Improvement of insurance agents performance using data mining in OV agency

A N Habyba*, and R Fitriana
Department of Industrial Engineering, Trisakti University, Jakarta, Indonesia

*E-mail: anik@trisakti.ac.id

Abstract. One problem that occurs in the insurance industry is the high turnover rate of insurance agents. It makes many companies must pay a large fee for the recruitment and training of new agents. OV Agency had high insurance agent turnover rate. This company wants to improve the agents performance to avoid the turnover. This research aims to identify prospective insurance agents who have better performance at recruitment process by using classification task of data mining. There are three classification methods were used. They were decision tree (J48), Neural Network (MLP) and Naive Bayes. Based on the result, MLP has the best average accuracy value 84.2% and ROC area value 0.87 (good classification). MLP was chosen to predict the 25 insurance agent candidates and 3 agents were selected because they entered class A (better performance) in all performance criteria. The prediction results show that three agents are men, aged 25-32 years old, divorced, senior high school education and master degree, work status in other companies, and have work experience of 0-3 years. The OV Agency can consider these result in new age nt recruitment process.

1. Introduction
The high turnover rate of insurance agents is a common problem for insurance companies like OV Agency. In the previous study the OV agency’s agent performance was greatly influences their customer satisfaction. It was necessary to improve the agents service quality. Some agents were not considered to provide a detailed explanation of the product and causing customer misunderstanding [1]. The agents turnover become something that cannot be avoided by the OV Agency. The agent turnover make companies recruiting new agents frequently and also increasing the recruitment and training costs. If the companies could not predict the agents performance before recruit them, it could make other problems to the company. OV Agency must be able to select new agents with good performance in the future. Prevention of agent turnover can be done by predicting the prospective agents performance. The prediction can be done by using the old agents performance data. Based on this data, OV Agency can recruit only agents who are predicted to have good performance. The OV Agency assess agent performance based on length of service, total turnover, total policy, turnover persistency, and policy persistency. The agent performance can be divided into classes according to the company goals. Predictive data mining can be used to solve this problems.

Data mining has been widely used to assist insurance companies in processing their data to becomes useful information. Specifically for the use of classification and prediction, data mining has been used to predict consumer behavior, customers classification, the type of policy most desirable, segment performance and factors that will affect the insurance sector, also the classification of trends in organizational movements for successful/unsuccessful customer historical records [2]. In addition, data mining has also been used for the selection of insurance agents based on performance criteria for
length of service, total policy, and policy persistency. Decision tree, discriminant analysis, and artificial neural networks were used in this previous study [3]. Furthermore naive bayesian classification is also used to predict the life insurance dataset where predictions can be made on customer perceptions of existing life insurance products [4]. Research to predict insurance fraud has also been carried out using data mining. [5] implemented naive bayesian, decision tree, and rule-based classification to predict and analyze fraud patterns in auto insurance data. And the results show that the reliable model to be used to detect fraud is another type of insurance. These various studies prove that data mining has many benefits in predictive research in the insurance industry and can also be applied by the OV Agency.

A study has been conducted to analyze the best classification algorithm in predicting slow learners in the education sector. Five classification algorithms such as multilayer perceptron, naive Bayes, SMO, J48, and REPtree are used to test student academic records performance. The results show that multilayer perceptron has the highest accuracy rate of 75%, then J48 69.73%, SMO 68.42%, REPtree 67.76%, and Naive Bayes 65.13% [6]. Further research needs to be done whether these results will be the same if applied to insurance agent data. The performance of OV Agency insurance new agents candidates can also be predicted using the classification algorithm. Furthermore, this study aims to identify prospective insurance agents who have better performance at recruitment process by using predictive data mining models using a classification based algorithm.

2. Research methodology

Data mining is also commonly known as knowledge discovery from data (KDD). Data mining is part of the knowledge discovery process which is an intelligent method that can be used to extract data patterns or information needed from large data sets [7]. The purpose of using data mining in this study was not only for classification but also prediction which is a high-level goal in the application of data mining. Prediction results can be used to see the possibilities that will occur in the future. In predictive modeling, systems will guess unknown values based on previously owned data patterns [8]. Predictive data mining in this study was used to predict the class value of the five insurance agent performance criteria, namely length of service, total turnover, total policy, turnover persistency and policy persistency. The value of each class from each insurance performance criteria can be seen in Table 1.

