Malaysian Journal of Computing, 7 (2): 1150-1161, 2022
Copyright © UiTM Press
eISSN: 2600-8238 online

PREDICTION OF LIFE EXPECTANCY FOR ASIAN POPULATION USING MACHINE LEARNING ALGORITHMS

Nurul Shahira Pisal[1], Shuzlina Abdul-Rahman[2*], Mastura Hanafiah[3] and Saidatul Izyanie Kamarudin[4]

[1,2,4]*Faculty of Computer and Mathematical Sciences,*
Universiti Teknologi MARA (UiTM), Shah Alam, MALAYSIA
[3]*Accenture Solutions Sdn. Bhd., Kuala Lumpur, MALAYSIA*
[1]shahirapisal99@gmail.com, [2*]shuzlina@uitm.edu.my, [3]mastura.hanafiah@accenture.com,
[4]saidatulizyanie@uitm.edu.my

ABSTRACT

Predicting life expectancy has become more important nowadays as life has become more vulnerable due to many factors, including social, economic, environmental, education, lifestyle, and health condition. A lot of studies on life expectancy have been carried out. However, studies focusing on the Asian population are limited. This study presents machine learning algorithms for life expectancy based on the Asian population dataset. Comparisons are made between tree classifier models, namely, J48, Random Tree, and Random Forest. Cross validations with 10 and 20 folds are used. Results show that the highest accuracy is obtained with Random Forest with 84% accuracy with 10-fold cross-validation. This study further identifies the most significant factors that influence life expectancy prediction, which includes socioeconomic factors and educational status, health conditions and infectious disease.

Keywords: *Life Expectancy, Data Classification, Data Mining, Asian Population*

Received for review: 01-06-2022 Accepted: 26-09-2022; Published: 01-10-2022
DOI: 10.24191/mjoc.v7i2.18218

1. Introduction

The lifespan of a person is a mystery that nobody has ever known. People do not know how long they are going to live or how they will live their lives. Life expectancy seems to be the critical metric for assessing the health of a population. Life expectancy refers to the number of years a person can expect to live. Estimates suggest that life expectancy was about 30 years around the world during the pre-modern era, but it has increased more than doubled since 1900 (Roser *et al.*, 2019). Between 2000 and 2019, life expectancy increased by more than six years globally between the year 2019 and 2000 (World Health Organization, n.d.-b).

Life expectancy can be influenced by several factors such as socioeconomic status, healthcare availability and quality, lifestyle, nutrition, social factors, genetic factors, and environmental factors. In a study done on the Asian population of countries including Singapore, Malaysia and Thailand, the results showed that the higher levels of socioeconomic advantage and more excellent healthcare resources of the people were more likely to enhance life expectancy (Chan & Kamala Devi, 2015).

With the growth of big data and analytics, many available data can be collected and analysed for researchers to predict life expectancy. A process of exploring large datasets to identify unknown patterns or relationships, or even anomalies that are present in the data, is called data mining. Data mining and machine learning might be used interchangeably, but there are differences between the two. Data mining investigates the patterns in the data. In contrast, machine learning goes beyond what has happened in the past to predict the outcomes. The outcome is based on what the machine has learned from the pre-existing data. Recent developments in data mining have been applied in many classifications works, such as in disease prediction (Nalluri *et al.*, 2020; Shuja *et al.*, 2020; Verma *et al.*, 2020), social studies (Song & Song, 2021; Vanlalawmpuia & Lalhmingliana, 2020), agriculture (Kaur *et al.*, 2021), education (Ahmad Tarmizi *et al.*, 2019; Basheer *et al.*, 2019; Mohammad Suhaimi *et al.*, 2019) and many more. Several data mining classifications have been used in previous studies on life expectancy, including Decision Tree, Naïve Bayes, k-Nearest Neighbor and Support Vector Machine (Mohammad Suhaimi *et al.*, 2019; Sharma *et al.*, 2016).

This study examines the prediction of life expectancy based on the dataset on life expectancy for Asian countries. The objective of this study is two-fold: the first is to investigate the factors influencing life expectancy, and the second is to compare the performance of the classification of life expectancy of the Asian population using machine learning algorithms (MLAs) approaches based on the identified factors. Three MLAs, namely J48, Random Tree, and Random Forest, are used to develop the predictive models. The performances of these MLAs are evaluated with several accuracy metrics. The following sections discuss the past works related to life expectancy factors and MLAs, the methodology, the results, and findings, and finally, the conclusion and possible future works.

