Ensemble Modeling Method to Predict Life Expectancy of Population in High-Income Countries: Japan and Finland

Nittaya Kerdprasop, Kittisak Kerdprasop
Suranaree University of Technology
Nakhon Ratchasima, Thailand
{nittaya,kerdpras}@sut.ac.th

Paradee Chuaybamroong
Thammasat University
Pathum Thani, Thailand
paradee@tu.ac.th

***Abstract*—Life expectancy at birth is an indicator defined by the United Nations Development Program (UNDP) as the number of years, on average, an infant is expected to live. This indicator is a proxy of good health. The health index together with the education and income indices are used by UNDP for measuring the development level of the member countries. In addition to improve human development along the health dimension, most governments also need the accurate projection of life expectancy of their populations for the effective social services and decent pension planning. In this work, we propose a data-driven modeling method to predict life expectancy. Our method is based on the ensemble scheme in which a combination of classification and regression tree (CART) and the chi-square automatic interaction detection (CHAID) algorithms are applied for making a cooperative prediction. We empirically prove that the proposed ensemble scheme is more accurate than a single model prediction. We experiment our modeling methodology with the life expectancy data of the two high-income countries: Japan and Finland. This selection is due to the fact that these two countries are in the group of very high human development according to the latest UNDP ranking report. The CART and CHAID models reveal that both economic and environmental factors share their contributions to forecasting life expectancy of populations in the two countries. Forest depletion, agricultural methane and CO_2 emissions, particulate emission damage, national income, and education expenditure are factors affecting longevity of Japanese population. To predict the Finn's life expectancy, the ensembled models consider several factors including exports and imports of goods and services, electric power consumption, energy use, national income, GDP growth, education expenditure, forest area, agricultural methane emission, and particulate emission damage.**

I. Introduction

From the second half of the 2oth century, development of the nations had been measured based on gross domestic product (GPD) per capita as a sole indicator [1]. Since 1990 up to this current year in the 21st century, the United Nations Development Program (UNDP) has expanded the indicator to cover the three main dimensions related human well-being [2]. These dimensions include long and healthy life, knowledge through education, and sufficient income to reach sensible living standard.

UNDP has introduced the measurement called the Human Development Index (HDI) [3] that is the geometric mean of indices: $(\text{health\_index} \times \text{education\_index} \times \text{income\_index})^{1/3}$. Indices along these three dimensions are computed from the normalized values of the four indicators: life expectancy at birth, expected years of schooling, means years of schooling, and gross national income (GNI) per capita. Among these indicators, life expectancy at birth is the main focus of our research because it is quite a subtle index compared to the education and income indices.

Life expectancy at birth has been defined by UNDP [4] as number of years, on average, a newborn baby is expected to live. This measurement is based on mortality pattern across all age groups and it is assumed that these mortality rates remain the same throughout the life of the newborn baby. Life expectancy at birth is a standardized measurement often used as an indicator to gauge population health [5], [6] and to assess longevity trends of people in the nation and across nations [7], [8], [9], [10]. Life expectancy is also one of the most important factors to consider for optimal actuarial and pension planning [11], [12], [13].

The association of life expectancy as a main part of HDI and economic growth level had been studied by Suri *et al.* [14] using path analysis to derive causal relationships, and explored by Wang *et al.* [15] using correlation analysis. Correlations between life expectancy at birth and household energy consumption in China are also investigated by several researchers [16], [17]. Based on the energy consumption analysis results, domestic coal usage has negative impact on life expectancy, whereas household electricity utilization shows positive correlations to life expectancy at birth. The authors point out that the negative impacts of household coal are more serious in the western provinces than in the east. This spatial analysis is in accordance to the distribution of economic growth areas in China.

Besides economic and socio-economic impacts [18], [19], [20], environment is another important factor affecting life expectancy of populations. A wide range of environmental factors that show some impact to life expectancy include carbon dioxide (CO_2) emission [21], [22], [23], [24], particulate matter (PM_{10}) and sulfur dioxide (SO_2) concentrations [25], [26], and climate conditions [27], [28].

The impacts of economics, environment, and other factors on the change in life expectancy trend had been mostly studied through modeling methods. The most simple form of modeling applied to forecast life expectancy is linear regression [29]. Another statistical-based modeling method adopted for life expectancy forecasting is the autoregressive integrated moving average, or ARIMA [30], [31], [32]. The advanced machine learning methods are also applied to model life expectancy of populations. These machine learning forecasting models are based on the feedforward neural network [33] and extreme learning machine [34], which is the extension of neural network algorithm. Some researchers also consider applying the ensemble strategy to make a forecast by firstly creating many forecasting models, and then combining the results through the averaging technique [35].

In this work, we explore the ensemble scheme using the classification and regression tree (CART) [36] and the chi-square automatic interaction detection (CHAID) [37] as the base algorithms for predicting the number of years a newborn baby is expected to live, which is the target of our prediction. The fifteen economic and environmental attributes are used as predictors. The selection of CART and CHAID algorithms is due to the support for reasoning, which is the advantage inherent in most tree-based learning algorithms [38]. Our data source, selected data attributes, and modeling methods are explained in the next section. Results from preliminary data exploration and the generated models are demonstrated in Section 3. Performance of the models are then evaluated and shown in Section 4. We finally conclude our work in Section 5.

II. Material and method

A. Data source and attribute meaning

The life expectancy of populations in Japan and Finland together with other fifteen development indicators used as training data for creating a predictive model are extracted from the databank of World Bank [39]. The time-series data range from 1970 to 2017. The trends in life expectancy of people in Japan and Finland during these 48 years are graphically compared in Fig. 1. Details of data attributes and the meaning of each development indicator are summarized in Table I.