| Agents performance criteria          | Class                              |
|--------------------------------------|------------------------------------|
| Length of service                    | Class A (length of service ≥7 years), Class B |
| Total turnover                       | Class A (turnover ≥ IDR 50000000), Class B |
| Total policy                         | Class A (policy ≥ 10), Class B    |
| Persistency turnover                 | Class A (persistency ≥ 90%), Class B |
| Persistency policy                   | Class A (persistency ≥ 90%), Class B |

Data collection was carried out using secondary data from the OV Agency. A total of 250 data from insurance agents owned by the company were used as training data and classified based on five performance criteria. The demographic data of insurance agents was used as an independent variable that will guess the performance of agents in the future. Six variables such as sex, age, marital status, academic level, employment status, and working experience are considered to affect the performance of insurance agents. This variable was chosen based on the results of previous studies [3] and adjusted for information that can be obtained from OV Agency. The classification was performed for each agent performance criteria and uses three classification methods in data mining, namely decision tree (JT8), ANN (multilayer perceptron), and naive Bayes in Weka (Waikato Environment for Knowledge Analysis) version 3.8.1. Weka is developed by the University of Waikato New Zealand as an open-source Java application that has many algorithms to analyze data sets. The C4.5 algorithm discovered by Ross Quinlan 1993 is called J48 in Weka [9]. This algorithm uses the concept of the gain ratio
which is the development of information gain. So the decision tree becomes a method that can automatically determine the most important variables in classifying data into correct class categories. The resulting tree model is also easy to understand [10].

Multilayer perceptron (MLP) is one type of ANN that also includes supervised learning which is good for classification. Multilayer feedforward networks are the most important and most widely used type of ANN. MLP has the characteristics that the model includes non-linear, sigmoidal or hyperbolic activation functions, the network can have more than one hidden layer that can learn complex tasks, the network shows a high level of connectivity between layers [11]. MLP can be used to categorize more than two output classes, in contrast to decision trees and naive Bayes, naive Bayes are opportunity-based classification methods that add up the frequency and combination of values from a given dataset. There are two assumptions in naive Bayes that all variables are priority or equally important and are independent (Bayes theorem). Naive Bayesian Classifier is a classification method that can be compared with decision trees and neural networks. This method can have a minimum error rate if the assumptions used are accurately such as independence classes and enough available probability data [7]. This research compared three methods decision tree, MLP and Naive Bayes in OV Agency insurance agent classification.

Accuracy models were made that were formed from each method and observed that the demographic variables of the agent that most influenced the results of the data classification. The classification model accuracy is judged good also based on the values of correcting classified and ROC area. The ROC (Receiver Operating Characteristic) curve shows accuracy and visually compares classifications. ROC expresses the confusion matrix. ROC is a two-dimensional graph with false positives as horizontal lines and true positive as vertical lines [12]. Accuracy based on ROC as follows [13]:

a. The accuracy is 0.90 - 1.00 = excellent classification
b. The accuracy is 0.80 - 0.90 = good classification
c. The accuracy is 0.70 - 0.80 = fair classification
d. The accuracy is 0.60 - 0.70 = poor classification
e. The accuracy is 0.50 - 0.60 = failure

Predictive data mining was done to select the new OV Agency insurance agent based on information held in the recruitment process. The results of the prediction of the performance of new agents in the future were used to be considered as acceptance of new agents. A total of 25 new agent data were used as testing data in this study. The results of the three methods were compared and analyzed.

3. Results and discussion

3.1. Classification of insurance agents performance

The results of the data mining classification for the three methods based on five insurance agent performance criteria can be seen in Table 2. The goodness of the classification model can be seen from the ROC area and accuracy. The results show that the highest ROC area value is using multilayer perceptron in each performance criterion. The highest ROC area MLP value is 0.90 obtained from total turnover with an accuracy of 84.0%, these values are indicated by boldface. The ROC value of the area of 0.90 indicates good classification.

Naive Bayes classification results have the lowest ROC area and accuracy values for each performance criteria. The lowest yield is ROC area 0.59 and accuracy 57.6%. The value of ROC area 0.59 shows the accuracy of failure classification. These results also indicate if the Naive Bayes classification fails to classify the performance criteria, namely length of service correctly. The results of the classification of the three methods look good in the total policy performance criteria where the decision tree is 78.8%, MLP 88.4%, and Naive Bayes 72.4%. In this total policy performance criteria, Naive Bayes gets the highest accuracy value compared to the other four performance criteria, which is 72.4% and ROC area value 0.64. The value of ROC area 0.64 indicates the accuracy is poor.
classification. Overall results show that MLP can be used to predict the performance of new insurance agents well. The same results [14] also show that compared to J48, MLP is a better algorithm used for many data cases. MLP algorithm based on neural networks has good learning skills and is suitable for classification problems.