2. Related Works

2.1 Life Expectancy Influencing Factors

There are many studies on factors influencing human life expectancy. Some aspects found to have associations with life expectancy were genomics, environment, socioeconomics, and patient behaviour (Kang & Adibi, 2018). Monsef and Mehrjardi (2015) studied the determinants of life expectancy according to social, economic and environmental factors. The study also revealed that unemployment and inflation were the key economic factors that had a negative impact on life expectancy, but gross capital formation and gross national income had a favourable effect. Financial and education were also contributing factors toward better life expectancy as it was believed that a better financial state provided a better quality of life (Kaplan *et al.*, 1996; Walczak *et al.*, 2021). In addition, better education helps the person to live a better life with a fully equipped home, better education and the ability to have high-quality medical care (Luy *et al.*, 2019).

A person's life expectancy could also be influenced by demographic characteristics, lifestyle, and health and disease indicators (Walter *et al.*, 2012; Li *et al.*, 2020). In most countries, studies on genomic elements associated with organisms, including the structure of the human genetic code, revealed that women lived longer than men (Le *et al.*, 2015; Bin-Jumah *et al.*, 2022). There was also a link between parents' life expectancy and their children's life expectancy, which is slightly higher than the link between their spouses' life expectancy (Fire & Elovici, 2015). Lifestyle was another contributor to human life expectancy. Early exposure to unhealthy daily habits such as smoking, drinking alcohol, and drug abuse could make the person's mortality rate shorter than those who receive it later; this could be because early exposure could increase the chance of getting killing diseases such as cancer, stroke, diabetes, and dementia (Rizzuto & Fratiglioni, 2014). Several studies also indicated that immunization had contributed to life expectancy and quality of life improvement (Andre *et al.*, 2008; Destefano *et al.*, 2019; Gagneur *et al.*, 2019).

2.2 Life Expectancy Prediction with Machine Learning

Studies on life expectancy depend mainly on the labelled dataset in which supervised approaches are more relevant. Further studies on tree-based classification models have been used in several life expectancy studies (Karacan *et al.*, 2020; Meshram, 2020; Vydehi *et al.*, 2020). Karacan *et al.* (2020), a survey of life expectancy across countries based on the World Health Organization (WHO) dataset using a decision tree revealed that 9 out of 25 attributes significantly influence life expectancy. A recent study comparing life expectancy in rich and developing nations using three regression models, Linear Regression, Decision Tree Regressor, and Random Forest Regressor, found that the Random Forest Regressor was the best model with R2=0.99(training) and 0.95(testing), along with 4.43 and 1.58 as the mean squared error and mean absolute error.

Artificial Neural Networks (ANN) have also been a common choice for life expectancy prediction studies. Beeksma *et al.* (2019) found that by utilizing an extended short-term recurrent neural network and unstructured clinical free-text, the model achieved a level of accuracy comparable to human accuracy (20% accuracy), and the keyword model boosted prediction accuracy to 29%. The researchers believed that this model tends to make pessimistic forecasts, whereas doctors tend to make optimistic predictions. In Wang *et al.* (2017), several models were also used to predict chronological age based on patients' medical records, including Random Forest, Elastic Nets and deep ANN. It was found that ANN yielded the best performance with an increase in sample size (377,686 medical records from patients aged 18-85 years old) with an accuracy of 90%.

Agarwal *et al.* (2019), in their study on predicting the life expectancy of the population over various continents of the world, and classifying the likelihood of occurrence of long-standing diseases, applied simple linear regression and multiple linear regression used for the former and decision tree and random forest algorithms for the latter. The study showed that Multiple Linear Regression produced the most accurate results for predicting average population life expectancy given current continent features. In contrast, Random Forest had better results for indicating the likelihood of the five diseases (HIV/AIDS, measles, diphtheria, hepatitis B, and polio) occurring across continents.

