Prediction Medical Problem of Elderly People by Using Machine Learning Technique

A Tongkaw\textsuperscript{1} and S Tongkaw\textsuperscript{2}

Abstract. This paper describes the machine learning technique for classification and prediction of the possibility of medical problem occurring in elderly people using sklearn in Python. The research chose 16 attributes out of 23 attributes. Overall, 353 samples can be classified and divided into category 0, no medical condition, 166 samples, and category 1, have medical condition, 187 samples. The root node is level of education. It can be classified with the number of cigarettes per day and have exercise or not. An accuracy value after removing the attribute hobby is about 67.80\%, considered as good accuracy with the real data. The graph representing the decision tree results is drawn by using pydotplus. This work can improve for future work by finding more parameters or other algorithms such as neural network for improving the accuracy.

1. Introduction

The elderly are people who have limitations in their daily activities or social involvement. Because of their effectiveness in seeing, hearing, performance, movement efficiency, meaninglessness, communication, efficiency began to diminish affects the psychological, emotional, behavioral, and has limited memory and learning or other aspects. It is necessary to be assisted in one way to be able to practice daily activities or have happy social participation. Getting help these elderly people do not fully work due to the shortage of resources and truly reach those in need. At present, various technologies have played a significant role in everyday life, whether it is computer technology, smartphone technology, a trendy technology. Those are also widely used today. Elderly health is important. Due to the increasing number of patients present from various causes such as diabetes, renal failure, or even toxic substances contaminated in food. The food of the elderly is also essential. We should consider the appropriateness of the food menu of the elderly quantity of drinking water consumption of alcohol, soft drinks, also must be considered. Food hygiene, cleanliness, knowledge about the benefits of drinking water, such as reducing body temperature, time of drinking water, and balance control of the body. The elderly should be in the right environment, such as using clean water, no garbage clogged in the drain. With clean water, consumed and sufficient in a clean house. Community access, hospitals, public access services, close to home in the issue of providing knowledge about actions of the elderly in the community, such as the use of Thai traditional medicine, exercise, smoking, participation in social activities, as well as various hobbies.

Decision tree is a classification technique used for various classification problems such as classification of land cover from remotely sensed data. (1) Learning of decision tree is a learning by classification data into various groups using the attributes. The decision tree obtained from learning makes it known that what features are the classification criteria, each feature is important and what is

\textsuperscript{1} Faculty of Science and Technology, Songkhla Rajabhat University, Songkhla, Thailand
\textsuperscript{2} Faculty of Science and Technology, Songkhla Rajabhat University, Songkhla, Thailand
the difference. A research used K-Means clustering and C4.5 decision tree algorithm for network detection. The result from the research show that the K-Means clustering and C4.5 decision tree algorithm can give impressive detection accuracy and precision. The decision tree has three components, first, an internal node is a feature of data that, when any data falls, the node will use this feature to decide which direction the data will go. By the inner node that is the beginning of the tree is called the root; second, the branch, link is the value of the feature in the internal node that breaks this branch, in which, the inner node will be branched to the number equal to the number of attributes in that inner node; third, leaf node is a group which results in data classification. There is limitation of the research which use the real situation data. A research had done about evaluate the susceptibility of an area to landslide by using certainty factor and index of entropy with alternating decision tree models. (2) Another important research is using decision tree methodology to predict employment after moderate to severe traumatic brain injury. It suggested that the decision tree is ease-to-use and can provide prognostic information on long-term competitive employment outcomes related with the moderate to severe traumatic brain injury. It concluded that there were a set of factors including length of posttraumatic amnesia, a clinical maker of injury severity, and preinjury education and employment status led to the result of prediction. (3) Many researches talk about using decision tree to classification health care data such as a research that enhancing the performance of decision tree by using NSUM technique, Novel Symmetrical Uncertainty Measure, for diabetes patients. The research use WEKA tool. (4) A research used WEKA to predict heart decease with hybridization technique with decision tree and artificial neural network. Although, the techniques were accuracy sensitivity and specificity of the classification individually, the data was from UCI repository not from the real situation. (5) A research confirmed that Naïve Bayes and Decision tree with information gain gave better results in the diagnosis of heart disease and better accuracy result compared with other classifiers. (6) For elder people, giving high-quality services is important, a paper use adaptive decision support system by using the text mining to integrate in case adaptation process to up-to-date medical information from the internet. (7) This research try to extend the idea for elder people in the health care services.

