서울과학기술대학교
SEOULTECH SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

Applied Artificial Intelligence

MACHINE LEARNING 1

A Predictive Model for Suicidal Ideation of Elderly Using Random Forests Machine Learning Algorithm

Cho Minji (21102390),
Jeong Dayena (21102418)

A Predictive Model for Suicidal Ideation of Elderly Using Random Forests Machine Learning Algorithm

Cho Minji[i], Jeong Dayena[ii]

[i] Seoul National University of Science And Technology, Department of Applied Artificial Intelligence S.Korea

[ii] Seoul National University of Science And Technology, Department of Applied Artificial Intelligence S.Korea

27 May 2

Abstract

This study set out to present a predictive model for suicidal ideation of the elderly using a random forest and identified the characteristics of predictors. The Survey on the Elderly data, from 2014, 2016, and 2020 conducted by the Ministry of Health and Welfare, was used in the empirical analysis. The subjects of the study were 18,258 elderlies. Predictive models were estimated using 31 factors, which were to predict suicidal thoughts of the elderly reported in previous studies and variables to be added in subsequent studies, and random forests machine learning algorithm. The evaluation of predictive models was shown to be an accuracy of 80.70%, a sensitivity of 26.46%, and a specificity of 96.21%, respectively. The relative importance of the predictors was that the importance of variables in terms of economic conditions was generally higher than the others. This study tries to reflect the risk assessment of suicide attempts and discuss intervention methods for these variables, and it is thought that countermeasures are required to prevent suicidal thoughts or attempts.

Key words: machine learning, random forests, elderly, suicidal ideation, predictive model

1. Introduction

Individuals attain longevity as human life expectancy increases, while society is aging due to a rise in the senior population. Furthermore, due to our society's low birth rate and the retirement of a big number of baby boomers, the share of the old population is increasing, causing new social difficulties. Suicide of the elderly is a common occurrence, and the rise in the senior population in Korean culture has resulted in a new societal problem known as the suicide of the elderly(Yeong-Kyung and Moo-Sik, 2021).

According to the Ministry of Health and Welfare's "White Paper on Suicide Prevention" issued in 2021, Korea's suicide rate for elderly adults aged 65 and older per 100,000 population was 57.9 in Daejeon and 46.6 nationwide, considerably above the OECD average (17.2) and Slovenia (36.9).

Developing measures to prevent suicide in the elderly has become critical, and it has been determined that developing a model to predict whether suicidal ideation will occur, as well as analyzing the predictive factors, will aid in the development of measures to prevent suicide in the elderly.

Therefore, the goal of this study is to estimate the elderly's suicide thought prediction model through the random forest algorithm, a technique not used in previous studies, and to identify the predictive factors.

2. Related Work

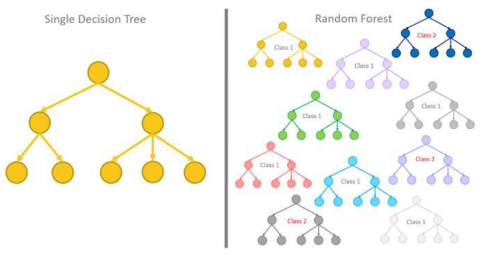

Fig 1. Random Forest

Random forests analyze and aggregate multiple decision trees to generate a final prediction model and go through the process of forming a forest with multiple randomly sampled decision trees. Random forests repeatedly generate independent decision trees by providing maximum randomness in model-specific sample selection and variable selection, thus reducing prediction errors while maintaining low bias in the decision tree(Dangeti, 2017; Géron, 2017). In addition, high-dimensional data containing multiple explanatory variables are stable without causing errors because it considers the interaction and nonlinearity between explanatory variables(Dangeti, 2017; Géron, 2017).

It can be inferred that random forest can be used for research to predict and understand the characteristics of factors that contribute to predicting subjective individual experiences that cannot be objectively grasped, such as suicidal thoughts. Therefore, the purpose of this study is to contribute to establishing measures for suicide prevention by estimating models that predict suicide thoughts in the elderly using random forest machine learning algorithms and identifying factors that contribute to prediction(Ki-Hye, 2020).

3. Dataset and Features

1) Data and Research subjects

The data used in this study came from a three-year elderly survey conducted by the Ministry of Health and Welfare in 2014, 2016, and 2020. The subjects of each year's survey were studied, and a total of 18,258 among 30,847 respondents, excluding those who did not respond and outliers (12,589), were analyzed.

2) Label and Feature

1. Label

The label of this study is suicidal thoughts. It is a binary variable that sets "I have never thought of suicide" to 0 and "I have thought of suicide" to 1(Ki-Hye, 2020). <Table 1> is a frequency table of the elderly for the presence or absence of suicidal thoughts, recoded based on the entire sample.

