systems, 31(2), pp. 859-872.

[40]. De Lima, B. V. A., Machado, V. P., and Lopes, L. A. (2015) 'Automatic labeling of social network users Scientia. Net through the machine learning supervised application', Social Network Analysis and Mining, 5(1), pp. 1-10.

[41]. Kim, Y. B., Park, N., Zhang, Q., Kim, J. G., Kang, S. J., and Kim, C. H. (2016)'Predicting virtual world user population fluctuations with deep learning', PlosOne, 11(12), pp. 1-12.

[42]. Yahya, A. A., Osman, A., Taleb, A., and Alattab, A. A. (2013)'Analyzing the cognitive level of classroom questions using machine learning techniques', Procedia-Social and Behavioral Sciences, 97, pp. 587-595.

[43]. Gillingham, P. (2015)'Predictive risk modelling to prevent child maltreatment

and other adverse outcomes for service users: Inside the 'Black Box'of machine learning', The British Journal of Social Work, 46(4), pp. 1044-1058.

[44]. Lanfranchi, V. (2017)'Machine Learning and Social Media in Crisis Management: Agility vs Ethics', In Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management. IMT Mines Albi-Carmaux (École Mines-Télécom).

[45]. Van Zoonen, W., and Toni, G. L. A. (2016) 'Social media research: The application of supervised machine learning in organizational communication research', Computers in Human Behavior, 63, pp. 132-141.

[46]. Madlon-Kay, S., Brent, L., Montague, M., Heller, K., and Platt, M. (2017)'Using machine learning to discover latent social phenotypes in free-ranging Macaques', Brain sciences, 7(7), pp. 1-24.

[47]. Burscher, B., Vliegenthart, R., and De Vreese, C. H. (2015)'Using supervised machine learning to code policy issues: Can classifiers generalize across contexts', The ANNALS of the American Academy of Political and Social Science, 659(1), pp. 122-131.

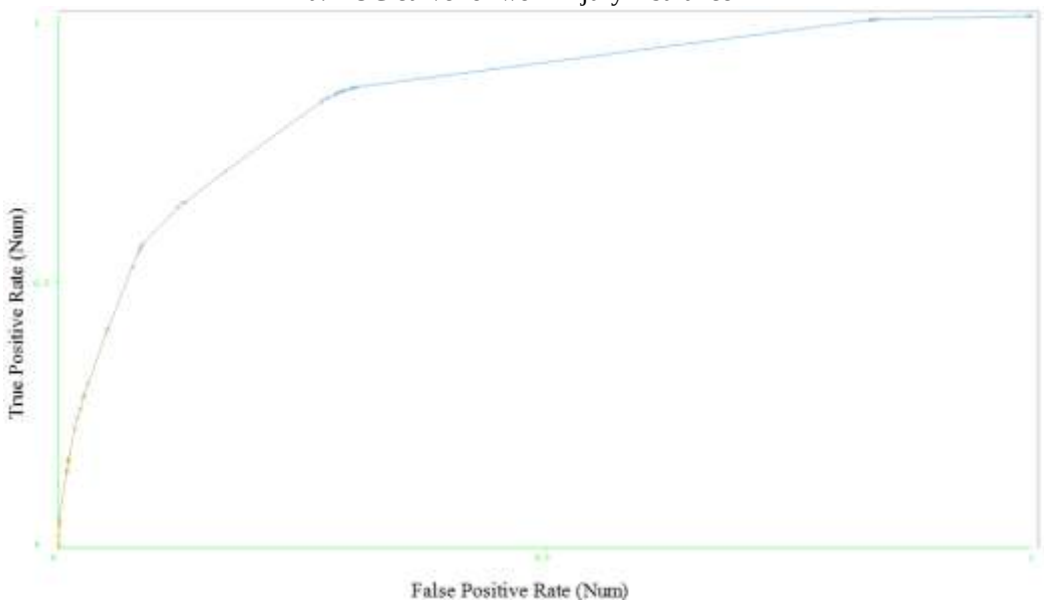
Table 1 Demographic information of participants