It is true that predicting the occurrence of health problems can consult a doctor at the hospital because doctors can have a detailed diagnosis and more accurate medical information from hospitals is more accurate. However, the hospital has many patients and screening for screening to increase the burden on the hospital and some seniors do not regularly check-up or refuse to go to the hospital. In addition, some seniors understand that going to the hospital only when the health problem has really happened, therefore, this research creates a preliminary prediction model. Which gives an idea of the possible important conditions that may cause health problems for the elderly in this area, and also can identify the medical problems that often occur with the elderly from the interview, the benefit from this research is that it can be a guideline for advising the elderly to receive medical treatment in the hospital. In the event that there is a chance of a problem causing the elderly to receive treatment from the start. Causing the problem to not spread and reduce the cost of screening the problem. Therefore, classification is useful, allowing people to analyse data and make decisions more accurately.

2. Objective

This research aims to describes the decision tree of machine learning technique for classification and prediction of the possibility of medical problem occur of elder people by using Python. The result from sklearn library from Python will describe. The important of the variables or attributes will show in the results which influence of the medical problem of elder people.

3. Method

The decision tree algorithm has a tree structure that looks like a hierarchy chart, in which the internal nodes represent characteristics or attribute fields represent decision rules and each leaf node represents the result. The top node in the decision tree is called the root node. The algorithm will learn to partition based on the attribute values based on the data. By partitioning the trees in each section will be called repeatedly by repartitioning. Starts partitioning tree building repeating this process recursively for each child until one of the condition will match, all of the tuples belong to the same attribute value, or there are no more remaining attributes, or there are no more instances. This
hierarchical structure helps you to make decisions helps to create a visual information that mimics human-level thinking. The elder people data were collected from the real situation by interview a district in Songkhla province. The data will clean in cleaning process and separated into training and testing data set. Then, select best attribute using attribute select measures such as Information Gain or Gini index or Gain Ratio. This break data set into small subset. The process of select the best attribute and break data set recursively repeat the process for each child. After training data was learned, it needs to evaluation by compare with the testing data set. Then, performance evaluation will show in the form of accuracy, precision and recall. Figure 1 shows decision tree process.
3.1 Data Collection
The participant are 471 samples in total and 23 attributes. They have medical problem 254 samples and did not have medical problem 217 samples. The research use sklearn and pandas library in python.

3.2 Cleansing data
Because the data is come from real situation, the data should be clean, (8) no missing value. Some of missing value fill with the mean. The research chose 16 attributes, includings: Sex, Age, Edu, Mental, Exercise, Food, taste, Water_glass, Drink, Alcohol, Smoke, Num_smoke, Exer_type1, Hobby, Social and Careyn. The description of the attributes is in followed.

- Sex Female=0 Male=1
- Age year
- Edu below and Pratomsuksa=1, Matthayomsuks=2, Bachelor=3, upper Bachelor=4
- Mental No=0 Yes=1
- Exercise No=0 Yes=1
- Food Veg=1 Meat=2
- taste Normal=0 Fresh=1 Strong=2
- Water_glass Number of fresh water
- Drink (Coffee or Tea) No=0 Yes=1
- Alcohol No. of glass per day
- Smoke No=0 Yes=1
- Num_Smoke No. of smoke per day
- Exer_type1 stand=1, walk=2, sit=3, lie=4, altogether=5
- Hobby No=0 Yes=1
- Social No=0 Yes=1
- Class: Medproyn have medical problem No=0 Yes=1

3.3 Choose model hyperparameters
The researcher create two model, first, the model with criterion = ‘entropy’ for maximum of the layer, second, the model with criterion = ‘entropy’ with the max_depth =3.

```python
model = tree.DecisionTreeClassifier(criterion='entropy')
model = tree.DecisionTreeClassifier(criterion='entropy', max_depth=3)
```

3.4 Import data
Save data to csv file separate by comma (,). Then, Import library in python This paper use sklearn and pandas library in python for create decision tree and random forest.

```python
import pandas as pd
df = pd.read_csv('elder_tree_backup.csv')
print(df.head())
```

3.5 Separate in training and testing data set
For decision tree and random forest, this research we separated the data by random of sample into two section for training 70 percent and for testing 30 percent. We assigned X is training data set and y is testing data set. Then, the training data set need to drop class name ‘Medproyn’, medical problem. The axis = 1 is mean column.

\[
X = \text{df.drop('Medproyn', axis=1)}
\]
\[
y = \text{df['Medproyn']}
\]

3.6 Choose attributes
First, the researcher selected some attributes that possibly affect the medical problem such as sex, age, education, medicine, mental health, exercise, food, taste, number of glass of water, occasionally drink, alcohol, smoking habit, number of smoke, exercise type, have hobby, social, have carer.

\[
df = \text{df[['Sex', 'Age1', 'Edu', 'Medproyn', 'Mental', 'Exercise', 'Food', 'taste', 'Water_glass', 'Drink', 'Alcohol', 'Smoke', 'Num_Smoke', 'Exer_type1', 'Hobby', 'Social']]}\]

After the researcher checked correlation matrix, the result is the hobby is seeming lightly related with other attributes. Then, we drop attribute hobby. Figure 2 shows correlation matrix all selected attributes and Figure 3 shows correlation matrix used attributes without hobby attribute.

