COMPARISON OF CLASSIFICATION METHODS FOR ANALYZING ELSA AND CLHLS AGEING DATASETS

Yuan Fangzheng, Assoc.Prof. Dr. Suhaila Zainudin

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor Malaysia.

p123027@siswa.ukm.edu.my, suhaila.zainudin@ukm.edu.my

ABSTRACT

With the advent of the new century, the world economy has become increasingly integrated, the accelerated ageing of the population worldwide, more healthcare resources may be needed to meet the needs of society better, and it is becoming increasingly important to plan current and future healthcare resources based on the health status of older adults. Conventional methods perform poorly in the face of increasingly large amounts of data and cannot accurately capture some nonlinear relationships. This study focus on the comparison of various classification methods in analyzing ageing datasets from English Longitudinal Study of Ageing (ELSA) and Chinese Longitudinal Healthy Longevity Survey (CLHLS). Potential health influencing factors of older adults were mined through machine learning methods, particularly physiological factors and physical factors. The performance of algorithms such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and XGBoost (XGB) on two datasets (CLHLS and ELSA) were compared to find the most effective model for predicting the health status of the ageing population. The project's results proved machine learning models' effectiveness in predicting older adults' health status. The logistic regression model for CLHLS and ELSA was ultimately found to predict the best healthy ageing with the highest accuracy (0.776 and 0.774) and the lowest mean square error (0.253 and 0.224). This demonstrates its ability to capture the underlying patterns and relationships in the data and make accurate predictions. This study supports the application of machine learning methods to analyze the health status for ageing datasets.

Keyword: Data Mining, Maching Learning, Healthy Ageing, Data Preprocessing

I. INTRODUCTION

As the global population ages, research on health and ageing becomes increasingly essential, especially in planning current and future healthcare resources (OECD & Organization 2020). The proportion of people aged 65 and above is increasing, particularly in Europe and Asia (Keating, 2022).

A. Research background

Healthy ageing refers to maintaining physical, psychological, and emotional health as individuals age. It involves taking measures to prevent or delay the onset of age-related diseases, preserving functional

capacity, and improving the overall quality of life. Healthy ageing is achieved through a combination of factors, including regular physical activity, a balanced diet, avoiding harmful habits like smoking and excessive alcohol consumption, maintaining a healthy weight, adequate sleep, stress management, and seeking regular medical care. Keeping mental and social engagement and a positive outlook on life can also promote healthy ageing (Rudnicka et al. 2020).

However, care resources are limited, so prioritizing the effective utilization of these resources becomes necessary for older adults. To address this issue, predicting the health status of older adults is needed to provide appropriate services based on their care needs and risks. Given that older adults' health is influenced by various factors, including genetics, lifestyle, environmental factors, and chronic diseases, predicting the health status of older adults becomes a complex task (Feng et al. 2020).

B. Problem statement

Most of the current prediction problems of healthy ageing are based on traditional prediction methods, which may not be able to achieve the efficiency and effectiveness of machine learning methods in the face of massive amounts of data because conventional methods need to manually select the features, which causes great inconvenience and also challenging to find the nonlinear relationship between the elements and the target variable (health status), which makes it challenging to achieve the most accurate prediction results (Ngiam & Khor 2019).

Despite the availability of large amounts of data, the data collected from different sources may need more consistency and standardization, making it easier to compare and analyze the data. Furthermore, the lack of data on specific subgroups of older adults population, such as older adults living in rural areas or from ethnic communities, can lead to biased predictions, hindering the development of effective policies and interventions (Müller et al. 2021).

Feature selection is an essential part of machine learning. It aims to select the most influential features for predicting outcomes from extensive data. However, the health outcomes of older adults may be influenced by various factors, including lifestyle, genetic, and environmental factors, which may interact with each other in complex ways (Lim et. al 2020). Predicting the health status of older adults may involve many potential influencing factors, making it a highly challenging task to select the most relevant features (Chekroud et al. 2021).

Choosing an appropriate machine learning model is crucial for predicting the health status of older adults. Different models have different characteristics and application scopes, so it is necessary to consider the models' accuracy, efficiency, and interpretability (Qiu et al. 2022).

Machine learning models have solid predictive capabilities but need more interpretability, meaning explaining why the model makes confident predictions is difficult (Elshawi et al. 2019). For predicting the health status of older adults, it is essential to understand the reasons and basis behind the model's predictions to take appropriate intervention measures.

To address these problems, this study aims to study the performances of various classification methods on ageing datasets (ELSA and CLHLS). By doing so, it aims to improve the understanding of the health status of older adults and facilitate the development of personalized health management strategies and interventions in the long run.