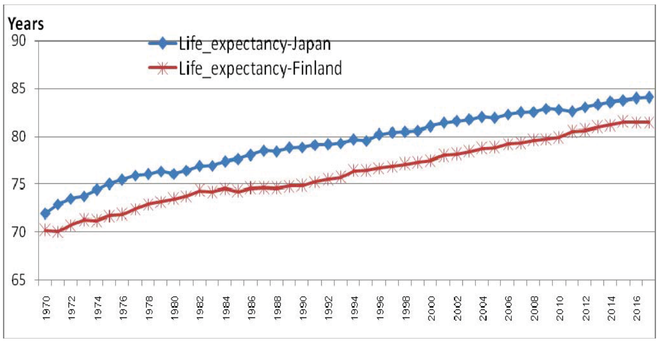

Fig. 1. Life expectancy of populations in Japan and Finland during 1970-2017

TABLE I. Data attributes and their explanation

Attribute name	Development indicator and meaning
National_income	Adjusted net national income per capita (annual % growth) -- It is GNI minus consumption of fixed capital and natural resources depletion
Education_expense	Adjusted savings: education expenditure (% of GNI) -- including wages and salaries, excluding capital investments in buildings and equipment
Forest_depletion	Adjusted savings: net forest depletion (% of GNI) -- calculated as the product of unit resource rents and the excess of roundwood harvest over natural growth; if growth exceeds harvest, this figure is zero
Particulate_emission_damage	Adjusted savings: particulate emission damage (% of GNI) -- the damage due to exposure to ambient concentrations of particulates measuring less than 2.5 microns in diameter ($PM_{2.5}$), ambient ozone pollution, and indoor concentrations of $PM_{2.5}$ in households cooking with solid fuels. Damages are calculated as foregone labor income due to premature death.
Agri_methane_emission	Agricultural methane emissions (% of total) -- emissions from animals, animal waste, rice production, agricultural waste burning
CO2_emission	CO_2 emissions (metric tons per capita) -- from the burning of fossil fuels and the manufacture of cement, including carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring
Electric_power_consumption	Electric power consumption (kWh per capita) -- the production of heat and power plants
Energy_use	Energy use (kg of oil equivalent per capita) -- use of primary energy before transformation to other end-use fuels
Exports	Exports of goods and services (% of GDP) -- the value of all merchandise, freight, insurance, transport, travel, royalties, license fees, and other services provided to the rest of the world
Forest_area	Forest area (% of land area) -- land under natural or planted stands of trees of at least 5 meters in situ, whether productive or not, and excludes tree stands in agricultural production systems and trees in parks and gardens
GDP_growth	GDP growth (annual %) -- sum of gross value added by all resident producers in the economy
GNI	GNI per capita growth (annual %) -- sum of value added by all resident producers plus any product taxes plus net receipts of primary income from abroad
Hi-tech_exports	High-technology exports (% of manufactured exports) -- products with high R&D intensity, such as in aerospace, computers, pharmaceuticals, scientific instruments, and electrical machinery
Imports	Imports of goods and services (% of GDP) -- the value of all goods and other market services received from the rest of the world
Industry	Industry, value added (% of GDP) -- value added in mining, manufacturing, construction, electricity, water, and gas
Life_expectancy	Life expectancy at birth, total (years) -- the number of years a newborn infant would live if patterns of mortality at the time of its birth were the same throughout its life

B. Modeling method for life expectancy prediction

Our ensemble modeling method is composed of four main phases: data preparation, data exploration, data modeling, and model evaluation. Details in each phase are as follows.

Data preparation phase covers the first two steps as shown in Fig.2. The first step is database access and data extraction. From the World Bank data source [39], there are in total 1,599 indicators to assess development of the nation. We extract only 16 indicators including life expectancy at birth that is going to be used as the target of our prediction models. The fifteen world development indicators used as predictors are those concerning economics, manufacturing, health, education, and environment. Details of these indicators can be found in Table I.

Data exploration phase is the next step (step 3 in Fig. 2) following data extraction. This phase is for the understanding of data characteristics, importance of attribute towards the value of target attribute, and relations that exist among data attributes. We apply correlation analysis to study associations of the data attributes.

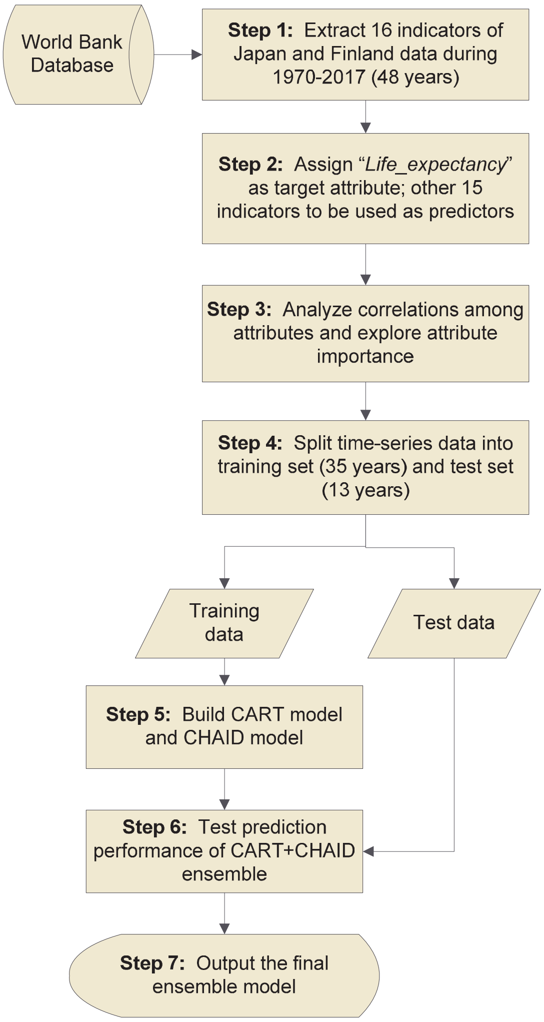
World Bank Database
Step 1: Extract 16 indicators of Japan and Finland data during 1970-2017 (48 years)
Step 2: Assign "*Life_expectancy*" as target attribute; other 15 indicators to be used as predictors
Step 3: Analyze correlations among attributes and explore attribute importance
Step 4: Split time-series data into training set (35 years) and test set (13 years)
Training data
Test data
Step 5: Build CART model and CHAID model
Step 6: Test prediction performance of CART+CHAID ensemble
Step 7: Output the final ensemble model