![Figure 2. Correlation matrix all selected attributes](image)

![Figure 3. Correlation matrix remove a hobby attribute](image)

Final attributes that in X for training is 14 attributes including : Sex = X_1, Age1=X_2, Edu=X_3, Mental=X_4, Exercise=X_5, Food=X_6, taste=X_7, Water_glass=X_8, Drink=X_9, Alcohol=X_10, Smoke=X_11, Num_Smoke= X_12, Exer_type1= X_13, Social= X_14.

3.7 splits out the data into a training set and a test set
Each instance will decide to be in training set or testing set, usually, we divide 70 percent of instance to be in training set and the rest is in testing set. The following is the code of sklearn_model_selection for split X_train and X_test for learning process and y_train and y_test for prediction.

from sklearn.model_selection import train_test_split

\[
X\_train, X\_test, y\_train, y\_test = \text{train_t}
\]
\[
test\_split(X, y, \text{random_state=99})
\]

The entropy changes when put the instances into smaller subsets. The value of change of entropy called information gain. Information gain calculates the error or the difference between entropy before split node and average entropy after split of the dataset based on given attribute values. The formula used in this research came from scikit-learn.org.
A decision tree recursively partitions, the substances with the same attributes are grouped together. Giving train vectors $x_i \in \mathbb{R}^n$, $i=1,...,l$ and an attribute vector $y \in \mathbb{R}^l$. The data at node $m$ represented by $Q$. For each substance split $\theta = (j, t_m)$ consisting of a feature $j$ and threshold $t_m$, partition the data into $Q_{left}(\theta)$ and $Q_{right}(\theta)$ subsets as in equation (2).

$$Q_{left}(\theta) = (x, y)|x_j \leq t_m$$
$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta)$$

(1) (2)

The impurity at $m$ is computed using an impurity function $H()$, the choice of which depends on the task being solved classification as in equation (3).

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))$$

(3)

Select the parameters that minimizes the impurity and recursive for both subsets until the maximum allowable depth is reached, $N_m < \text{min\_samples}$ or $N_m = 1$.

3.8 Train decision tree classifier.

model.fit(X_train, y_train)

The result from sklearn is

```python
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None,
 max_features=None, max_leaf_nodes=None,
 min_impurity_decrease=0.0, min_impurity_split=None,
 min_samples_leaf=1, min_samples_split=2,
 min_weight_fraction_leaf=0.0, presort=False,
 random_state=None, splitter='best')
```

3.9 Predict labels for unknown data.

y_predict = model.predict(X_test)

4. Results and Discussion

4.1 Create confusion metrics

from sklearn.metrics import confusion_matrix

confusion_matrix(y_test, y_predict)

results
before remove attribute ‘hobby’

```python
array([[34, 17],
       [23, 44]])
```

After remove attribute ‘hobby’
```
array([[34, 17],
       [21, 46]])
```

**Accuracy score**

```python
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_predict)
```

4.2 The accuracy score before remove attribute hobby

An attribute was removed after least correlation with other attribute. In this case we remove attribute ‘hobby’. The accuracy before remove attribute ‘hobby’ is 66.01%. The accuracy after remove attribute hobby is increasing to 67.80%.

4.3 The accuracy result from the max_depth

**Classification report**

```python
from sklearn.metrics import classification_report
print(classification_report(y_test, y_predict))
```

4.4 Before and after remove hobby

| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0         | 0.60   | 0.67     | 0.63    | 51      |
| 1         | 0.72   | 0.66     | 0.69    | 67      |

accuracy: 0.66, 118

| macro avg | 0.66 | 0.66 | 0.66 | 118 |
| weighted avg | 0.67 | 0.66 | 0.66 | 118 |

4.5 After decrease max_depth=3

| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0         | 0.67   | 0.47     | 0.55    | 51      |
| 1         | 0.67   | 0.82     | 0.74    | 67      |

accuracy: 0.67, 118

| macro avg | 0.67 | 0.65 | 0.64 | 118 |
| weighted avg | 0.67 | 0.67 | 0.66 | 118 |
The accuracy from after remove the hobby and use maximum depth gives the highest value about 68%.