C. Research questions and objectives

The following are the objectives of this study:

1. To select features that contributed towards healthy ageing in ELSA and CLHLS.
2. To compare classification models based on diferent approaches to classifify the ageing dataset.

D. Research significance

With the acceleration of global population ageing, ageing has become an important issue that is faced globally. In the context of an ageing society, more medical resources are needed to meet the health needs of older adults, and it is becoming increasingly important to plan current and future medical resources with the health status of older adults. Therefore, the feasibility and development of machine learning models in predicting healthy ageing are of great value to providing information for health prevention and treatment planning for older adults, improving their health and quality of life, and planning current and future healthcare resources.

II. LITERATURE REVIEW

A. Overview of Healthy Ageing

According to the literature review, the physiological factors affecting healthy ageing mainly include age, gender, physical function, chronic Illness, etc. With the increase in age, the decline of physical activity ability and the change in the social role of older adults restrict their social participation, which is not conducive to healthy ageing (Rodriguez-Laso et al. 2018).

In addition, women have long been responsible for physical work such as household cleaning and cleaning, and their minds are delicate and sensitive, so they are more likely to have physical and psychological problems. At the same time, elderly women lack independence in the society, and the healthy ageing level of elderly women is lower than that of elderly men (Naah, et al. 2020). Physical health is the fundamental guarantee for older adults to realize healthy ageing. The good physical function ensures good self-care ability and social interaction in older adults (Chan et. al 2020). older adults plagued by chronic diseases for years cannot participate in social activities, and their mental health will also be affected (Chew et al. 2021).

Xu et al. (2020) conducted a longitudinal analysis on the relationship between abdominal circumference and healthy ageing based on an 8-year follow-up survey data in the United States, and the results showed that the healthy ageing score of the abdominal obese elderly was significantly lower than that of the standard group. However, abdominal circumference had no apparent effect on the annual change rate of healthy ageing.

Su et al. (2021) investigated the effectiveness of logistic regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM) in predicting the health status of older adults on the CLHLS dataset. The LR method showed higher accuracy than the DT and RF methods, while the SVM method had the fastest convergence rate. The accuracy of the DT method (75.9%) was 10% higher than that of the LR (65.9%) but 5% lower than that of the SVM （56.4%）. These results show that SVM methods are efficient and accurate in predicting the health status of older adults.

Tarekegn et al. (2020) examined the effectiveness of LR, DT, RF, and SVM in predicting frailty in older adults on the Piedmontese Longitudinal Study dataset. The LR method (69%) showed higher accuracy than the DT and RF methods, while the SVM method (68%) had the fastest convergence rate. The RF method (68%) was 1% more accurate than LR (69%) but 2% less accurate than DT (67%). These results suggest that the LR method is efficient and accurate in predicting frailty in older adults.

Cuaya-Simbro et al. (2020) stated that DT and SVM were used for prediction on a dataset from the National Rehabilitation Institute (INR) and community centres in Mexico City. The SVM method showed higher accuracy compared to the DT method. The SVM method (65.45%) was 0.9% less accurate than the DT (66.36%). These results show that the DT method is efficient and accurate in predicting the health status of older adults.

Tongkaw et al. (2020) used decision tree modelling to focus on predicting medical problems in the elderly. The method obtained an accuracy of 0.67 on an unspecified dataset. These results show that the decision tree method is efficient and accurate in predicting the medical problems of the elderly.

Fazakis et al. (2021) studied the effectiveness of DT, RF and SVM in predicting long-term cholesterol risk in older adults was investigated on the ELSA dataset. The RF method (61.36%) showed similar accuracy compared to DT (61.39%) and SVM (59.51%) methods. These results suggest that RF methods are efficient and accurate in predicting long-term cholesterol risk in the elderly.

Wu & Fang (2020) used the CLHLS dataset and focused on predicting stroke using LR and SVM. The LR method (71%) showed higher accuracy compared to the SVM method (67%).

Zulfiker et al. (2021) an in-depth analysis of the application of machine learning methods in behavioral science research, especially the XGBoost model. The method obtained an accuracy of 82.64% on an unspecified dataset. These results show that the XGBoost method is efficient and accurate in behavioral science research.

Qin et al. (2020) analyzed the effectiveness of LR, RF, SVM and XGBoost in predicting the health status of older adults was investigated on the CHARLS dataset. The LR method (67.2%) showed similar accuracy to SVM (67.2%) compared to RF (66.7%) and XGBoost (63.5%) methods.