Variables		Frequency (N=8200)	Percentage (%)	Variables		Frequency (N=8200)	Percentage (%)
Gender	Male	3849	46.94	**Flow time (year)**	<3	5104	62.24
	Female	4351	53.06		4-9	2138	26.07
Age	<20	665	8.11		10-19	850	10.37
	21-24	1192	14.54		20-29	104	1.27
	25-33	2779	33.89		≥30	4	0.05
	34-41	2298	28.02	**Employment status**	Employed	7338	89.49
	42-54	1181	14.40		Housekeeping	706	8.61
	≥55	85	1.04		At school	46	0.56
Marital status	Unmarried	1907	23.26		Unemployed	87	1.06
	Married	6213	75.77		Retired	23	0.28
	Divorced	70	0.85	**Working days per week**	0	871	10.62
	Others	10	0.12		1	10	0.12
Education	High school	1413	17.23		2	12	0.15
	Junior college	540	6.59		3	16	0.20
	Secondary school	4356	53.12		4	69	0.84
	Undergraduate	275	3.35		5	1060	12.93
	Technical secondary school	592	7.22		6	2442	29.78
	Primary school	928	11.32		7	3720	45.37

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

	Postgraduate	22	0.27	**Working hours per day**	0-7	1085	13.23
	No schooling	74	0.90		8-9	3299	40.23
Type of household	Non-agricultural	1149	14.01		10-12	3376	41.17
	Agricultural	7051	85.99		≥13	440	5.37
Flow range	Province-to-province	4177	50.94	**Number of children**	0	2343	28.57
	City-to-city within one province	3627	44.23		1	3665	44.70
	County-to-county within one city	396	4.83		2	1981	24.16
Family monthly income	≤24002.4RMB	8142	99.29		3	195	2.38
	24002.5RMB - 48004.8RMB	45	0.55		4	15	0.18
	48004.9RMB - 72007.2RMB	8	0.10		5	1	0.01
	>72007.3RMB	5	0.06				