4.6 Visualize the tree
The easy to explain the result of decision tree is the picture of the tree. In Scikit-learn can optimize of decision tree classifier by pre-pruning, another model we used maximum depth equal three to control for pre-pruning.

```python
from sklearn.tree import export_graphviz
from sklearn.externals.six import StringIO
from IPython.display import Image
import pydotplus

dot_data = StringIO()
export_graphviz(model, out_file=dot_data,
                filled=True, rounded=True,
                special_characters=True, class_names=['0','1'])
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
graph.write_png('ElderTree2.png')
Image(graph.create_png())
```

The result showed that the education (X3) is the root of the first classification. If the value of education is below or equal 0.5, true, entropy is 0.997, and they are in category 1, have medical condition. Overall samples can be classified are 353 samples, divided into category 0, no medical condition, 166 samples, and category 1, have medical condition, 187 samples. If root node (X3) it true, next node attribute that can be classified is X12, number of smokes. If root node (X3) is false, next node attribute that can be classified is X5, exercise. See Figure 4. Decision Tree graph with maximum depth.
Figure 4. Decision Tree graph with maximum depth

After plot a decision tree on the same data with max_depth=3 the result shows as below. Figure 5 shows Decision tree graph with maximum depth=3 layers. The result showed that the education (X_3) is the root of the first classification. If the value of education is below or equal 0.5, true, entropy is 0.997, and they are in category 1, have medical condition. Overall samples can be classified are 353 samples, divided into category 0, no medical condition, 166 samples, and category 1, have medical condition, 187 samples. If root node (X_3) it true, next node attribute that can be classified is X_{12}, number of smokes. If root node (X_3) is false, next node attribute that can be classified is X_5, exercise.
5. Conclusion and future work

This research tried to use decision tree which is a machine learning technique for classification and prediction of the possibility of medical problem occur of elder people. The elderly people data came from interview in the real situation. The research chose 16 attributes, including: sex, age, education, mental, exercise, food, taste, number of glass of water, drink coffee or tea, alcohol habit, smoke habit, number of smoke, exercise type, hobby, social and carer. An accuracy value after remove attribute hobby is about 67.80%, considered as good accuracy with the real data. This work can improve for future work by find out more parameters or other algorithm such as neural network for improving the accuracy. In addition, the future work can do about comparison among data mining techniques such as Naïve Bayes, Decision tree, k-NN, or ANN.

6. References

1. Muniyandi AP, Rajeswari R, Rajaram R. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm. Procedia Engineering. 2012 Jan 1;30:174–82.

2. Chen W, Pourghasemi HR, Kornejady A, Xie X. GIS-Based Landslide Susceptibility Evaluation Using Certainty Factor and Index of Entropy Ensembled with Alternating Decision Tree Models. In: Pourghasemi HR, Rossi M, editors. Natural Hazards GIS-Based Spatial Modeling Using Data Mining Techniques [Internet]. Cham: Springer International Publishing; 2019. p. 225–51. Available from: https://doi.org/10.1007/978-3-319-73383-8_10

3. Stromberg KA, Agyemang AA, Graham KM, Walker WC, Sima AP, Marwitz JH, et al. Using Decision Tree Methodology to Predict Employment After Moderate to Severe Traumatic Brain Injury. The Journal of Head Trauma Rehabilitation [Internet]. 2019;34(3). Available from: https://journals.lww.com/headtraumarehab/Fulltext/2019/05000/Using_Decision_Tree_Methodology_to_Predict.16.aspx

4. Settu N, Rajasekharar Babu M. Enhancing the Performance of Decision Tree Using NSUM Technique for Diabetes Patients. In: Krishna PV, Gurumoorthy S, Obaidat MS, editors. Internet of Things and Personalized Healthcare Systems [Internet]. Singapore: Springer Singapore; 2019. p. 13–20. Available from: https://doi.org/10.1007/978-981-13-0866-6_2
5. Maji S, Arora S. Decision Tree Algorithms for Prediction of Heart Disease. In: Fong S, Akashe S, Mahalle PN, editors. Information and Communication Technology for Competitive Strategies. Springer Singapore; 2019. p. 447–54.

6. Nikhar S, Karandikar AM. Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Advanced Engineering, Management and Science (IJAEMS). 2016 Jun;2(6):617–21.

7. Tang V, Siu PKY, Choy KL, Lam HY, Ho GTS, Lee CKM, et al. An adaptive clinical decision support system for serving the elderly with chronic diseases in healthcare industry. Expert Systems. 2019 Apr 1;36(2):e12369.

8. Smitha Rao MS, Pallavi M, Geetha N. Conceptual Machine Learning Framework for Initial Data Analysis. In: Peng S-L, Dey N, Bundele M, editors. Computing and Network Sustainability. Springer Singapore; 2019. p. 51–9.