Table 1 Performance comparison of different machine learning models

Citation	Model	Dataset	Result				
			LR	DT	RF	SVM	XGBoost
(Su et al. 2021)	LR, DT, RF, SVM	CLHLS	0.659	0.759	0.482	0.564	/
(Tarekegn et al. 2020)	LR, DT, RF, SVM	Piedmontese Longitudinal Study	0.69	0.67	0.68	0.68	/
(Cuaya-Simbro et al. 2020)	DT, SVM	The National Institute of Rehabilitation and at a community center in Mexico City.	/	0.6636	/	0.6545	/
(Tongkaw & Tongkaw 2020)	DT	/	/	0.67	/	/	/
(Fazakis et al. 2021)	DT, RF, SVM	ELSA	/	0.6139	0.6136	0.5951	/
(Wu & Fang 2020)	LR, SVM	CLHLS	0.71	/	/	0.67	/
(Zulfiker et al. 2021)	XGBoost	/	/	/	/	/	0.8264
(Qin et al. 2020)	LR, RF, SVM, XGBoost	CHARLS	0.672	/	0.667	0.672	0.635

Note: '/' means unknown or none.

B. Overview of Data Mining

New or non-intuitive knowledge can be obtained by cleaning, integrating, approximating, transforming, mining and evaluating large amounts of data(Olarte et al. 2019). It is a technique for

finding usable information and knowledge from rich data resources and synthesizes statistics, machine learning and artificial intelligence techniques. Data mining consists of the following steps：

Yang et al. (2020) show that in the real world, the collected data can not be directly used for data mining. Most of them are incomplete dirty data, which requires data cleaning, mainly dealing with problems such as inconsistent format, non-standard data, outliers, missing values, etc. Data integration is mainly responsible for integrating data from multiple sources into a unified data warehouse. The confirmation of the same entity in different data sources and the problem of redundant data need to be dealt with. Data reduction mainly simplifies the amount of data by selecting attributes and quantities. Data transformation mainly transforms data into a form suitable for processing, and its strategies mainly include removing noise in the data, data aggregation, generalization, and normalization. Developing a data mining model is to analyze and build models for different problems, build relevant models for different scenarios, test and evaluate the models, and interpret and analyze the results.

C. Conclusion

In this chapter, provides a review of the data mining literature, highlighting the scope, advantages, and general flow of data mining techniques. This section aims to present papers on ML and data mining in analyzing ageing datasets.

III. METHODOLOGY

This chapter is divided into the following sections: firstly, data collection was conducted by collecting the 2018 Chinese Longitudinal Health Factor Survey (CLHLS) (opendata.pku.edu.cn) and the 2018 English Longitudinal Study of Ageing (ELSA) (www.elsa-project.ac.uk). A data preview was undertaken to obtain the original dataset and observe its structure. Data preprocessing was then performed, which focused on cleaning the data, dealing with outliers and missing values, and selecting and normalizing data features. Data visualization is the visual analysis of data, which provides data analysis and organization for machine learning. The essential elements of the data can be displayed more intuitively through visual analysis. Then feature selection and extraction, through the maximum information coefficient (MIC) and Pearson's coefficient, analyze the intrinsic relationship between the features and the target variables as well as between the components and retain the most conducive to the regression (or classification) of the practical elements. Then the data plan is divided, and the model is built and trained: the data set is divided into the training set and the test set; the training set is used to train ten machine learning models, and then the test data set is used to conduct comparative experiments. Then perform ten-fold cross-validation: randomly divide the

entire dataset into ten groups, designate one group as the validation set, and the remaining groups as the training set. A total of ten validations are performed. Figure 1 represents the research methodology that is used in this study.