	Frequency(people)	Ratio(%)
Suicidal idea(No)	660	86.96(%)
Suicidal idea(Yes)	99	13.04(%)
Total	759	100(%)

Table 1. the frequency table of suicidal ideas in the elderly

2. Feature

The feature of this study is a variable used to predict the suicide of the elderly reported in previous studies and a variable that requires follow-up studies. The explanatory variables are 31 variables, consisting of a total of 8 categories, and the items are shown in Table 2.

	Categories and scales	Features
1	Demographic factor(4)	Total number of household members, Elderly household type, Form of residence, Types of housing
2	Health behavior factor(3)	frequency of drinking over the past year, Nutritional status, Physical condition
3	Health level factor(2)	Subjective health conditions, Total number of diagnosed chronic diseases
4	Functional factors(3)	Vision problems, Hearing problems, and Oral problems
5	Economic condition(3)	National Basic Living Security beneficiary status, Personal debt status, Economic status
6	Relational factor(7)	Conflict with children in the past year, Spousal health status, Number of close relatives, Including siblings, Number of close friends/neighbors/acquaintances, Relationship with spouse, Relationship with children, Relationship with friends and community
7	Quality factors in life(8)	Satisfaction, whether you experienced discrimination in your daily life, whether you suffered physical pain from others, whether you were hurt by others, whether you suffered financial damage from others, whether you were not cared for by your family or guardians, Whether family members or guardians are neglecting or not supporting living expenses
8	Living environment factors(2)	Experience in falls over the past year, social/leisure/cultural activities
	Total	30

Table 2. Features categories, scales, and detailed items

Of these, 10 variables are continuous variables, and all 21 are categorical variables.

4. Methods

R was used for data analysis in this study, and data preprocessing, model development, performance evaluation, and results from the analysis were conducted using Python(Ki-Hye, 2020).

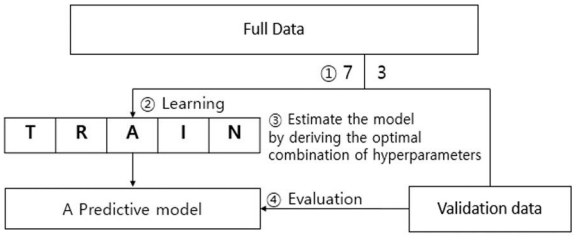

Fig 2. Machine Learning Procedure for Prediction Model Estimation and Model Evaluation

1) Data Preparation

1. Categorical variables were converted using 'Label Encoder', and continuous variables were normalized using 'Min-Max Scaling'.

2. The training and test data were separated at random in a 7:3 ratio, which is a usual practice(Young-Sik et al., 2019; Pil-Sun and In-Sik, 2018).

2) Predictive Model Assessment

1. The training data was split into 5 groups, and k-fold cross-validation was used.

2. Create a set of hyper-parameters from 500 random searches to determine the most appropriate model.

3. The prediction model is adjusted to the verification data to assess prediction performance for elderly people who consider suicide versus those who do not.

3) Analyzing Features

1. Analyze and visualize the importance of explanatory Features

2. Analyze the type of functional relationship between the prediction of a feature and the model. (PDP)

4) Model Evaluation Indicators

1. Confusion Matrix

<Figure 3> is a table showing the confusion matrix for evaluating the prediction model. The performance indicators for evaluating the model include accuracy, recall, specificity, precision, F1-score, and ROC-AUC score, which means that the higher the value, the higher the predictive power of the model.

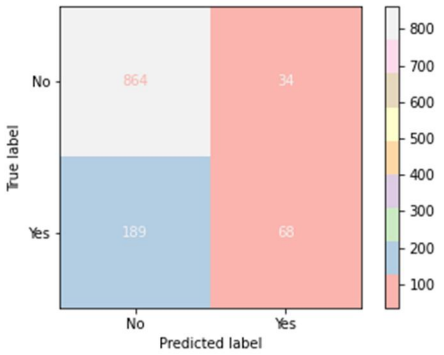

Fig 3. Confusion matrix

Accuracy is the ratio of correctly predicting the elderly who think about suicide and the elderly who do not think about suicide.

$$Accuracy = TN+TP \;/\; (TN+FN+FP+TP)$$

Sensitivity or Recall is the ratio of predicting that the elderly who think about suicide are the elderly who think about suicide.

$$Sensitivity\ or\ Recall = TP \;/\; (FN+TP)$$

Specificity is the ratio of predicting that the elderly who do not actually think of suicide are the elderly who do not think of suicide.

$$Specificity = TN \;/\; (TN+FP)$$

Precision is the ratio of the elderly who actually think about suicide over the elderly who predict that they are thinking about suicide.

$$Precision = TP \;/\; (FP+TP)$$

F1-score is a harmonic average of sensitivity and precision, which is high when sensitivity and precision are similar, and 1 when it has perfect sensitivity and precision.

$$F_1\text{-}score = 2 \;/\; (1/sensitivity) + (1/precision)$$

2. The ROC-AUC score

It represents the AUC-area as a percent. The ROC-AUC score becomes an accurate model as the AUC-area widens and converges to 100%.