Fig. 2. Steps in predictive model creation

Data modeling phase comprises of steps 4 and 5 in Fig. 2. In step 4, we split the dataset into two subsets with the proportion approximately 70:30. The larger proportion is for training the algorithms to build models, while the smaller subset is held out for testing model in a later phase. The training dataset contains 35 data records which are used as input for the CART and CHAID algorithms to build the prediction model in step 5. Therefore, both algorithms use the same set of data for model building.

Model evaluation occurs at step 6 of our modeling process. We adopt the out-of-sample method in which the test data are unseen by the learning algorithms to evaluate the performance of CART and CHAID models. To assess model performance, we make a comparison of our CART and CHAID ensemble against the model built with statistical learning algorithm, that is regression, and other ensemble scheme of the base models. After the confirmation of model performance, the ensemble of CART and CHAID is produced as the output of our modeling process in step 7.

III. DATA EXPLORATION AND MODELING RESULTS

A. Data exploration results

The results from exploratory phase in step 3 of our modeling method are the predictor importance and correlation analyses. Predictor importance measures the contribution of independent variables toward the value of a target variable. The importance scale is in the range [0,1]. The higher the value, the more important the variable. Results of predictor importance analysis are graphically shown in Fig. 3. To consider life expectancy of Japanese population, the four most important predictors are *CO2_emission*, *National_income*, *Forest_depletion*, and *Agri_methane_emission*. For the Finn people, a single most important predictor is *Agri_methane_emission*.

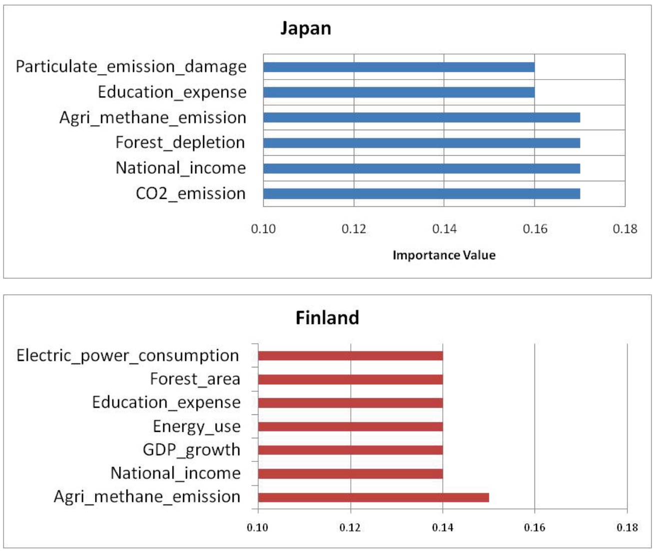

Fig. 3. Analysis results of predictor importance to the prediction of life expectancy of population in Japan (above) and Finland (below)