**Table 2. Model performance of the classification (use training set)**

| Performance criteria | Classification methods | Class | TP rate | FP rate | Precision | Recall | F-measure | ROC area | Accuracy (correctly classified) |
|----------------------|------------------------|-------|---------|---------|-----------|--------|-----------|----------|--------------------------------|
| Length of service    | J48                    | A     | 0.87    | 0.59    | 0.67      | 0.87   | 0.76      | 0.67     | 67.6%             |
|                      |                        | B     | 0.41    | 0.13    | 0.69      | 0.41   | 0.52      | 0.67     |                  |
|                      | MLP                    | A     | 0.82    | 0.12    | 0.90      | 0.82   | 0.86      | 0.83     | **84.4%**         |
|                      |                        | B     | 0.88    | 0.18    | 0.78      | 0.88   | 0.83      | 0.83     |                  |
|                      | Naive Bayes            | A     | 0.85    | 0.80    | 0.59      | 0.85   | 0.70      | 0.59     | 57.6%             |
|                      |                        | B     | 0.20    | 0.15    | 0.49      | 0.20   | 0.28      | 0.59     |                  |
| Total turnover       | J48                    | A     | 0.26    | 0.03    | 0.86      | 0.26   | 0.40      | 0.65     | 70.4%             |
|                      |                        | B     | 0.97    | 0.74    | 0.68      | 0.97   | 0.80      | 0.65     |                  |
|                      | MLP                    | A     | 0.77    | 0.12    | 0.80      | 0.77   | 0.79      | 0.90     | **84.0%**         |
|                      |                        | B     | 0.88    | 0.23    | 0.86      | 0.88   | 0.87      | 0.90     |                  |
|                      | Naive Bayes            | A     | 0.27    | 0.12    | 0.59      | 0.27   | 0.37      | 0.61     | 65.2%             |
|                      |                        | B     | 0.88    | 0.73    | 0.67      | 0.88   | 0.76      | 0.61     |                  |
| Total policy         | J48                    | A     | 0.36    | 0.29    | 0.84      | 0.36   | 0.51      | 0.73     | 78.8%             |
|                      |                        | B     | 0.97    | 0.64    | 0.78      | 0.97   | 0.87      | 0.73     |                  |
|                      | MLP                    | A     | 0.81    | 0.09    | 0.80      | 0.81   | 0.81      | 0.88     | **88.4%**         |
|                      |                        | B     | 0.91    | 0.19    | 0.92      | 0.91   | 0.92      | 0.88     |                  |
|                      | Naive Bayes            | A     | 0.11    | 0.11    | 0.80      | 0.11   | 0.19      | 0.64     | 72.4%             |
|                      |                        | B     | 0.99    | 0.89    | 0.72      | 0.99   | 0.83      | 0.64     |                  |
| Persistency turnover | J48                    | A     | 0.81    | 0.40    | 0.75      | 0.81   | 0.78      | 0.77     | 72.4%             |
|                      |                        | B     | 0.60    | 0.19    | 0.67      | 0.60   | 0.64      | 0.77     |                  |
|                      | MLP                    | A     | 0.95    | 0.41    | 0.78      | 0.95   | 0.85      | 0.85     | **80.4%**         |
|                      |                        | B     | 0.59    | 0.05    | 0.88      | 0.59   | 0.71      | 0.85     |                  |
|                      | Naive Bayes            | A     | 0.87    | 0.78    | 0.63      | 0.87   | 0.73      | 0.61     | 60.8%             |
|                      |                        | B     | 0.22    | 0.13    | 0.53      | 0.22   | 0.31      | 0.61     |                  |
| Persistency policy   | J48                    | A     | 0.82    | 0.28    | 0.79      | 0.82   | 0.80      | 0.85     | 77.6%             |
|                      |                        | B     | 0.72    | 0.18    | 0.76      | 0.72   | 0.74      | 0.85     |                  |
|                      | MLP                    | A     | 0.94    | 0.29    | 0.81      | 0.94   | 0.87      | 0.87     | **84.0%**         |
|                      |                        | B     | 0.71    | 0.06    | 0.91      | 0.71   | 0.80      | 0.87     |                  |
|                      | Naive Bayes            | A     | 0.74    | 0.63    | 0.60      | 0.74   | 0.66      | 0.63     | 57.6%             |
|                      |                        | B     | 0.37    | 0.26    | 0.53      | 0.37   | 0.44      | 0.63     |                  |