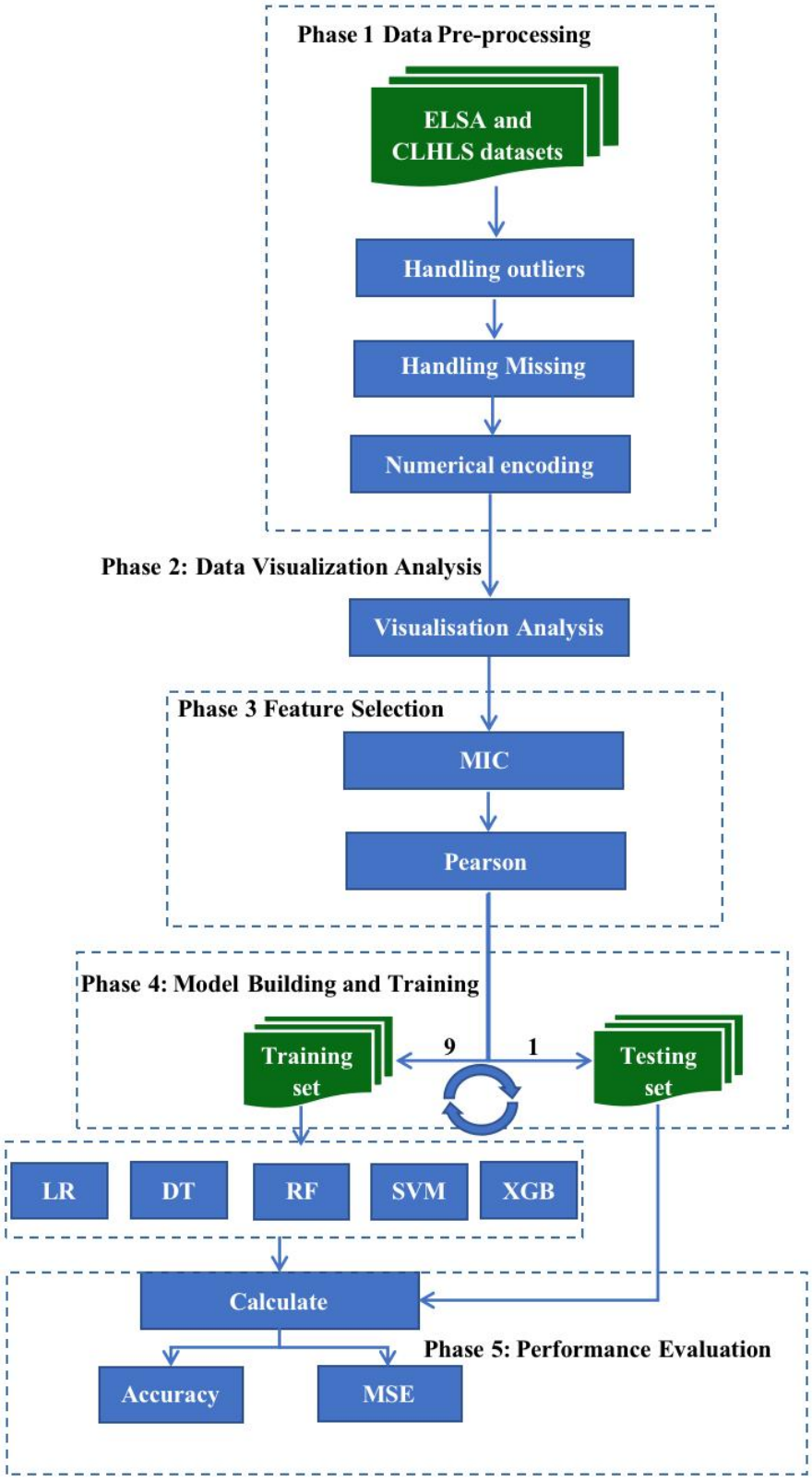


Figure1 Steps of Data Mining

A. Phase 1 Data Pre-Processing

To deal with outliers, the standard deviation method was used to detect and eliminate them. This is because some characteristic variables may be "999" to express uncertainty or other extreme equivalence uncertainties, which may significantly affect the prediction results. By addressing outliers, the accuracy and consistency of the data can be assured, thus mitigating any adverse effects they may have on the model. Regarding the treatment of missing eigenvalues, I use the median for interpolation. Filling the missing values with the mean of the feature preserves the distribution of the data, but may introduce distortion. Conversely, using the median helps minimize this distortion. In addition, for non-numeric features, the One-Hot Encoding method was used to convert to numeric form for further analysis. One-Hot Encoding converts categorical data into binary vectors, making it suitable for various machine learning algorithms and statistical analysis.