2.3 Machine Learning Algorithms

Among the many Machine Learning Algorithms (MLAs), decision tree techniques such as J48, Random Tree, and Random Forest are commonly used. These techniques are available in WEKA (Eibe *et al.*, 2016) and are widely used for classification studies. J48 is a classification algorithm that generates a decision tree. It is one of the best machine learning algorithms to examine the data categorically and continuously (Saravana & Gayathri, 2018). Random Tree is another supervised classifier that is available in WEKA. It is an ensemble learning algorithm that generates individual learners or a set of tree predictors that is called a forest. Employing a bagging approach produces a collection of data to create a decision tree. However, using Random trees or bagged decision trees has a disadvantage whereby the decision trees constructed using the greedy approach select the best split point at each step in the tree-building process. The process causes the trees to appear very similar, and consequently, the variance of the prediction from all the bags is reduced; therefore, the robustness of the prediction could be deteriorated (Brownlee, 2019). Random Forest is an extension of a Random Tree that can be used for classification or regression. Instead of employing greedy splitting like in a bagged decision tree, the split points can only be chosen randomly from a subset of input attributes during tree building. This single change reduces the similarity between the bagged trees and the predicted outcome (Brownlee, 2019).

3. Methodology

This study has gone through several phases, as shown in Figure 1, which are: i) Data collection, ii) Data preparation and pre-processing, iii) Model development and evaluation. The descriptions of these phases are described in the following sections.

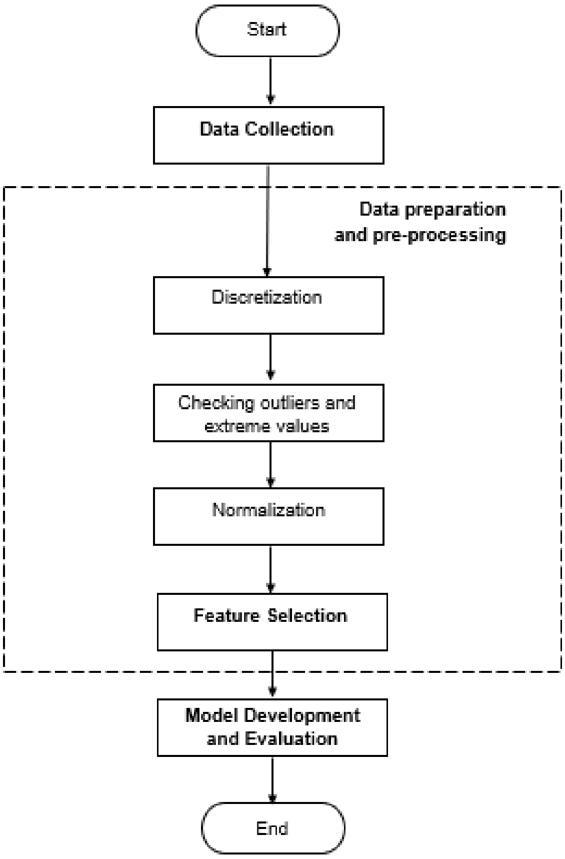
Start
Data Collection
Data preparation and pre-processing
Discretization
Checking outliers and extreme values
Normalization
Feature Selection
Model Development and Evaluation
End

Figure 1. Flowchart of the research methodology

3.1 Data Collection

The data used in the study was the dataset on life expectancy that is available on the Kaggle website (https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who). The data consists of life expectancy data from 193 countries, with 22 attributes and 2938 instances or records, from 2000 to 2015. Table 1 summarises the Life expectancy attributes and their descriptions.

Table 1. Profile of Life Expectancy Attributes

Attribute	Description
Country	Country
Year	Year
Status	Develop or Developing Country
Adult mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant deaths	Number of Infant Deaths per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita (%)
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Measles - number of reported cases per 1000 population
BMI	Average Body Mass Index of the entire population
Under-five deaths	Number of under-five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	The population of the country
Thinness 10-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9(%)
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Number of years of Schooling(years)
Life expectancy (class/output)	Life expectancy in age

3.2 Data preparation dan pre-processing

The dataset was downsized to the Asia continent, which consists of five regions, namely Eastern Asia, Southern Asia, Central Asia, Western Asia, and Southeast Asia, and 48 countries. In the dataset, however, only data from 46 countries were available. The final instances used in the study were 736. Data preparation and pre-processing, classification and evaluation were handled using WEKA (Waikato Environment for Knowledge Analysis) software and can be downloaded from the WEKA website (https://waikato.github.io/weka-wiki/downloading_weka/). It is a collection of advanced machine learning algorithms and pre-processing data tools written in Java. It can perform data mining tasks like data categorization, clustering, regression, attribute selection, and many more.