3.2. Prediction of new insurance agents performance

The purpose of OV Agency is to get a new insurance agent that is predicted to have good performance, namely class A. Prediction results of 25 new insurance agents used variables of sex, age, marital status, academic level, employment status, and working experience on each performance criterion showing results differently. The prediction result can be seen in Table 3. The highest accuracy multilayer perceptron method also shows the highest number of agents entering class A. At the length of service, as many as 11 agents are predicted to enter class A which has a work duration of \( \geq 7 \) years. Based on the total turnover of 7 agents will reach a turnover of \( \geq \) IDR 50,000,000 and 8
agents will reach ≥10 policies. Whereas based on the 15 agent turnover persistency and the 18 agent policy persistency will have a persistency of ≥ 90% or fall into class A.

Persistency is one of the determinants of agent performance that is very influential on the sustainability of insurance companies. The persistency value refers to the volume of turnover and policy that the agent can maintain. The persistency is obtained from the proportion of active policies remaining at the end of the period out of the total policies at the beginning of the period [15]. OV Agency has turnover data and non-active policies from each agent in order to monitor agent performance through persistency. High persistency value is a necessity that must be achieved by each agent for the continuity of the company, therefore OV Agency expects persistency ≥ 90% of each agent-owned and new agents to be recruited. Small persistency will have a negative impact on the company. If there are a lot of inactive policies or turnover, the company will have difficulty bearing the death rates of the existing policyholders. In addition, it also reduces the profits of the company, the increase in capital costs due to a decrease in revenue as well as the increase in operational costs and sales of each policy, the price of a new policy becomes expensive and more difficult to sell and policyholders will stop paying if it is too expensive [16].

| Performance criteria | Classification method | Predictive class | Total | Accuracy (correctly classified) |
|----------------------|-----------------------|-----------------|-------|--------------------------------|
| Length of service    | J48                   | A               | 14    | 67.6%                          |
|                      |                       | B               | 11    |                                |
|                      | MLP                   | A               | 11    | 84.4%                          |
|                      |                       | B               | 14    |                                |
|                      | Naive Bayes           | A               | 10    | 57.6%                          |
|                      |                       | B               | 15    |                                |
| Total turnover       | J48                   | A               | 4     | 70.4%                          |
|                      |                       | B               | 21    |                                |
|                      | MLP                   | A               | 7     | 84.0%                          |
|                      |                       | B               | 18    |                                |
|                      | Naive Bayes           | A               | 0     | 65.2%                          |
|                      |                       | B               | 25    |                                |
| Total policy         | J48                   | A               | 0     | 78.8%                          |
|                      |                       | B               | 25    |                                |
|                      | MLP                   | A               | 8     | 88.4%                          |
|                      |                       | B               | 17    |                                |
|                      | Naive Bayes           | A               | 0     | 72.4%                          |
|                      |                       | B               | 25    |                                |
| Persistency turnover | J48                   | A               | 18    | 72.4%                          |
|                      |                       | B               | 7     |                                |
|                      | MLP                   | A               | 15    | 80.4%                          |
|                      |                       | B               | 10    |                                |
|                      | Naive Bayes           | A               | 5     | 60.8%                          |
|                      |                       | B               | 20    |                                |
| Persistency policy   | J48                   | A               | 8     | 77.6%                          |
|                      |                       | B               | 17    |                                |
|                      | MLP                   | A               | 18    | 84.0%                          |
|                      |                       | B               | 7     |                                |
|                      | Naive Bayes           | A               | 14    | 57.6%                          |
|                      |                       | B               | 11    |                                |
3.3. **Multilayer perceptron prediction results**

Multilayer perceptron was used to evaluate 25 new OV Agency insurance agent candidates. The results of the prediction of the class five performance criteria for each prospective insurance agent can be seen in Table 4. The results show that there are three agents who classified as class A in all performance criteria, namely Agent 1, Agent 15, and Agent 22. These three agents are men, aged 25-32 years old, divorced, senior high school education and master degree, work status in other companies, and have work experience of 0-3 years. These results are in accordance with MLP learning of training data, so the accuracy of the results is strongly influenced by the quality of the training data used.

