1. Introduction

As they get older, some people experience general deterioration in their health and a gradual loss of the ability to perform basic but important daily activities such as bathing, washing, and eating. In recent decades, the demographic structure of many countries has changed rapidly. The elderly were the fastest growing group in the population. At the same time, huge amounts of public money are spent each year to meet the care needs of older people.

Health and social services help older people recover from illness, relieve symptoms of illness, enable them to live independently in their own homes, and improve people's quality of life and general well-being. However, care resources are limited, making effective allocation and provision of quality services to older people a formidable challenge for both developing and developed countries.  
It is therefore important to be able to accurately predict the health status of older people so that services can be tailored to their care needs.

To accurately predict the health status of older people, we first need to analyse the factors most relevant to their health. Against this background, a vast body of literature has addressed this issue, and knowledge and expertise in the field are rapidly accumulating. In particular, the availability of large datasets in the information age offers researchers the opportunity to understand more clearly the complex relationships that exist between variables, allowing them to conduct more rigorous studies. . However, as the amount of information accessible to researchers has increased exponentially, how to effectively process that information (i.e., identify the most relevant predictors and use them to generate accurate  
The practical challenge of making predictions) arises. It takes a long time for researchers to familiarize themselves with these large His  
datasets and identify His predictors not found in the existing literature (Manual feature engineering). Once this task is completed, new data will be available and the processed information would already be out of date.

Machine learning techniques can be used to extract non-linear and seemingly insignificant constituents that are difficult to find with traditional techniques, allowing the feature selection process to be completed more accurately. Machine learning techniques can also significantly reduce the time it takes to extract key elements from large datasets, ultimately improving prediction accuracy.

This paper is organized as follows. First, a brief overview of the topic and background for consideration. The second section reviews the existing literature on determinants of health in older populations. The third part presents a data source description and test plan. Part 4 provides an overview of the test results. Part 5 discusses the methodology and theoretical implications of the research based on the research findings. Part 6 summarizes the main conclusions of this study. In machine learning literature, input variables are usually called features, whereas in social science research, input variables are called predictors or independent variables.

2. Literature review

The World Health Organization (2015) defines health as "a state of complete physical, mental and social well-being, not merely the absence of disease or infirmity" . Health is a multidimensional structure. For older adults, physical health, mental health, and cognitive performance are closely related to quality of life. Regarding physical health, there are three commonly used indicators in the literature. Presence of disease, level of functioning, and subjective assessment of health (O'Donnell et al, 2008). This study focuses on two indicators of physical health.

Many studies have investigated the factors that influence physical performance and self-reported health in older adults. These factors fall into four categories: elderly demographic factors, socioeconomic factors, lifestyle factors, and disease factors.

Research in the field of demography has focused on the age, gender and lifestyle of older people. Age is considered the most important cause of functional limitations in the elderly.

Previous studies have shown that the socioeconomic environment in which older people live can have a significant impact on their level of functioning and overall health. A study on the functioning of the elderly in Bangladesh showed that location has a significant impact. Higher mortality and disability rates have been observed in rural areas compared to urban areas.

3. Experiment

This experiment is mainly split up into the following sections:  
l Data Preview. Take the original data set, import it into the  
database, and observe its data structure.  
lI Data Merge. In the original dataset, the questionnaire for each topic is separate. It makes sense to merge them into a single dataset to simplify the analysis.  
lII Feature selection. Apply machine learning algorithms to extract the factors that have the greatest impact on the health of older adults.  
lV Data processing. Process the data until you can use the data to train a machine learning model.  
V Training model. Train multiple machine learning models  
using the comparative training set.  
Vl 5-fold cross-validation. The entire data set is randomly divided into 5 groups, one group is called the validation set, the rest are called the training set, and the validation is run 5 times.  
VIl Result. Get results and further summarize conclusions



WORKFLOW

*Data Preview*

Data extraction is the first step throughout the  
experiment. Using multistage sampling and measurement of  
samples.