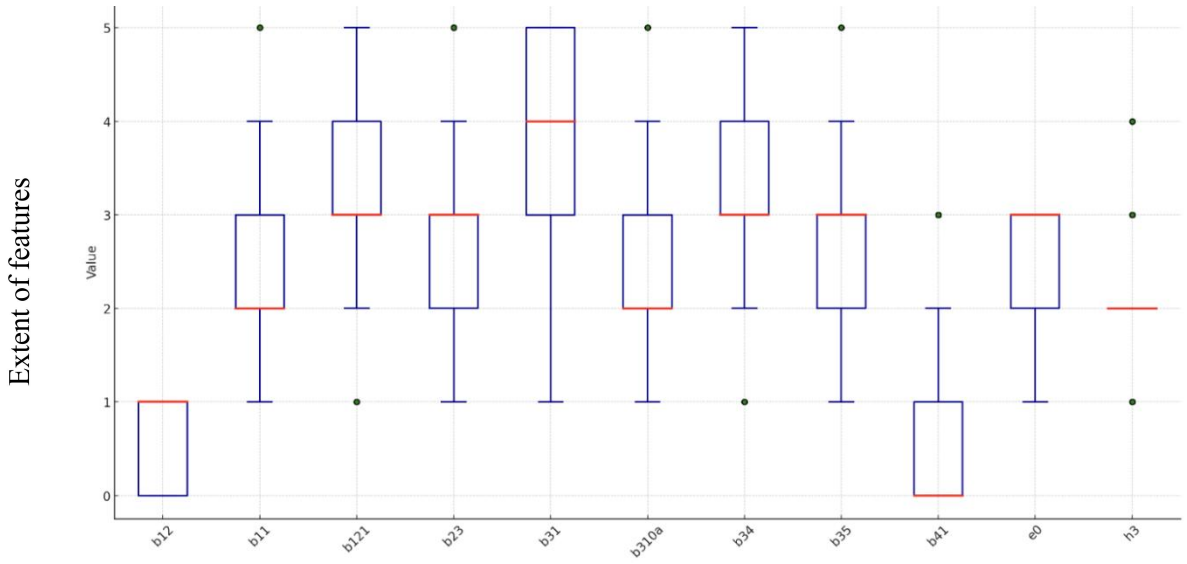
B. PHASE 2 DATAVISUALIZATION ANALYSIS

Boxplots provide essential information about the data distribution and help us understand various aspects of the dataset. Boxplots can provide valuable insights into the distribution of critical variables and their relationship to health when exploring factors related to population health and ageing.

Boxplots can help us detect outliers, such as data points that clearly fall outside the expected range, when examining population health and ageing factors. Outliers may indicate outliers or extreme values that require further investigation. For example, Sala et al. (2020) boxplots can show individuals with abnormally high or low disease incidence, indicating unique health status or potential measurement error.

In addition, box plots enable us to observe the expansion of the data. Wider boxplots indicate more excellent dispersion of the data, indicating more significant variability or heterogeneity in the health or ageing factors of the population under investigation. On the other hand, a narrower boxplot means a more concentrated distribution, implying lower variability or a more homogeneous population.

Boxplots can help us identify associations between age and other variables of interest in predicting population health and ageing factors. We may observe the effects of certain health conditions or risk factors on healthy ageing.



Frature

Case: blue

Median line: red

Whiskers and whisker ends: blue

Outliers: green

Figure 3.2 Box plots of CLHLS

In the Figure 3.2, the horizontal coordinates represent the features, and the vertical coordinates represent the extent of the features.

Box plots can be used to look at the overall distribution of data, using statistics such as median, 25% quartile, 75% quartile, upper boundary, and lower boundary to describe the widespread distribution of data. By calculating these statistics, a box plot is generated. The box contains most of the standard data, while those outside the upper and lower boundaries of the box are abnormal data, and no outliers can be found in the figure. The upper and lower limits of the box, which are the upper and lower quartiles of the data, respectively, mean that the package contains 50% of the data (for example, 'b11' ranges from 2.0-3.0). Therefore, the length of the box reflects the degree of fluctuation of the data to some extent. The flatter the chest, the more concentrated the data, and the shorter the end lines, the more focused the data ('h3'). In some extreme cases, the box is so flattened that there is only one line left, and there are also many outliers. There are two common reasons for these cases. Firstly, there are exceptionally large or small outliers in the sample data, and this outlier behaviour causes the boxes to be compressed as a whole, highlighting these anomalies instead. Secondly, the sample data is minimal, so the effects of the individual data magnify the boxes.

C. PHASE 3 FEATURE SELECTION

To determine the most relevant factors, two different methods were used: Maximum Information Coefficient (MIC) and Pearson correlation coefficient. These methods calculate the strength of the relationship between each feature and the target variable. After applying these measures, ten features were selected in the CLHLS dataset as the most relevant factors potentially affecting the health and ageing of the population.

D. PHASE 4 MODEL BUILDING AND TRAINING

Logistic regression is a standard binary classification method that provides the probability of each predicted category, which provides us with additional information, such as the likelihood that a person is predicted to be healthy. Logistic regression also has the advantage of being computationally efficient and easy to understand and interpret, which is essential for us to understand the factors that influence the health status of older people.

Decision trees are a non-parametric supervised learning method that learns decision rules from a set of features to predict the target value of a sample. An essential advantage of decision trees is that the model is highly interpretable, and it is clear which features have the most significant impact on the predicted outcome, which is very helpful in understanding and improving older people's health status.