Before proceeding with data pre-processing, the data needed to be cleaned; some missing values were detected in the dataset. Thus, the missing values were treated using the 'ReplaceMissingValue' function. The missing values were replaced with the mean value. However, some missing values remained untreated for the attribute 'schooling' for the country Republic of Korea. This missing value was manually filled with the value of '12'; which is the total number of years that Korean students usually spend during schooling (elementary, middle and high school) (National Center on Education, n.d.; Education system, 2020). Korean education ministry adapted 6-3-3-4 in their education system, whereby 6 years are spent for elementary study, while 3 years are allocated for middle school, 3 years for highs school and 4 years in university or college.

The next step is, performing binning or discretization on the attribute 'Life Expectancy', the expected class output. Binning is done to transform a continuous or numerical variable into a categorical feature. The values for the Life Expectancy attribute were the numbers representing age. Thus, it had to be categorized into a smaller number of age intervals. In WEKA, binning was done using the discretization function, where; the bins were set to be '5'. Binning is also performed on the attribute 'BMI' to categorize the values into five intervals.

Afterwards, the next step was checking the presence of outliers and extreme values in the dataset. The function used was 'InterQuartileRange'. The outliers and extreme values were then removed using a function in WEKA called 'RemoveWithValue', where the process is set to remove attributes with outliers and extreme values. The attributes were mostly numerical and had varying scales; thus, data normalization was performed. Data normalization is rescaling one or more attributes to the range of 0 to 1. The normalization was carried out with WEKA by selecting the Normalize filter and applying it to the dataset.

The purpose of feature selection was to find the features that substantially impact the model's performance. The selection procedure entailed searching the data for all possible combinations of attributes to determine which subset of attributes would be most beneficial for the prediction. With the WEKA function' Select Attribute', the selection was made using the Wrapper evaluator and 'BestFirst' method. The classifier used in selecting the attribute was the Random Forest, as the classification process would use this classifier. The 'BestFirst' method was used in the backward direction since the selection was to find the most influencing attributes towards the output prediction.

3.3 Model Development and Evaluation

The models J48, Random Tree, and Random Forest were developed and evaluated with the cross-validation technique, a resampling approach for evaluating machine learning models on a small sample of data. It involves the parameter called *k,* which represents the number of groups into which the dataset is to be split. This approach divides the dataset into approximately equal sizes of *k* groups of folds, and the first fold is used as a validation set while the remaining *k-1* folds are used as training (James *et al.*, 2013). In this study, 10 and 20 folds were chosen. For the tree-based classifier models, i.e., Random Forest, Random Tree and J48, the performance analysis is conducted against their accuracy, Root Means Squared Error (RMSE), Relative Absolute Error (RAE) and Receiver Operating Characteristic (ROC) Area. RMSE is the measure of the standard deviation of residuals, which is calculated with the square root of the mean squared area. RAE is another metric normally used to measure the performance of a predictive model. It is the measure in percentage of how much the result deviates from the actual value. ROC Area summarises the overall diagnostic accuracy of the test, which takes the value between 0 (perfectly inaccurate) and 1 (perfectly accurate). The ROC area measures discrimination, that is, the ability of the test to perform the classification correctly. The value of ROC area between 0.7-0.8 are acceptable, 0.8-0.9 are excellent, and above 0.9 is outstanding (Hosmer & Lemeshow, 2000).

4. Results and Discussion

4.1 Tree Classifier Model Performance with Full Features

Before the feature selection was conducted, the data were first analyzed with all the available features by applying the tree classifier models: Random Forest, Random Tree and J48. These algorithms were run with 10-fold and 20-fold cross-validation techniques. Table 2 summarizes the performance result for these classifiers.