Length of service is an important agent performance criteria because predicting agents who can have a long service life will reduce agent turnover. Agent turnover in an insurance company is caused by a number of things including high-pressure work that requires agents to understand customer needs and offer suitable products, poorly planned recruitment processes and insurance companies have agents from various sectors so that the likelihood that agents will switch jobs to other sectors is high [17]. MLP in this study was successfully used to predict the length of service as one of the important criteria for agent performance.

| Insurance agent candidates | Length of service | Total turnover | Total policy | Persistency turnover | Persistency policy |
|----------------------------|-------------------|----------------|-------------|----------------------|-------------------|
| 1                          | Class A           | Class A        | Class A     | Class A              | Class A           |
| 2                          | Class B           | Class B        | Class B     | Class A              | Class A           |
| 3                          | Class B           | Class B        | Class B     | Class A              | Class A           |
| 4                          | Class A           | Class B        | Class B     | Class B              | Class A           |
| 15                         | Class A           | Class A        | Class A     | Class A              | Class A           |
| 16                         | Class B           | Class A        | Class B     | Class B              | Class A           |
| 17                         | Class A           | Class B        | Class A     | Class B              | Class B           |
| 22                         | Class A           | Class A        | Class A     | Class A              | Class A           |
| 23                         | Class B           | Class B        | Class A     | Class B              | Class A           |
| 24                         | Class B           | Class B        | Class B     | Class B              | Class B           |
| 25                         | Class B           | Class B        | Class B     | Class B              | Class B           |

4. **Conclusion and recommendation**

In this research, there are three classification methods used to train insurance agents data of OV Agency, they are decision tree (J48), Neural Network (MLP), and Naive Bayes. Based on the result, multilayer perceptron has the best average accuracy value 84.2% and ROC area value 0.87 (good classification). The accuracy values are high and have the highest number of agents entering class A for five agent performance criteria. MLP was chosen to predict the 25 insurance agent candidates and 3 agents were selected because they entered class A (better performance) in all performance criteria. The prediction results show that three agents are men, aged 25-32 years old, divorced, senior high school education and master degree, work status in other companies, and have work experience of 0-3 years. The OV Agency can consider these result in new agent recruitment process. The results were under MLP learning of training data, so the accuracy of the results is strongly influenced by the quality.
of the training data used. For future work more classification methods can be used and much more dataset from OV Agency should be taken.

Acknowledgments
The authors pay homage to the Faculty of Industrial Technology, Trisakti University for continuous encouragement, guidance, and support for their work. This paper is the research funded by Trisakti University in the academic year 2019/2020.

References
[1] Fitriana R, Kurniawan W and Anwar M R 2019 Measurement and proposal of improving marketing process to improve the quality of aftersales services with fuzzy quality function deployment and data mining methods in ov agency IOP Conference Series: Materials Science and Engineering
[2] Umamaheswari K and Janakiraman S 2014 Role of data mining in insurance industry An Int. J. Adv. Comput. Technol. 3 961–6
[3] Cho V and Ngai E W T 2003 Data mining for selection of insurance sales agents Expert Syst. 20 123–32
[4] Balaji S and K. Srivatsa S 2012 Naïve bayes classification approach for mining life insurance databases for effective prediction of customer preferences over life insurance products Int. J. Comput. Appl. 51 22–6
[5] Bhowmik R 2011 Detecting auto insurance fraud by data mining techniques J. Emerg. Trends Comput. Inf. Sci. 2 156–62
[6] Kaur P, Singh M and Josan G S 2015 Classification and prediction based data mining algorithms to predict slow learners in education sector Procedia Computer Science 57 500–8
[7] Han J, Kamber M and Pei J 2012 Data mining: concepts and techniques (Waltham: Morgan Kaufmann)
[8] Sumathi S and Sivanandam S N 2006 Introduction to data mining and its applications (Heidelberg: Springer)
[9] Drazin S 2010 Decision tree analysis using weka Mach. Learn. II, Univ. Miami 1–3
[10] Kohavi R and Quinlan R 1999 Decision tree discovery Handb. Data Min. Knowl. Discov. 3 267-276
[11] Kantardzic M 2019 Data mining: concepts, methods, and algorithms (Piscataway: IEEE Press)
[12] Vercellis C 2009 Business intelligence: data mining and optimization for decision making (West Sussex: John Wiley & Sons Ltd.)
[13] Gorunescu F 2011 Data mining: concepts, models and techniques (Heidelberg: Springer)
[14] Arora R and Suman S 2012 Comparative analysis of classification algorithms on different datasets using weka Int. J. Comput. Appl. 34 21-25
[15] Life Insurance Council 2010 Persistency of life insurance policies – an exposure draft 1–5
[16] Capgemini 2013 Persistency management framework for life insurance 1–20
[17] Pathak S and Tripathi V 2010 Sales force turnover: an explanatory study of the indian insurance sector Manag. Univ. Primarosk. Fac. Manag. Koper. 5 3–19