Random Forest is an integrated learning method that constructs multiple decision trees and averages them to improve prediction performance. Random forests can effectively handle high-dimensional data, and there may be many features to consider when predicting the health of older adults. In contrast, random forests construct multiple decision trees and average their results, which can effectively reduce the variance of the model and prevent overfitting.

Support Vector Machine is a binary classification model that has the advantage of being able to handle high dimensional data efficiently and has excellent generalisation capabilities. In our problem, each feature (e.g., self-reported quality of life, change in health over the last year, etc.) can be considered as a dimension, and the SVM can effectively find a hyperplane to separate the two categories (healthy and unhealthy).

XGBoost is an integrated model based on gradient boosting, which improves the accuracy of predictions by combining the predictions of multiple weak classifiers (usually decision trees). In the project, the health status of older people may be affected by various factors, which means that we may need to consider a large number of features. XGBoost can efficiently deal with high-dimensional data, an essential advantage for our problem.

During the training process, each model will learn the relationship between feature vectors and the self-assessed health status of older adults population. Appropriate hyperparameters and cross-validation techniques will be used to fine-tune the models, ensuring optimal performance and preventing overfitting.

E. PHASE 5 PERFORMANCE EVALUATION

Accuracy is an intuitive measure of the correctness of a model's predictions. In this project, accuracy is the percentage of older adults whose health is correctly predicted by the model. However, accuracy is not always a perfect evaluation metric. Accuracy may give too optimistic evaluation results, especially in the uneven distribution of categories. Therefore, other evaluation metrics are needed to measure the model's performance entirely. Mean square error (MSE) is used to measure the error between the predicted and actual values of the model. In this study, we aim to categorise, treating the category labels as numerical values, in order to calculate the MSE.

IV. EXPERIMENT AND ANALYSIS

This chapter will explain how to improve the accuracy of the model's predictions by applying various machine learning algorithms. The performance of five algorithm models, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XGBoost Classifier, was evaluated on two datasets: the Chinese Longitudinal Healthy Longevity Survey (CLHLS) and the English Longitudinal Study of Aging (ELSA).

A. Experiment environment

Experimental environment specification plays a significant role in implementing machine learning techniques. Before we discuss the results of this work, it is worth explaining the specification of the environment used in the various stages of implementing the machine learning model. We used Python as our programming language because of its many advantages. These advantages include a large community, simplicity, platform independence, abundant libraries, and support for deep learning frameworks such as Scikit-learn, Pandas, Keras, and TensorFlow.

To implement Python code, we found it most convenient to use an interactive environment called Colab Notebook, which executes the Python language in a Web browser. With the Colab environment, we can use popular Python libraries to analyze and visualize data, such as NumPy and Matplotlib. Colab is widely used in the AI community. Colab notebooks can store and retrieve data in Google Drive. When using a Colab notebook, the code is executed on Google's cloud servers, which means that the execution process will be accelerated by Google hardware, including Gpus and Cpus.

B. Analysis of experiment results

In this experiment, we evaluate the performance of five classification models:Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and XGBoost (XGB). The accuracy of each model was calculated using ten-fold cross validation. The experimental results are as follows:

Table 2 CLHLS Accuracy

Algorithms	Accuracy
Logistic regression	0.776
Decision tree	0.683
Random forest Classifier	0.757
Support Vector Machine	0.729
XGBoost Classifier	0.755

Table 4.1 (CLHLS Accuracy): This table shows the accuracy of each algorithm on the CLHLS dataset. From this table, we can infer that Logistic regression performed best, with an accuracy of

0.776 (or 77.6%). On the other hand, the decision tree model performed the worst with an accuracy of 0.683 (68.3%).

Table 3 ELSA Accuracy

Algorithms	Accuracy
Logistic regression	0.774
Decision tree	0.709
Random forest Classifier	0.770
Support Vector Machine	0.753
XGBoost Classifier	0.741

Table 4.2 (ELSA Accuracy): This table shows the accuracy of each algorithm on the ELSA dataset. The random forest classifier performs best on this dataset with an accuracy of 0.770 (or 77%). The decision tree model again performs worst with an accuracy of 0.709 (or 70.9%).

In both tables, higher precision values indicate better performance. These tables help you understand how well each algorithm predicts the target variable for each dataset. The choice of the best algorithm often depends on the specific dataset and the task at hand. In this case, logistic regression and random forest performed the best in both datasets.