a. ROC curve for work injury insurance



International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

b. ROC curve for unemployment insurance

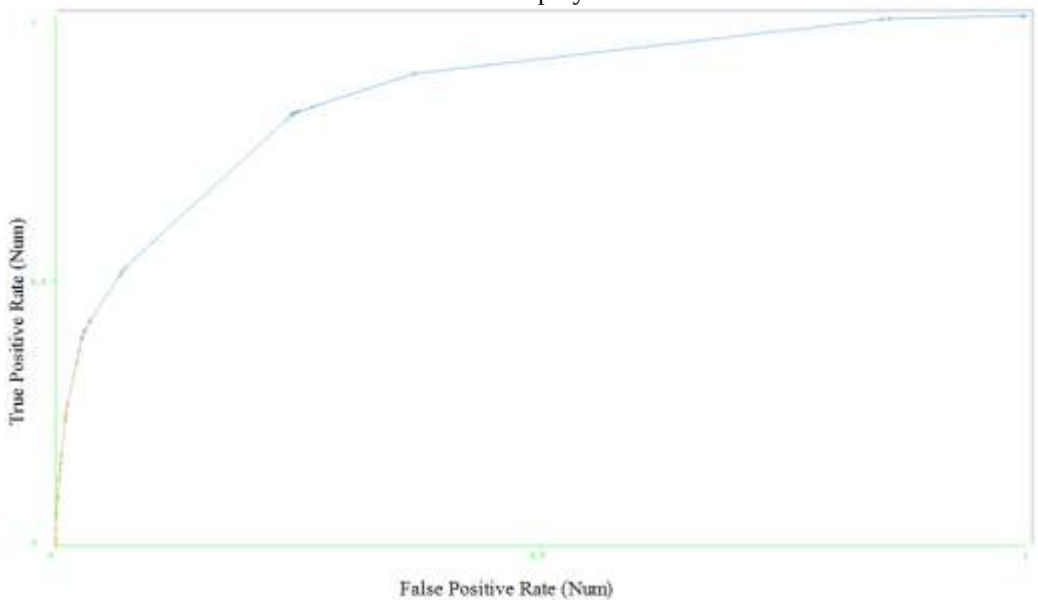


Figure 1 ROC curves a. work injury insurance b. unemployment insurance

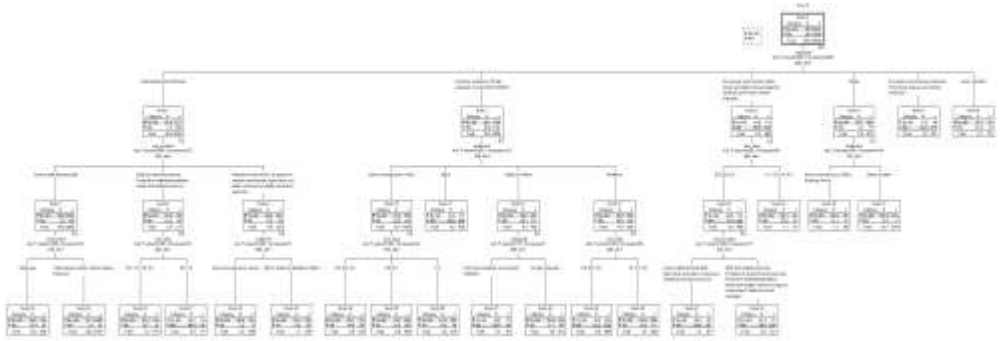


Figure 2 Decision tree prediction model of work injury insurance

(Methods of growth: CHAID, cross validation)

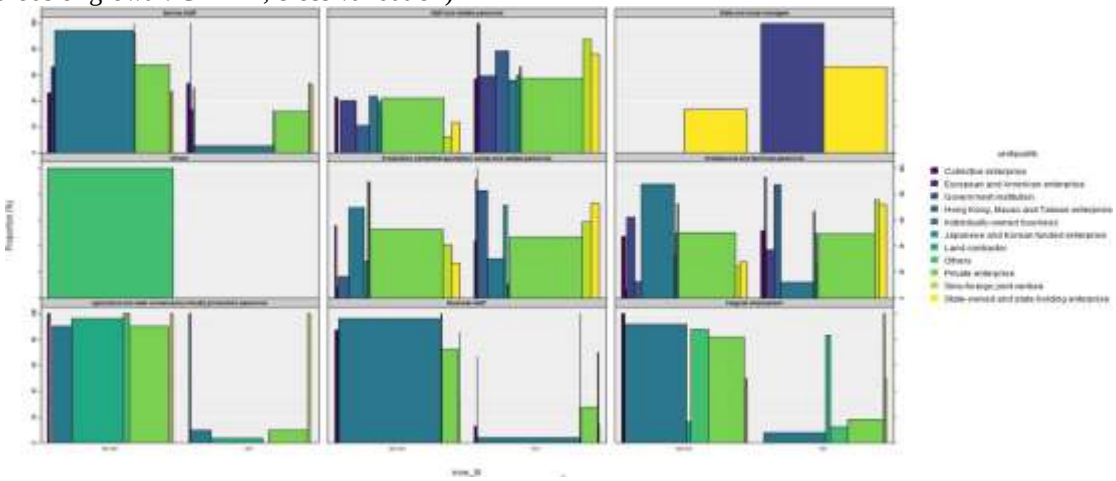


Figure 3 Distribution of WII byenterprise ownership subset byoccupation

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 886-898 www.ijaem.net ISSN: 2395-5252