TABLE II. CORRELATIONS AMONG DATA ATTRIBUTES

	NI (*National_income*)	EE	FD	PD	AM	CE	EC	EU	EX	FA	GG	GN	HE	IM	IN
EE (*Education_expense*)															
JAP	**0.9**														
FIN	**0.9**														
FD (*Forest_depletion*)															
JAP	**0.9**	**1.0**													
FIN	**0.9**	**1.0**													
PD (*Particulate_emission_damage*)															
JAP	0.4	0.4	0.4												
FIN	0.4	0.4	0.4												
AM (*Agri_methane_emission*)															
JAP	0.5	0.6	0.6	0.0											
FIN	0.5	0.6	0.6	0.0											
CE (*CO2_emission*)															
JAP	-0.5	-0.5	-0.5	-0.3	-0.6										
FIN	-0.5	-0.5	-0.5	-0.1	**-0.7**										
EC (*Electric_power_consumption*)															
JAP	-0.6	-0.6	-0.6	-0.6	-0.4	**0.9**									
FIN	-0.6	-0.5	-0.5	-0.6	-0.4	**0.7**									
EU (*Energy_use*)															
JAP	-0.6	-0.6	-0.6	-0.5	-0.5	**0.8**	**0.9**								
FIN	-0.6	-0.5	-0.5	-0.5	-0.4	**0.8**	**0.9**								
EX (*Exports*)															
JAP	-0.3	-0.3	-0.3	-0.3	0.1	-0.0	0.0	0.0							
FIN	-0.6	-0.5	-0.5	**-0.8**	-0.1	0.3	**0.7**	0.6							
FA (*Forest_area*)															
JAP	0.4	0.4	0.4	**0.9**	0.0	-0.4	**-0.7**	-0.6	-0.2						
FIN	0.4	0.4	0.4	**0.9**	0.0	-0.2	**-0.7**	-0.6	**-0.7**						
GG (*GDP_growth*)															
JAP	0.2	0.2	0.2	0.1	0.1	-0.2	-0.2	-0.2	-0.0	0.1					
FIN	0.2	0.2	0.2	0.1	0.1	-0.2	-0.1	-0.2	-0.1	0.1					
GN (*GNI*)															
JAP	**1.0**	**0.9**	**0.9**	0.4	0.5	-0.5	-0.6	-0.6	-0.3	0.4	0.2				
FIN	**1.0**	**0.9**	**0.9**	0.4	0.5	-0.5	-0.6	-0.6	-0.6	0.4	0.2				
HE (*Hi-tech_exports*)															
JAP	0.5	0.4	0.4	**0.9**	0.0	-0.3	-0.6	-0.5	-0.2	**0.9**	0.1	0.5			
FIN	0.5	0.4	0.4	**0.9**	0.0	-0.2	-0.6	-0.5	-0.7	**0.9**	0.1	0.5			
IM (*Imports*)															
JAP	-0.2	-0.2	-0.2	-0.2	0.3	-0.1	0.0	-0.0	**0.9**	-0.2	-0.0	-0.2	-0.2		
FIN	-0.5	-0.5	-0.5	-0.6	0.0	0.1	0.5	0.4	**0.8**	-0.6	-0.1	-0.5	-0.6		
IN (*Industry*)															
JAP	0.4	0.3	0.3	**0.8**	-0.0	-0.2	-0.5	-0.4	-0.4	**0.8**	0.1	0.4	**0.8**	-0.4	
FIN	**0.8**	**0.7**	**0.7**	0.5	0.4	-0.4	-0.6	-0.6	-0.6	0.5	0.2	**0.8**	0.6	-0.6	
LE (*Life_expectancy*)															
JAP	**-0.7**	**-0.7**	**-0.7**	**-0.8**	-0.2	0.4	0.6	0.6	0.5	**-0.7**	-0.2	**-0.7**	**-0.8**	0.4	**-0.7**
FIN	-0.6	-0.6	-0.6	**-0.8**	-0.0	0.1	0.6	0.5	**0.8**	**-0.7**	-0.2	-0.6	**-0.8**	**0.8**	**-0.7**

Another result of data exploratory step is the correlation analysis of all variables used in this study, which are shown in Table II. Variable correlations of the two countries are displayed in consecutive lines for the ease of comparison. It is noticeable that correlations of the four variables (*Education_expense*, *Forest_depletion*, *Particulate_emission_damage*, *Agri_methane _emission*) to other variables of Japan and Finland are exactly the same. Correlations of *GDP_growth* and *GNI* to other variables of the two countries are also quite resemble.

For the strong correlation cases, we highlight them with bold red font. The strongest correlation appears to be the same pair of variables in both Japan and Finland, which is the positive relation between *Education_expense* and *Forest_depletion*.

The strongest negative impacts (correlation = -0.8) to life expectancy of population in the two countries are the same factors which are *Particulate_emission_damage* and *Hi-tech_exports*. For Finn people, and *Exports* and *Imports* show strong positive impact (correlation = 0.8) toward longevity.

B. Life expectation prediction models for Japanese population

To predict life expectancy of Japanese population, both CART and CHAID models (in Fig. 4) are applied. The models are in a form of decision tree structure in which the root node (on the leftmost) is the first condition to consider. The next descendents are in the next indentation level, and so on. The leaf nodes appear on the rightmost of a tree indicating the predicted number of years of life expectancy.

Prediction from the CART+CHAID ensemble can be done independently using a CART and a CHAID model, then averaging the results to be the final output. For instance, the longest life expectancy, on average, of people in Japan predicted by the CART model is 83.658 years of age, while the prediction from the CHAID model is 83.985. Therefore, the final prediction from the ensemble model is 83.8215 years.

The CART+CHAID ensemble also delivers information regarding factors contributing to the longest life expectancy of population. In Japan, these factors are:

CART: *Forest_depletion* ≤ 0.014 % of GNI, and
Agri_methane_emission > 70.628 % of total, and
Particulate_emission_damage ≤ 0.092 % of GNI.

CHAID: *Forest_depletion* ≤ % of GNI, and
Agri_methane_emission > 71.711 % of total, and
National_income in a range [0.524,1.648] of annual % growth.