Table 2. Performance Results of Tree Classifiers (with full features)

	Random Forest		**Random Tree**		**J48**	
Metrics	**10-fold**	**20-fold**	**10-fold**	**20-fold**	**10-fold**	**20-fold**
Accuracy (%)	**88.24**	**87.62**	81.42	82.04	78.64	75. 85
RMSE	**0.193**	**0.193**	0.27	0.264	0.271	0.288
RAE (%)	31.34	31.10	26.53	**25.49**	36.98	41.78
ROC Area	0.973	**0.974**	0.868	0.877	0.889	0.864

Based on the table above, it can be concluded Random Forest is the best technique among the three techniques of tree classifiers. The method achieved an accuracy of 88.24% when tested with 10-fold and 87.62% with 20-fold cross-validation. RMSE is the lowest, with 0.193 for both settings. However, the lowest RAE is found for Random Tree of 20 folds (25.49 %). Although the value of Random Forest's RAE is higher (31%) than that of the Random Tree, Random Forest's ROC Area is higher (0.97). Based on these results, it can be concluded that Random Forest with 10-fold is the best classifier model when experimenting with all attributes.

4.2 Tree Classifier Model Performance with Selected Features

Feature selection was then performed to identify the attributes that influence life expectancy. The feature selection done with WEKA revealed that 11 attributes have an influence on the outcome. This included 'year', 'adult mortality', 'infant deaths', 'BMI', 'under-five deaths', total expenditure, 'diphtheria', 'HIV/AIDS', 'GDP', 'income composition of resources' and 'schooling' attributes. The data was re-modelled again with the same tree classifiers based on these identified attributes. Table 3 shows the experiment's results with the tree classifiers based on the 11 attributes, and Figure 2 shows the accuracy results. It is shown that between 10-fold and 20-fold cross-validation, there is not much difference in the accuracy values, especially on Random Forest and Random Tee.

Table 3. Performance Results of Tree Classifiers (with selected features)

	Random Forest		**Random Tree**		**J48**	
Metrics	**10-fold**	**20-fold**	**10-fold**	**Metrics**	**10-fold**	**20-fold**
Accuracy (%)	**84.83**	84.21	78.64	78.33	81.11	80.50
RMSE	0.212	**0.208**	0.292	0.295	0.2656	0.2663
RAE (%)	39.95	38.31	**30.65**	31.29	34.06	34.44
ROC Area	0.962	**0.967**	0.847	0.838	0.887	0.877

Random Forest and J48 have higher accuracy with 10-fold while Random Tree has higher accuracy with 20-fold. Among the three classifiers, the highest accuracy is obtained with Random Forest with 10-fold cross-validation, which is 84.83%. Figure 2 shows the RMSE, RAE and ROC Area chart combination. The figure shows that lower RMSEs are achieved with Random Forest, and the lowest is with 20-fold cross-validation, which is 0.2081, although the difference with 10-fold is relatively minimal.

RAE value is the lowest for Random Tree (30.65%). Although the value of Random Forest's RAE is slightly higher than the RAE of the Random Tree, Random Forest's ROC Area is higher (0.967). The results show that Random Forest is still the best technique among the three tree classifiers. This result is in sync with the findings from Karacan *et al.* (2020), in which Random Forest was found as the best model for life expectancy prediction.