5. Results

1) Predictive Model Performance

In machine learning, model assessment

examines prediction performance by assessing the model's generalizability by fitting the prediction model to data that was not included in the model formation. The verification data prediction result for the model in which the elderly's suicidal thoughts are assigned as dependent variables is shown in Table 3. (Ki-Hye, 2020). The accuracy, which represents the model's performance, is 80.7%, suggesting that the predictive model's performance is dependable.

	Model(%)
Accuracy	80.70
Sensitivity or Recall	26.46
Specificity	96.21
Precision	66.67
F1-score	37.88
ROC-AUC-Score	78.56

Table 3. Suicidal Thoughts in the Elderly: Random Forest Prediction Performance

The specificity is 96.21%, indicating that the model's accuracy in predicting that the elderly who do not consider suicide are those who do not consider suicide is excellent. This also indicates that it can be used to define older people who do not consider suicide without looking into their suicidal ideas. As a result, the predictive model can be used to identify the elderly who should be excluded from the target population while developing programs for the elderly who are contemplating suicide in the social welfare practice area.

The sensitivity, on the other hand, was 26.46%, which was lower than the specificity. This could indicate that the model's performance in predicting the elderly who truly consider suicide as the elderly who consider suicide is insufficient. If this is taken further, it can be deduced that the performance is poor in locating old people who are contemplating suicide while being concealed or neglected. Just like the sensitivity, the F1-score (37.88%) is low, implying that more data is needed because the ratio (22%) of the elderly considering suicide is low.

However, the precision, which is the probability that the elderly is thinking about suicide, is 66.67%, and the ROC-AUC-score is 78.56%, showing reliable results.

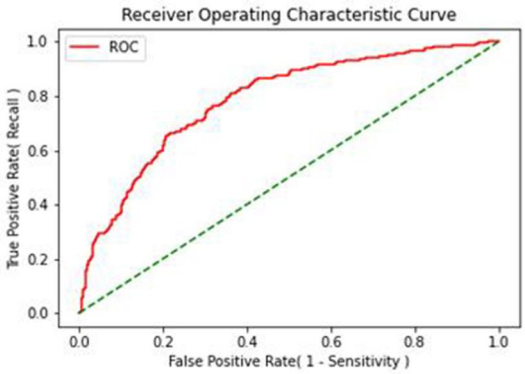

Fig 4. ROC-Curve and AUC-Area

The predictive performance of the above model can be seen as a result that is comparable to the predictive performance of the model reported in social science studies analyzed with random forest machine learning algorithms. This can be said to confirm the possibility that the prediction of individuals based on machine learning can play a role in diagnosing the elderly who are not motivated to live(Ki-Hye, 2020).

2) Importance of Feature

Although the machine learning model does not perform statistical significance diagnosis, the random forest can analyze the relative importance of labels that contributed to prediction(Ki-Hye, 2020). <Figure 5> is a visualization of relative importance centering on labels constituting one category and scale. The characteristics shown in the importance given according to the degree of contribution of label predicting suicidal thoughts of the elderly are as follows.

1. The label with the highest importance in predicting suicidal thoughts of the elderly was found to be 'Satisfaction with current economic conditions.'