CART Model	CHAID model

Forest_depletion ≤ 0.014
  Particulate_emission_damage ≤ 0.127
    Agri_methane_emission ≤ 70.628 **=> 81.49**
    Agri_methane_emission > 70.628
      Particulate_emission_damage ≤ 0.092 **=> 83.658**
      Particulate_emission_damage > 0.092 **=> 82.639**
  Particulate_emission_damage > 0.127
    Agri_methane_emission ≤ 62.339
      Education_expense ≤ 4.091 **=> 78.827**
      Education_expense > 4.091 **=> 78.316**
    Agri_methane_emission > 62.339 **=> 79.918**
Forest_depletion > 0.014
  Education_expense ≤ 3.200
    National_income ≤ 5002.667
      National_income ≤ 0.854 **=> 74.394**
      National_income > 0.854 **=> 73.507**
    National_income > 5002.667 **=> 71.95**
  Education_expense > 3.200
    Agri_methane_emission ≤ 61.480 **=> 76.414**
    Agri_methane_emission > 61.480 **=> 76.065**
	Forest_depletion ≤ 0
	  Agri_methane_emission ≤ 71.711
	    National_income ≤ -0.441 **=> 81.563**
	    National_income > -0.441 and ≤ 0.524 **=> 81.417**
	    National_income > 0.524 **=> 81.76**
	  Agri_methane_emission > 71.711
	    National_income ≤ -0.441 **=> 82.591**
	    National_income > -0.441 and ≤ 0.524 **=> 83.588**
	    National_income > 0.524 and ≤ 1.648 **=> 83.985**
	    National_income > 1.648 **=> 83.332**
	Forest_depletion > 0 and ≤ 0.004
	  Particulate_emission_damage ≤ 0.116
	    National_income ≤ -1.856 **=> 82.931**
	    National_income > -1.856 and ≤ 0.524 **=> 82.322**
	    National_income > 0.524 and ≤ 1.648 **=> 82.507**
	    National_income > 1.648 **=> 82.843**
	  Particulate_emission_damage > 0.116 and ≤ 0.135
	    National_income ≤ -1.856 **=> 80.501**
	    National_income > -1.856 and ≤ -0.441 **=> 80.571**
	    National_income > -0.441 **=> 81.076**
	  Particulate_emission_damage > 0.135]
	    Education_expense ≤ 3.400 **=> 79.294**
	    Education_expense > 3.400 and ≤ 3.572 **=> 79.536**
	    Education_expense > 3.572 **=> 79.687**
	Forest_depletion > 0.004 and ≤ 0.008
	  National_income ≤ -0.029 **=> 79.154**
	  National_income > -0.029 and ≤ 4.411 **=> 78.818**
	  National_income > 4.411 and ≤ 5.641 **=> 78.837**
	  National_income > 5.641 **=> 78.399**
	Forest_depletion > 0.008 and ≤ 0.026
	  Education_expense ≤ 4.082 **=> 76.33**
	  Education_expense > 4.082
	    National_income ≤ 2.462 **=> 78.065**
	    National_income > 2.462 **=> 78.484**
	Forest_depletion > 0.026 and ≤ 0.041
	  National_income ≤ -1.856 **=> 74.394**
	  National_income > -1.856 and ≤ -0.441 **=> 76.092**
	  National_income > -0.441 and ≤ 2.462 **=> 75.057**
	  National_income > 2.462 **=> 75.457**
	Forest_depletion > 0.041
	  CO2_emission ≤ 7.411 **=> 71.95**
	  CO2_emission > 7.411 and ≤ 7.773 **=> 72.883**
	  CO2_emission > 7.773 **=> 73.507**

Fig. 4. Ensemble model to predict life expectancy of population in Japan

C. Life expectation prediction models for Finn population

The interpretation of the CART+CHAID ensemble for the case of population in Finland can be done in the same manner as the Japanese case. The longest life expectancy, on average, predicted by the CART model is 81.202, but a little longer at 81.429 as predicted by the CHAID model. Thus, the final prediction of the longest life expectancy of Finn population is (81.202+81.429)/2, which is 81.3155 years. This is around 2.5 years shorter than the life expectancy of Japanese people.

The CART+CHAID ensemble reveals that the factors resulting in long life of Finn population are:

CART:

Exports of goods and services > 32.697 % of GDP, and
Imports of goods and services > 32.407 % of GDP, and
Particulate_emission_damage ≤ 0.036 % of GNI.

CHAID:

Agri_methane_emission > 33.784 % of total, and
National_income > 2.961 of annual % growth.

CART Model

```
Exports ≤ 32.697
    Electric_power_consumption ≤ 6735.301
        National_income ≤ 5002.840
            National_income ≤ 3.365                        => 71.135
            National_income > 3.365                        => 70.707
        National_income > 5002.840                         => 70.18
    Electric_power_consumption > 6735.301
        Industry ≤ 32.009
            Education_expense ≤ 6.400                      => 74.667
            Education_expense > 6.400                      => 75.705
        Industry > 32.009
            Electric_power_consumption ≤ 7813.100          => 72.897
            Electric_power_consumption > 7813.100          => 73.593
Exports > 32.697
    Imports <= 32.407
        Energy_use ≤ 6364.976                              => 76.797
        Energy_use > 6364.976
            National_income ≤ 2.024                        => 78.12
            National_income > 2.024                        => 77.966
    Imports > 32.407
        Particulate_emission_damage ≤ 0.040
            Particulate_emission_damage ≤ 0.036            => 81.202
            Particulate_emission_damage > 0.036            => 80.471
        Particulate_emission_damage > 0.040
            National_income ≤ 2.928                        => 79.795
            National_income > 2.928                        => 79.239
```

CHAID model

```
Agri_methane_emission ≤ 21.062
    Forest_area ≤ 72.916
        Agri_methane_emission ≤ 20.383                     => 79.215
        Agri_methane_emission > 20.383                     => 79.263
    Forest_area > 72.916 and ≤ 73.491
        National_income ≤ 0.062                            => 78.368
        National_income > 0.062 and ≤ 1.250                => 78.12
        National_income > 1.250                            => 77.966
    Forest_area > 73.491
        Education_expense ≤ 5.470                          => 77.466
        Education_expense > 5.470                          => 77.291
Agri_methane_emission > 21.062 and ≤ 22.906
    Energy_use ≤ 5960.693                                  => 76.41
    Energy_use > 5960.693 and ≤ 6262.192                   => 76.396
    Energy_use > 6262.192                                  => 77.091
Agri_methane_emission > 22.906 and ≤ 30.229
    GDP_growth ≤ -0.758                                    => 75.455
    GDP_growth > -0.758 and ≤ 0.344                        => 75.705
    GDP_growth > 0.344 and ≤ 1.680                         => 74.813
    GDP_growth > 1.680 and ≤ 2.772                         => 74.56
    GDP_growth > 2.772 and ≤ 4.207                         => 74.592
    GDP_growth > 4.207 and ≤ 5.185                         => 74.792
    GDP_growth > 5.185                                     => 74.577
Agri_methane_emission > 30.229 and ≤ 31.720
    National_income ≤ 1.250                                => 73.747
    National_income > 1.250 and ≤ 1.896                    => 74.201
    National_income > 1.896 and ≤ 2.961                    => 73.44
    National_income > 2.961                                => 73.155
Agri_methane_emission > 31.720 and ≤ 33.784
    Electric_power_consumption ≤ 4889.504
        National_income ≤ 1.250                            => 70.018
        National_income > 1.250                            => 70.18
    Electric_power_consumption > 4889.504 and ≤ 6545.747
        National_income ≤ -1.049                           => 71.674
        National_income > -1.049 and ≤ 0.062               => 71.135
        National_income > 0.062 and ≤ 1.896                => 71.813
        National_income > 1.896                            => 70.707
    Electric_power_consumption > 6545.747
        National_income ≤ -1.049                           => 72.35
        National_income > -1.049                           => 72.897
Agri_methane_emission > 33.784
    National_income ≤ -2.324                               => 79.72
    National_income > -2.324 and ≤ -1.049                  => 80.976
    National_income > -1.049 and ≤ 0.062                   => 81.18
    National_income > 0.062 and ≤ 1.250                    => 80.471
    National_income > 1.250 and ≤ 2.961                    => 79.871
    National_income > 2.961                                => 81.429
```