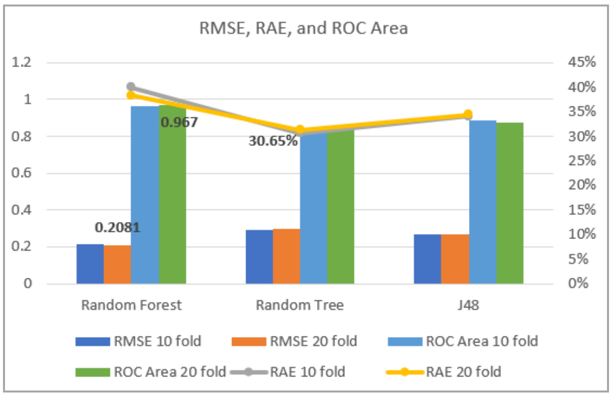

Figure 2. RMSE, RAE and ROC Area for tree classifiers

4.3 Factors Influencing Life Expectancy

During the attribute selection exercise, 11 attributes were found to have a strong association with life expectancy. The attributes' total expenditure', 'GDP' and 'income composition of resources were found to have an influence on life expectancy. Thus this finding is in sync with the works by Kaplan *et al.* (1996), Monsef and Mehrjardi (2015) and Walczak *et al.* (2021), whereby socioeconomic-related factors were found to have a positive impact on life expectancy. The attribute 'schooling' is also consistent with the findings by Luy *et al.* (2019), whereby the study demonstrated a strong association between education and overall population health. As for mortality or death-related attributes such as 'adult mortality', 'infant deaths', and 'under-five deaths', whether or not these factors have a strong influence on life expectancy, it was debatable (Murray, 1988). However, recent studies indicated that infant mortality levels lead to higher life expectancy at birth, suggesting the longevity of people in five EU countries (Miladinov, 2020). For the health-related attributes yielded in the attribute selection exercise, which were 'BMI' and 'HIV/AIDS', this finding is supported by several studies such as Walter *et al.* (2012) whereby this study indicated that life expectancy could be influenced by demographic characteristics, lifestyle and indicators of health and diseases. Although recent findings suggested that the life expectancy of adults with HIV infection may be comparable to that of people without HIV infection, more emphasis must be paid to preventing comorbidities in HIV patients (Marcus *et al.*, 2020). Another attribute found that has an association with life expectancy was 'diphtheria'. This finding is also supported by the research done by Agarwal *et al.* (2019), as discussed in the previous section. Diphtheria is a disease caused by particular bacteria and has caused many deaths, especially in developing countries (World Health Organization, n.d.-a). Studies have shown that the diphtheria vaccine has reduced the number of reported cases by more than 90%, consequently improving the population's life expectancy (Clarke, 2017; World Health Organization, n.d.-a; Sharma, *et al.*, 2019).

5. Conclusion and Future Works

This research has presented the classification results for life expectancy for the Asian population. Data mining approaches using several classifiers and regression models have been applied. The findings show that Random Forest performs better in terms of accuracy when compared with other tree classifier models, J48 and Random Tree. This study also found that several attributes strongly correlate with life expectancy, and these attributes are mainly related to economic and educational status, health conditions and infectious disease. Further studies could further explore more deeply on the correlation between the attributes and life expectancy. Additionally, a study on a recent dataset could also be conducted especially with the consequences of the current COVID-19 pandemic. The country-based analysis could also be established as different countries have different economic situations, education, and even healthcare facility status. Different algorithms such as artificial neural networks, support vector machines or logistic regression could also be applied to compare the results.

6. Acknowledgement

The authors would like to thank the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, for the research support.

References

Agarwal, P., Shetty, N., Jhajharia, K., Aggarwal, G., & Sharma, N. V. (2019). Machine learning for prognosis of life expectancy and diseases. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 1765–1771.

Ahmad Tarmizi, S. S., Mutalib, S., Abdul Hamid, N. H., Abdul-Rahman, S., & Md Ab Malik, A. (2019). A Case Study on Student Attrition Prediction in Higher Education Using Data Mining Techniques. In *International Conference on Soft Computing in Data Science* (pp. 181–192). Springer. https://doi.org/10.1007/978-981-15-0399-3_15

Andre, F. E., Booy, R., Bock, H. L., Clemens, J., Datta, S. K., John, T. J., Lee, B. W., Lolekha, S., Peltola, H., Ruff, T. A., Santosham, M., & Schmitt, H. J. (2008). Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bulletin of the World Health Organization*, *86*(2), 140–146.

Basheer, M. Y. I., Mutalib, S., Hamid, N. H. A., Abdul-Rahman, S., & Malik, A. M. A. (2019). Predictive analytics of university student intake using supervised methods. *IAES International Journal of Artificial Intelligence*, *8*(4), 367–374.

Beeksma, M., Verberne, S., van den Bosch, A., Das, E., Hendrickx, I., & Groenewoud, S. (2019). Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Medical Informatics and Decision Making*, *19*(1), 36.

Bin-Jumah, M. N., Nadeem, `M., Gilani, S., Al-Abbasi, F., Ullah, I., Alzarea, S., Kazmi, I. (2022). Genes and Longevity of Lifespan. *International Journal of Molecular Sciences, 23*(3), 1-27.