Table 4 CLHLS MSE

Algorithms	MSE
Logistic regression	0.253
Decision tree	0.303
Random forest Classifier	0.266
Support Vector Machine	0.241
XGBoost Classifier	0.245

Table 4.3 (CLHLS MSE): This table shows the MSE of each algorithm on the CLHLS dataset. The Support Vector Machine (SVM) model has the lowest MSE of 0.241, indicating that it performs best regarding prediction error. On the other hand, the decision tree model has the highest MSE of 0.303, suggesting that it serves the worst prediction error.

Table 5 ELSA MSE

Algorithms	MSE
Logistic regression	0.224
Decision tree	0.325
Random forest Classifier	0.238
Support Vector Machine	0.247
XGBoost Classifier	0.256

Table 4.4 (ELSA MSE): This table shows the MSE of each algorithm on the ELSA dataset. The Logistic regression model has the lowest MSE of 0.224, indicating that it performs best regarding prediction error. Meanwhile, the decision tree model has the highest MSE of 0.325, suggesting that it serves the worst prediction error.

Comparing the two datasets, we can see:

These models (SVM, logistic regression, random forest classifier, and XGBoost) exhibit decent accuracy levels, indicating that they generalize well to unseen data. This shows that they have successfully learned the underlying patterns and relationships in the data rather than simply memorizing the training samples. This is a positive sign that the model is not overfitting the data and will likely perform well in new instances. Regarding MSE, the Logistic regression model performs better on the ELSA dataset than the CLHLS dataset. Regarding MSE, the SVM model performs better on the CLHLS dataset than the ELSA dataset. Regarding MSE, the decision tree model performs the worst on both datasets. These results suggest that choosing Logistic regression may be the best model for these data sets.

C. *Discussion and analysis*

This chapter presents the experimental design for applying the above machine-learning algorithms to the CLHLS and ELSA datasets. Each model is evaluated based on two key metrics: accuracy and mean squared error (MSE). Based on the results, you can see that while all models show good accuracy and generalization ability, there are differences in their performance when considering prediction error (MSE). For example, the decision tree model offers the highest MSE on both datasets (0.303 and 0.325), indicating the worst performance in terms of prediction error. The final choice of the logistic regression model yielded higher accuracy (0.776 and 0.774) and lower mean squared error (0.253 and 0.224) on both datasets, indicating that it was more effective in predicting the target variable.

V. CONCLUSION

In this study, we apply a machine learning strategy to predict the health status of the elderly and compare the performance of several different models. The experimental results reveal that the logistic regression model exhibits the optimal performance among many models, which further confirms the effectiveness and robustness of the logistic regression model in predicting the health status of the elderly.

Our study further highlights the significant advantages of machine learning approaches over traditional social science methods. First, machine learning can capture nonlinear relationships in the data, which is crucial to predicting older adults' health status, as these nonlinear features may be determinants of health status. Second, machine learning methods can learn and select features directly from the data. This significantly saves time and avoids interference from a priori knowledge, resulting in more objective and accurate predictions.

However, while machine learning approaches have demonstrated advantages in predicting the health status of older adults, we should also recognize that traditional social science methods have not been rendered ineffective. Combining machine learning techniques with traditional social science methods allows for a more comprehensive understanding and explanation of the complexity of social issues and provides more accurate and in-depth results.

A. Research summarize

This study compares different classification methods to analyse two ageing datasets: the English Longitudinal Study of Ageing (ELSA) and the Chinese Longitudinal Healthy Longevity Survey (CLHLS). A machine learning approach was used to precisely tap into underlying factors affecting the health of older adults, such as life satisfaction, hope for the future, and depressive symptoms, as well as physical factors, such as sleep quality, ability to perform household tasks, and difficulty with activities of daily living. The performance of algorithms such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and XGBoost (XGB) on both datasets was also compared to find the most effective model to predict the ageing population's health status of the ageing population. The results show that the logistic regression model is the best predictor on the CLHLS and ELSA datasets, with an accuracy of 0.776 and 0.774 and a mean square error of 0.253 and 0.224, respectively, which suggests that it can capture the underlying patterns and relationships in the data and that it has good validity and robustness in predicting the health of older adults.

B. Future work

For future works, we can continue optimizing the models based on what we have, exploring integration methods and introducing more relevant features to further improve predictions' accuracy and stability. In addition, these models can also be applied to other areas related to healthy ageing, such as healthcare planning, decision-making, and personalized interventions, contributing to the well-being of the global ageing population. By utilizing computational intelligence, we can gain deeper insights into healthy ageing that will profoundly impact our society.

ACKNOWLEDGEMENT

The authors would like to thanks, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia and University of Malaya by giving the authors an opportunity to conduct this research.