Fig. 5. Ensemble model to predict life expectancy of population in Finland

It can be noticed from the CART+CHAID ensembles of both Japan and Finland that the highest level of agricultural methane emission but the lowest level of particulate emission damage show their significant contributions to the long living of populations. This information conveys the importance of agricultural sector in the two countries as well as reflects the damage caused by air pollution.

IV. Model evaluation

The final step of our modeling process is the comparative experimentation to confirm efficiency of the proposed CART+CHAID ensemble. The two metrics used for efficiency confirmation are prediction error and correlation of the model. For the prediction error test, we look for the lowest error measured in terms of mean absolute error (MAE). For the correlation test, on the contrary, we search for the highest one.

We compare MAE and correlation performances of the single modeling method against the ensemble scheme. Algorithms in the single model method include CART, CHAID, and regression. The ensemble scheme is the combinations of the single learning methods including CART+CHAID, CART+Regression, CHAID+Regression, and combination of all three algorithms, which is CART+CHAID+Regression. Comparative results of these learning schemes in the dimension of MAE are graphically presented in Fig. 6. Correlation is normally less concerned by most data analysts, but it more or less shows fitness of the model prediction and predictors. We thus also demonstrate correlation comparative results in Fig. 7.

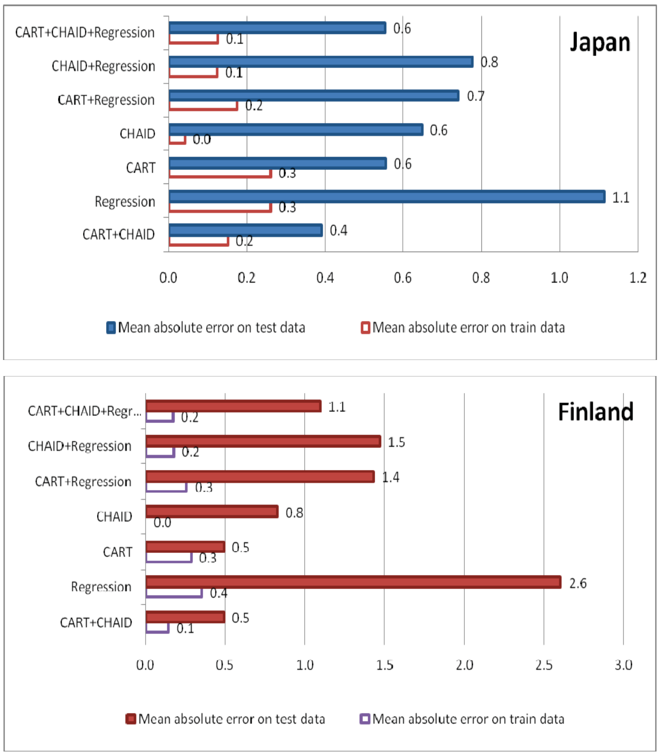

Fig. 6. Prediction performance of CART+CHAID ensemble against other learning schemes

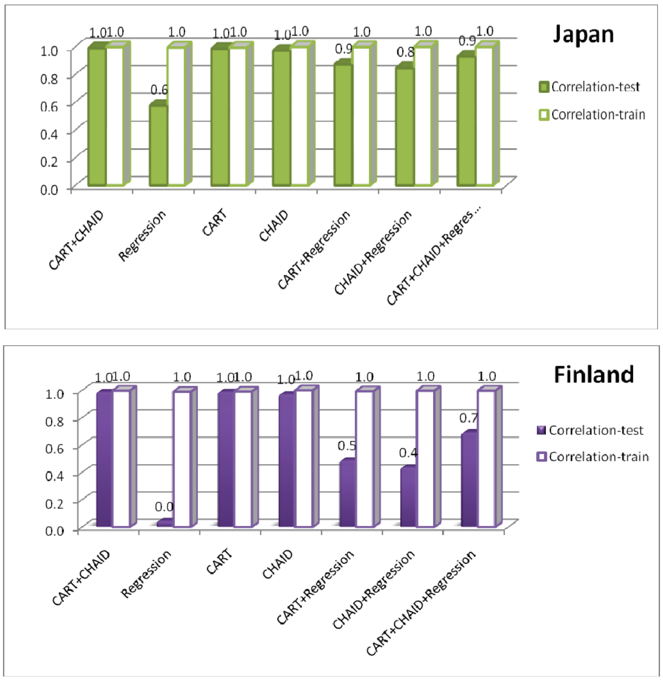

Fig. 7. Correlation comparison of the models built from the CART+CHAID ensemble and other algorithms

To make a fair comparison, we evaluate model performance with the same set of test data which are unseen by the learning algorithms. The focus of comparison is on the performance that model can achieve on predicting test data. However, model performance on training data are also presented with the intention to assess overfitting characteristic of the model. Overfitting is the problem that most learning algorithms try to avoid. This problem occurs when a model perform the best on the training data, but it work badly on the unseen test data. It can be notices from Fig. 6 that the overfitting problem occurs with regression algorithm and the ensembles of CART+Regression and CHAID+Regression. The CART+CHAID does not face such problem.