Brownlee, J. (2019). How to Use Ensemble Machine Learning Algorithms in WEKA, WEKA Machine Learning. A post at MachineLearningMastery available at https://machinelearningm*astery.com/use-ensemble-machine-learning-algorithms-WEKA/*

Chan, M. F., & Kamala Devi, M. (2015). Factors Affecting Life Expectancy: evidence from 1980-2009 data in Singapore, Malaysia, and Thailand. *Asia Pacific Journal of Public Health*, *27*(2), 136–146. https://doi.org/10.1177/1010539512454163

Clarke, K. (2017). Review of the epidemiology of diphtheria 2000-2016. *US Centeres for Disease Control and Prevention*. https://doi.org/10.1371/journal.pone.0044878

Destefano, F., Bodenstab, H. M., & Offit, P. A. (2019). Principal Controversies in Vaccine Safety in the United States. *Clinical Infectious Diseases*, *69*(4), 726–731.

Education system. (2020). Retrieved from Ministry of Education: http://english.moe.go.kr/sub/infoRenewal.do?m=0301&page=0301&s=english

Eibe, F., Mark A. H., & Ian H. W. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Fire, M., & Elovici, Y. (2015). Data mining of online genealogy datasets for revealing lifespan patterns in human population. *ACM Transactions on Intelligent Systems and Technology*, *6*(2), 1–22. https://doi.org/10.1145/2700464

Gagneur, A., Quach, C., Boucher, F. D., Tapiero, B., De Wals, P., Farrands, A., Lemaitre, T., Boulianne, N., Sauvageau, C., Ouakki, M., Gosselin, V., Gagnon, D., Petit, G., Jacques, M. C., & Dubé, È. (2019). Promoting vaccination in the province of Québec: The PromoVaQ randomized controlled trial protocol. *BMC Public Health*, *19*(1), 1–9.

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression 2nd edn John Wiley & Sons. *Inc.: New York, NY, USA*, 160–164.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kang, J., & Adibi, S. (2018). Systematic Predictive Analysis of Personalized Life Expectancy Using Smart Devices. *Technologies*, *6*(3), 74.

Kaplan, G. A., Pamuk, E. R., Lynch, J. W., Cohen, R. D., & Balfour, J. L. (1996). Inequality in income and mortality in the United States: Analysis of mortality and potential pathways. *British Medical Journal*, *312*(7037), 999–1003.

Karacan, I., Sennaroglu, B., & Vayvay, O. (2020). Analysis of life expectancy across countries using a decision tree. *Eastern Mediterranean Health Journal*, *26*(2), 143–151.

Kaur, P., Chahal, J. K., & Sharma, T. (2021). A DATA MINING APPROACH FOR CROP YIELD PREDICTION IN AGRICULTURE SECTOR. *Advances in Mathematics: Scientific Journal*, *10*(3), 1425–1430.

Le, Y., Ren, J., Shen, J., Li, T., & Zhang, C. F. (2015). The changing gender differences in life expectancy. *PLoS ONE*, *10*(4), 1–11.

Li, Y., Schoufour, J., Wang, D., Dhana, K., Pan, A., Liu, X., Hu, F. (2020, January 8). Healthy lifestyle and life expectancy free of cancer,. *BMJ, 368*(8228), 1-9.

Luy, M., Zannella, M., Wegner-Siegmundt, C., Minagawa, Y., Lutz, W., & Caselli, G. (2019). The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA. *Genus*, *75*(1).

Marcus, J. L., Leyden, W. A., Alexeeff, S. E., Anderson, A. N., Hechter, R. C., Hu, H., Lam,

J. O., Towner, W. J., Yuan, Q., Horberg, M. A., & Silverberg, M. J. (2020). Comparison of Overall and Comorbidity-Free Life Expectancy Between Insured Adults With and Without HIV Infection, 2000-2016. *JAMA Network Open*, *3*(6), e207954. https://doi.org/10.1001/jamanetworkopen.2020.7954

Meshram, S. S. (2020). Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning. *2020 IEEE Bombay Section Signature Conference (IBSSC)*, 6–10.

Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: evidence from the EU accession candidate countries. *Genus*, *76*(1).

Mohammad Suhaimi, N., Abdul-Rahman, S., Mutalib, S., Abdul Hamid, N. H., & Md Ab Malik, A. (2019). Predictive Model of Graduate-On-Time Using Machine Learning Algorithms. In *Communications in Computer and Information Science* (Vol. 1100, Issue September). Springer Singapore.