REFERENCE

Chan, E.-Y., Samsudin, S.A. & Lim, Y.J. 2020. Older patients' perception of engagement in functional self-care during hospitalization: A qualitative study. *Geriatric Nursing* 41(3): 297–304.

Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D. & Choi, K. 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry* 20(2): 154–170.

Chew, S.M., Lee, J.H., Lim, S.F., Liew, M.J., Xu, Y. & Towle, R.M. 2021. Prevalence and predictors of medication non-adherence among older community-dwelling people with chronic disease in Singapore. *Journal of Advanced Nursing* 77(10): 4069–4080.

Cuaya-Simbro, G., Perez-Sanpablo, A.-I., Muñoz-Meléndez, A., Uriostegui, I.Q., Morales-Manzanares, E.-F. & Nuñez-Carrera, L. 2020. Comparison of Machine Learning Models to Predict Risk of Falling in Osteoporosis Elderly. *Foundations of Computing and Decision Sciences* 45(2): 66–77.

Elshawi, R., Al-Mallah, M.H. & Sakr, S. 2019. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making* 19(1): 146.

Fazakis, N., Dritsas, E., Kocsis, O., Fakotakis, N. & Moustakas, K. 2021. Long-term Cholesterol Risk Prediction using Machine Learning Techniques in ELSA Database: *Proceedings of the 13th International Joint Conference on Computational Intelligence*, hlm. 445–450. SCITEPRESS - Science and Technology Publications: Valletta, Malta.

Lim, M.H., Eres, R. & Vasan, S. 2020. Understanding loneliness in the twenty-first century: an update on correlates, risk factors, and potential solutions. *Social Psychiatry and Psychiatric*

Epidemiology 55(7): 793–810.

Müller, S.R., Chen, X. (Leslie), Peters, H., Chaintreau, A. & Matz, S.C. 2021. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports* 11(1): 14007.

Naah, F.L., Njong, A.M. & Kimengsi, J.N. 2020. Determinants of Active and Healthy Ageing in Sub-Saharan Africa: Evidence from Cameroon. *International Journal of Environmental Research and Public Health* 17(9): 3038.

Ngiam, K.Y. & Khor, I.W. 2019. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology* 20(5): e262–e273.

Olarte, E., Panizzi, M. & Bertone, R. 2019. Market Segmentation Using Data Mining Techniques in Social Networks. *Computer Science – CACIC 2018*, hlm. 221–231. Springer International Publishing: Cham.

Qiu, W., Chen, H., Dincer, A.B., Lundberg, S., Kaeberlein, M. & Lee, S.-I. 2022. Interpretable machine learning prediction of all-cause mortality. *Communications Medicine* 2(1): 1–15.

Rodriguez-Laso, A., McLaughlin, S.J., Urdaneta, E. & Yanguas, J. 2018. Defining and Estimating Healthy Aging in Spain: A Cross-sectional Study. *The Gerontologist* 58(2): 388–398.

Sala, G., Chakraborti, R., Ota, A. & Miyakawa, T. 2020. Association of BCG vaccination policy and tuberculosis burden with incidence and mortality of COVID-19. medRxiv.:

Su, D., Zhang, X., He, K. & Chen, Y. 2021. Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. *Journal of Affective Disorders* 282: 289–298.

Tarekegn, A., Ricceri, F., Costa, G., Ferracin, E. & Giacobini, M. 2020. Predictive Modeling for Frailty Conditions in Elderly People: Machine Learning Approaches. *JMIR Medical Informatics* 8(6): e16678.

Tongkaw, A. & Tongkaw, S. 2020. Prediction Medical Problem of Elderly People by Using Machine Learning Technique. *Journal of Physics: Conference Series* 1529(3): 032083.

Wu, Y. & Fang, Y. 2020. Stroke Prediction with Machine Learning Methods among Older Chinese. *International Journal of Environmental Research and Public Health* 17(6): 1828.

Xu, F., Earp, J.E., Greene, G.W., Cohen, S.A., Lofgren, I.E., Delmonico, M.J. & Greaney, M.L. 2020.

Temporal Association between Abdominal Weight Status and Healthy Aging: Findings from the 2011–2018 National Health and Aging Trends Study. *International Journal of Environmental Research and Public Health* 17(16): 5656.

Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., Zheng, S., Xu, A. & Lyu, J. 2020. Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine* 13(1): 57–69.

Zulfiker, Md.S., Kabir, N., Biswas, A.A., Nazneen, T. & Uddin, M.S. 2021. An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences* 2: 100044.