The CART+CHAID shows the best prediction performance with the lowest MAE measured on the test dataset in the Japan population data. For Finland data, CART+CHAID performs as good as a single CART model. But on the same set of training data, CART generate a model with higher mean absoluter error than the ensemble CART+CHAID.

The correlations shown in Fig. 7 have been round up to a single decimal digit for a clean picture. The precise value is 0.99 for both CART+CHAID ensemble and CART algorithm on test data of Japan and Finland. This measurement confirms the fitness of the predictors on the target variable.

The regression algorithm shows the worst performance in terms of correlation as it can generate the model with correlation equals to 0.99 on training data. But on test data, the correlation drops to 0.58 for the Japan test set and 0.04 on the Finland test set. Much decrease in correlation from train to test set reveals unstability of the regression learning algorithm.

V. Conclusion

We present the ensemble model of the two tree-based learning algorithms, classification and regression tree (CART) and chi-square automatic interaction detection (CHAID), to cooperatively predict the number of years, on average, a newborn baby is expected to live, also known as life expectancy at birth. This knowledge is not only important for government and actuaries to plan social services and pension policies, but also necessary for the United Nations Development Program (UNDP) to be used as one indicator for evaluating human development level of the member countries. According to the 2018 ranking report of UNDP, Japan and Finland are in the top-twenty of very high human development countries. We are thus select the two countries as our case study for analyzing factors having prominent impact to life expectancy of the population.

Besides life expectancy at birth, we also extract other fifteen indicators from the World Bank database to perform modeling and comparative experimentation. The fifteen indicators used as predictors of our model cover economic, health, and environment factors. The advantages of tree-based models that we adopt for our ensemble modeling are their high accuracy on prediction and the reasoning facility such that the predicted result can be traced back to find out the reason or condition leading to such prediction.

Based on this reasoning ability, we can investigate from our CART+CHAID ensemble model that the two important factors leading to longevity of populations in Japan and Finland are high proportion of agricultural sector including animal farming and a low particulate emission damage, which is the damage due to exposure to $PM_{2.5}$ concentration and ozone pollution. Such damage is computed as percentage of gross national income loss due to premature death, based on mortality rates, of labor. Other factors affecting long life of population in Japan are forest depletion, national income, education expense. For population in Finland, other important factors are high values of import and export of goods and services, energy use, electric power consumption, GDP growth, and the value added in terms of percentage of GDP of manufacturing and industrial sector of the country. From these findings of longevity pattern, we thus plan to adopt machine learning techniques to further investigate a wide range countries across different group of incomes.

Acknowledgment

This work has been supported by National Research Council of Thailand and Suranaree University of Technology through the funding of Data and Knowledge Engineering Research Unit, in which the first and second authors are principal researchers.

References

[1] D.K. Despotis, “Measuring human development via data envelopment analysis: the case of Asia and the Pacific”, *Omega*, vol. 33, no. 5, Oct. 2005, pp. 385–390.

[2] E.A. Stanton, “The human development index: a history”, *PERI Working Papers*, no.127, Political Economy Research Institute, University of Massachusetts Amhurst, Feb. 2007, pp. 1-36.

[3] UNDP, “Human development indices and indicators: 2018 statistical update”, *Technical note*, 2018. http://hdr.undp.org/sites/default/files/hdr2018_technical_notes

[4] UNDP, “Human development report”, 2019. http://hdr.undp.org/en/composite/HDI

[5] D. Dicker *et al.*, “Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: A systematic analysis for the Global Burden of Disease Study 2017”, *The Lancet*, vol. 392, no. 10159, Nov. 2018, pp. 1684–1735.

[6] M. Kim and Y.-H. Khang, “Why do Japan and South Korea record very low levels of perceived health despite having very high life expectancies?,” *SSRN Electronic Journal*, no. 3276420, 2018.

[7] L. Wang *et al.*, “Regional aging and longevity characteristics in China”, *Archives of Gerontology and Geriatrics*, vol. 67, Nov. 2016, pp. 153–159.

[8] S. Wang, K. Luo, R. Ni, Y. Tian, and X. Gao, “Assessment of elemental background values and their relation with lifespan indicators: A comparative study of Jining in Shandong Province and Guanzhong area in Shaanxi Province, northern China”, *Science of The Total Environment*, vol. 595, Oct. 2017, pp. 315–324.

[9] L. A. Johnston, “The economic demography transition: is china’s ‘not rich, first old’ circumstance a barrier to growth?”, *Australian Economic Review*, Jul. 2019, pp. 1–21.

[10] R. Bai *et al.*, “Trends in life expectancy and its association with economic factors in the belt and road countries—evidence from 2000–2014”, *International Journal of Environmental Research and Public Health*, vol. 15, no. 12, Dec. 2018, p. 2890.

[11] E. Kurtbegu, “Replicating intergenerational longevity risk sharing in collective defined contribution pension plans using financial markets”, *Insurance: Mathematics and Economics*, vol. 78, Jan. 2018, pp. 286–300.