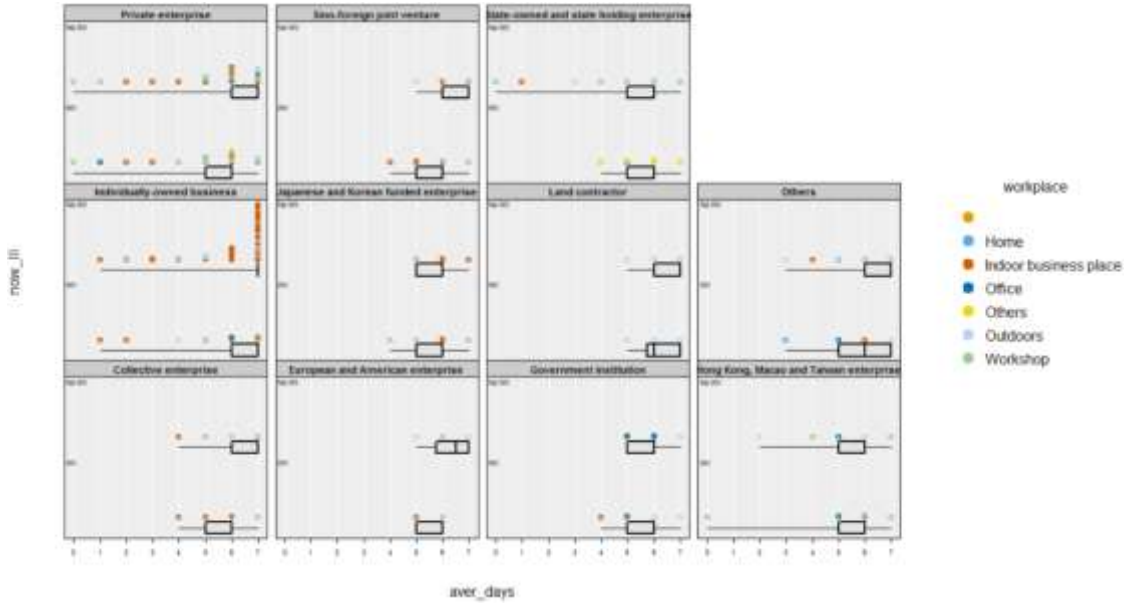


Figure 4 Distribution of WII by weekly working days subset by enterprise ownership
(Color points represents workplace.)

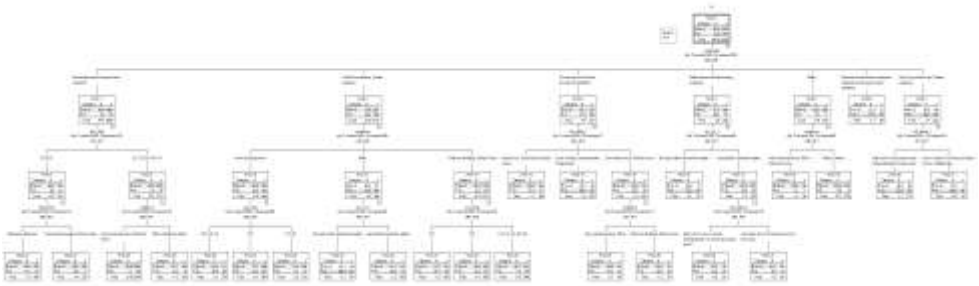


Figure 5 Decision tree prediction model of unemployment insurance
(Methods of growth: CHAID, cross validation)

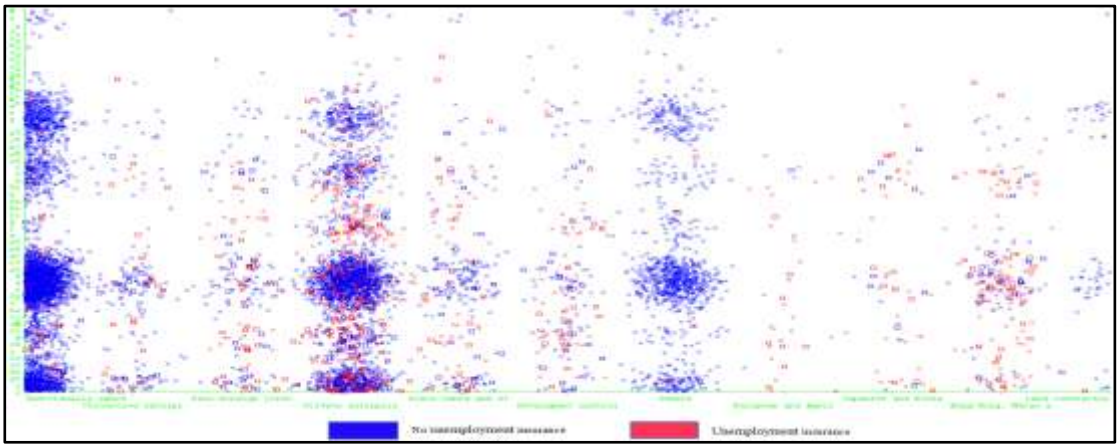


Figure6Plot of participation in unemployment insurance under different education levels and enterprise ownership