2. The overall importance of labels related to 'types damaged by others' or 'economic conditions' was found to be high.

3. We could see that the impact of health-related labels and personal debt was not at the top, contrary to predictions.

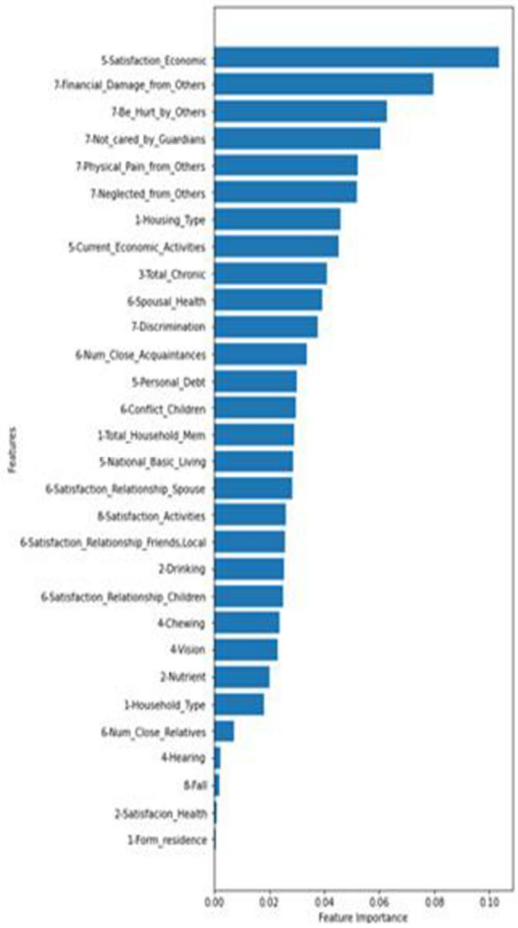

Fig 5. Importance of Feature

3) PDP: Partial Dependence Plots

Based on a KBS documentary dealing with the issue of elderly suicide under the theme of 'old bankruptcy', four labels related to 'economy' were selected and visualized.

<Figure 6> is a visualization of the type of relationship between labels related to 'economy' that affects suicidal thoughts. The relationship type between personal debt, basic livelihood recipients, current economic activity, and suicidal thoughts appeared in the positive (+) direction, and economic satisfaction and suicidal thoughts appeared in a straight line in the negative (-) direction.

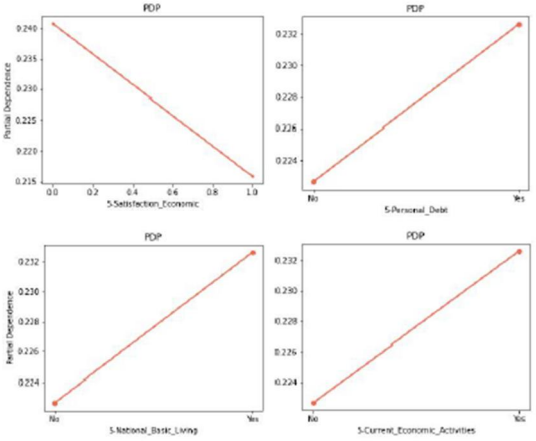

Fig 6. PDP of Economic conditions

6. Conclusion

The model's performance and results fluctuate each time it is run due to the random extraction of the "NO" option, which is challenging. Also, the data was biased and detailed, making it difficult to work with. This study tries to reflect the risk assessment of suicide attempts and discuss intervention methods for these variables, and it is thought that countermeasures are required to prevent suicidal thoughts or attempts.

References

• Dangeti, P., Statistics for Machine Learning. Birmingham, UK: Packt Publishing Ltd, 2017.
• Gam-Ram Park, A Study on the Effect of the Household Type on Health Behavior and Health Level in the Elderly over 65, December. 2020.
• Géron, A., Hands-On Machine Learning with Scikit-Learn and TensorFlow, Sebastopol, CA: O'Reilly Media, 2017.
• Ki-Hye Hong, A Predictive Model for Suicidal Ideation of Adolescents Using Random Forests Machine Learning Algorithm, August. 2020.
• Pil-Sun Choi, In-Sik Min, A Predictive Model for the Employment of College Graduates Using a Machine Learning Approach, March. 2018.
• Sun-Mi Kim, Gyung-Joo Lee, Risk Factors of Suicide Ideation in Younger-Old and Older-Old Persons:Using Data from the Korea Health Panel Survey, November. 2020.
• Yeong-Kyung Hong, Moo-Sik Lee, Factors Related to Attempts of Suicide in Korean Elderly -Using hierarchical regression-, October. 2021.
• Young-Sik Kim, Eun-Jeong Lee, Hyun-Jun Joo, Exploring a predictive variables for the university entrance through Korean-Type Early Decision Programs.The journal of Educational Studies, December. 2019.

Addendum

<Figure 7> shows that one decision tree derived from the random forest analysis process is one decision tree derived from the random forest machine learning algorithm analysis process of this study. Random Forest repeatedly creates an independent decision tree by giving maximum randomness to sample selection and variable selection, and analyzes it to have a final value. Since this study is a category for diagnosing suicidal thoughts, the results of repeated decision-making trees as follows are counted as voting to have the final result value(Ki-Hye, 2020).

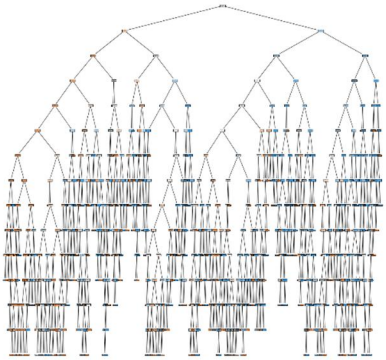

Fig 7. A Decision Tree Derived from the Analysis Process of Random Forest Machine Learning Algorithm