[12] E. Debonneuil, S. Loisel, and F. Planchet, “Do actuaries believe in longevity deceleration?”, *Insurance: Mathematics and Economics*, vol. 78, Jan. 2018, pp. 325–338.

[13] L. Mayhew, D. Smith, and D. Wright, “The effect of longevity drift and investment volatility on income sufficiency in retirement”, *Insurance: Mathematics and Economics*, vol. 78, Jan. 2018, pp. 201–211.

[14] T. Suri, M. A. Boozer, G. Ranis, and F. Stewart, “Paths to success: The relationship between human development and economic growth”, *World Development*, vol. 39, no. 4, Apr. 2011, pp. 506–522.

[15] S. Wang *et al.*, “Economic level and human longevity: Spatial and temporal variations and correlation analysis of per capita GDP and longevity indicators in China”, *Archives of Gerontology and Geriatrics*, vol. 61, no. 1, Jul. 2015, pp. 93–102.

[16] S. Wang and K. Luo, “Life expectancy impacts due to heating energy utilization in China: Distribution, relations, and policy implications”, *Science of The Total Environment*, vol. 610–611, Jan. 2018, pp. 1047–1056.

[17] S. Wang, Y. Liu, C. Zhao, and H. Pu, “Residential energy consumption and its linkages with life expectancy in mainland China: A geographically weighted regression approach and energy-ladder-based perspective”, *Energy*, vol. 177, Jun. 2019, pp. 347–357.

[18] W. C. Cockerham and Y. Yamori, “Okinawa: an exception to the social gradient of life expectancy in Japan”, *Asia Pacific Journal of Clinical Nutrition*, vol. 10, no. 2, 2001, pp. 154–158.

[19] J. Jiang, L. Luo, P. Xu, and P. Wang, “How does social development influence life expectancy? A geographically weighted regression analysis in China”, *Public Health*, vol. 163, Oct. 2018, pp. 95–104.

[20] C. Lee and M. Kim, “The relationship between internet environment and life expectancy in Asia”, *Review of Integrative Business & Economics*, vol. 8, no. 2, 2019, pp. 70–80.

[21] K. Hassan and R. Salim, “Population ageing, income growth and CO_2 emission: empirical evidence from high income OECD countries”, *Journal of Economic Studies*, vol. 42, no. 1, Jan. 2015, pp. 54–67.

[22] J.-C. Yeh and C.-H. Liao, “Impact of population and economic growth on carbon emissions in Taiwan using an analytic tool STIRPAT”, *Sustainable Environment Research*, vol. 27, no. 1, Jan. 2017, pp. 41–48.

[23] N. Kerdprasop and K. Kerdprasop, “Regression tree analysis of CO_2 emissions and environmental factors to the survival rate of population in Thailand and China”, in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2016, pp. 286–290.

[24] N. Kerdprasop and K. Kerdprasop, "Association of economic and environmental factors to life expectancy of people in the Mekong basin", in *Proceedings of 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, Siem Reap, Cambodia, 2017, pp. 1984–1989.

[25] L. Wang *et al.*, "A study of air pollutants influencing life expectancy and longevity from spatial perspective in China", *Science of The Total Environment*, vol. 487, Jul. 2014, pp. 57–64.

[26] W. Song, Y. Li, Z. Hao, H. Li, and W. Wang, "Public health in China: An environmental and socio-economic perspective", *Atmospheric Environment*, vol. 129, Mar. 2016, pp. 9–17.

[27] J. Lv, W. Wang, and Y. Li, "Effects of environmental factors on the longevous people in China", *Archives of Gerontology and Geriatrics*, vol. 53, no. 2, Sep. 2011, pp. 200–205.

[28] J.-M. Robine *et al.*, "Exploring the impact of climate on human longevity", *Experimental Gerontology*, vol. 47, no. 9, Sep. 2012, pp. 660–671.

[29] G. Gulis, "Life expectancy as an indicator of environmental health", *European Journal of Epidemiology*, vol. 16, 2000, pp. 161–165.

[30] T. Torri and J. W. Vaupel, "Forecasting life expectancy in an international context", *International Journal of Forecasting*, vol. 28, no. 2, Apr. 2012, pp. 519–531.

[31] K. J. Foreman *et al.*, "Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: Reference and alternative scenarios for 2016–40 for 195 countries and territories", *The Lancet*, vol. 392, no. 10159, Nov. 2018, pp. 2052–2090.

[32] M. D. Pascariu, V. Canudas-Romo, and J. W. Vaupel, "The double-gap life expectancy forecasting model", *Insurance: Mathematics and Economics*, vol. 78, Jan. 2018, pp. 339–350.

[33] M. Kanevski *et al.*, "Environmental data mining and modelling based on machine learning algorithms and geostatistics", *Environment Modelling & Software*, vol. 19, no. 9, Sep. 2004, pp. 845–855.

[34] M. Leuenberger and M. Kanevski, "Extreme Learning Machines for spatial environmental data", *Computers & Geosciences*, vol. 85, Dec. 2015, pp. 64–73.

[35] V. Kontis, J. E. Bennett, C. D. Mathers, G. Li, K. Foreman, and M. Ezzati, "Future life expectancy in 35 industrialised countries: Projections with a Bayesian model ensemble", *The Lancet*, vol. 389, no. 10076, Apr. 2017, pp. 1323–1335.

[36] L. Breiman, J. H. Friedman, and R. A. Olshen, *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[37] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data", *Applied Statistics*, vol. 29, no. 2, 1980, pp. 119–127.

[38] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*. Singapore ; Hackensack, NJ: World Scientific, 2008.

[39] The World Bank, "World development indicator", 2019. https://databank.worldbank.org/source/world-development-indicators