Monsef, A., & Mehrjardi, A. S. (2015). Determinants of Life Expectancy: A Panel Data Approach. *Asian Economic and Financial Review*, *5*(11), 1251–1257.

Murray, C. J. L. (1988). The Infant Mortality Rate, Life Expectancy at Birth, and a Linear Index of Mortality as Measures of General Health Status. *International Journal of Epidemiology*, *17*(1), 122–128. https://doi.org/10.1093/ije/17.1.122

Nalluri, S., Vijaya Saraswathi, R., Ramasubbareddy, S., Govinda, K., & Swetha, E. (2020). Chronic Heart Disease Prediction Using Data Mining Techniques. *Advances in Intelligent Systems and Computing*, *1079*(June), 903–912.

National Center on Education. (n.d.). Top Performing Countries. *Availabe at https://ncee.org/country/korea/*

Navidi, W., & Monk, B. (2015). *Elementary Statistics* (2nd ed). MCGraw-Hill Education.

Rizzuto, D., & Fratiglioni, L. (2014). Lifestyle factors related to mortality and survival: A mini-review. *Gerontology*, *60*(4), 327–335. https://doi.org/10.1159/000356771

Roser, M., Ortiz-Ospina, E., & Ritchie, H. (2019). Life Expectancy. *A post at OurWorldinData availabe at https://ourworldindata.org/life-expectancy*

Saravana, N., & Gayathri, D. V. (2018). Performance and classification evaluation of J48 algorithm and Kendall's based J48 algorithm (KNJ48). *Int. J. Comput. Trends Technol.(IJCTT)--Volume*, *59*, 73–80.

Sharma, N. C., Efstratiou, A., Mokrousov, I., Mutreja, A., Das, B., & Ramamurthy, T. (2019). Diphtheria. *Primer*, 1-18. doi:https://doi.org/10.1038/s41572-019-0131-y

Sharma, T., Sharma, A., & Mansotra, V. (2016). Performance analysis of data mining classification techniques on public health care data. *International Journal of Innovative Research in Computer and Communication Engineering*, *4*(6), 11381–11386.

Shuja, M., Mittal, S., & Zaman, M. (2020). *Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. January*, 195–211.

Song, T.-M., & Song, J. (2021). Prediction of risk factors of cyberbullying-related words in

Korea: Application of data mining using social big data. *Telematics and Informatics*, *58*, 101524. https://doi.org/10.1016/j.tele.2020.101524

Vanlalawmpuia, R., & Lalhmingliana, M. (2020). Prediction of Depression in Social Network Sites Using Data Mining. *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 489–495.

Verma, A. K., Pal, S., & Kumar, S. (2020). Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study. *Applied Biochemistry and Biotechnology*, *190*(2), 341–359.

Vydehi, K., Manchikanti, K., Satya Kumari, T., & Ahmad Shah, S. K. (2020). Machine learning techniques for life expectancy prediction. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4), 4503–4507.

Walczak, D., Wantoch-Rekowski, J., & Marczak, R. (2021). Impact of income on life expectancy: A challenge for the pension policy. *Risks*, *9*(4).

Walter, S., MacKenbach, J., Vokó, Z., Lhachimi, S., Ikram, M. A., Uitterlinden, A. G., Newman, A. B., Murabito, J. M., Garcia, M. E., Gudnason, V., Tanaka, T., Tranah, G. J., Wallaschofski, H., Kocher, T., Launer, L. J., Franceschini, N., Schipper, M., Hofman, A., & Tiemeier, H. (2012). Genetic, physiological, and lifestyle predictors of mortality in the general population. *American Journal of Public Health*, *102*(4), 3–10.

Wang, Z., Li, L., Glicksberg, B. S., Israel, A., Dudley, J. T., & Ma'ayan, A. (2017). Predicting age by mining electronic medical records with deep learning characterizes differences between chronological and physiological age. *Journal of Biomedical Informatics*, *76*, 59–68.

World Health Organization. (n.d.-a). Diphtheria reported cases. *Availabe at http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tsincidencediphtheria.html.*

World Health Organization. (n.d.-b). GHE: Life expectancy and healthy life expectancy. *Availabe at https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-life-expectancy-and-healthy-life-expectancy*