DEMOGRAPHIC BACKGROUND: name, date of birth,

address, etc.

l FAMILY: parent, childrearing and sibling information, time

transfers, etc.

l HEALTH STATUS AND FUNCTIONING: health status,

functional limitations, helpers, etc.

l HEALTH CARE AND INSURANCE: medical insurance,

health care costs and utilization.

l WORK, RETIREMENT, AND PENSION: job status, fringe

benefits, etc.

l INCOME, EXPENDITURE, AND ASSETS: household

income and expenditure, household assets, etc.

l HOUSING CHARACTERISTICS: total housing land area,

the year when the house was built, etc.

*Data Merging*

Once all the data was entered into the database, each data table  
was added as a large table based on the respondent's ID. The  
database chosen here is MongoDB.  
In MongoDB, you can store data in the document, and the data is stored as key-value pairs. The key is used to uniquely identify the document and is of type string, but the value can be a variety of complex file types. This form of storage is called BSON.

BSON is a JSON like binary storage format,  
is also called Binary JSON for short. Such characteristics mean that MongoDB performs better in this experiment compared to SQL-type databases, mainly because we used non-indexed add, delete, modify, and search operations.

*Feature Selection*

We used the maximum information coefficient (MIC) and the Pearson  
correlation coefficient to calculate the most relevant factors.  
MIC is a new correlation statistic that measures the strength of linear or nonlinear correlation between two variables X and Y. Pearson Correlation Coefficient (PCC) is a measure of linear correlation between two variables

*Data Processing*

Machine learning models such as linear regression and  
support vector machines (SVM) are very sensitive to missing  
values ​​in the data set. We imputed missing values ​​before function selection. We experimented with four machine learning models, including linear regression, k-nearest neighbor (kNN), decision tree, and XGBoost. We used XBGoost algorithm because it gave the best performance for imputation.

Health status was divided, with 1 representing good health and 0 representing poor health.

Data normalization has been an important part of data processing.  
Some machine learning models are sensitive to different scales in the data set. To prevent large values ​​from overwriting small values ​​in the data, we rescaled all features so that the values ​​are in a consistent range between 0 and 1. This is done by subtracting the original values ​​of the features from the mean and dividing them by the variance.

*Training*

Our study focuses on the performance of the nonlinear  
model in predicting health status in the elderly, so the linear  
model is mainly used as a complementary comparative study. Here we discuss the artificial neural network that played a central role in the experiment.



*Cross-validation*

To run a machine learning approach, you usually need to split your data set into a training set and a test (or validation) set. Models built on the training set are tested on the test set to ensure that the model generalizes well when new data is received and does not overfit the data. This  
experiment used 5-fold cross-validation as the validation approach. We randomly divided the sample into five equal-sized her subsamples. Of the 5 subsamples, 1 subsample (i.e. 20% of of all samples) was kept as validation data and the remaining subsamples were used as training data. We then repeated this procedure 5 times and took the average result of out of 5 repetitions as our final estimate. The final results of the review show that the model does not appear to be overfitted, indicating that the model is suitable for use in predicting unknown data.

Conclusion

We use multiple machine learning models to predict the health status of  
older people in recent years. Machine learning method differs from conventional method used. In summary, the overall workflow of machine learning techniques can reduce the impact of experience and rely more on data for accurate predictions. It can be used to build data-driven general paradigm research that can be transferred to any domain.

References

Aldaz, J. M., Barza, S., Fujii, M., & Moslehian, M. S. (2015).

Advances in Operator Cauchy-Schwarz inequalities and

their reverses, *Annals of Functional Analysis, 6*(3): 275-

295. https://doi.org/10.15352/afa/06-3-20.

Artaud, F., Dugravot, A., Sabia, S., et al. (2013). Unhealthy beha

viors and disability in older adults: Three-City Dijon cohort

study. *BMJ, 347*: f4240. https://doi.org/10.1136/bmj.f4240.

Avlund, K., Vass, M., Lund, R., Yamada, Y., & Hendriksen, C.

(2008). Influence of psychological characteristics and socia

l relations on receiving preventative home visits in older me

n and women. European Journal of Ageing, 5(3), 191-201.

https://doi.org/10.1007/s10433-008-0086-4.

Benfante, R., Reed, D., & Brody, J. (1985). Biological and social

predictors of health in an aging cohort. *Journal of Chronic*

*Diseases, 38*(5): 0-395.

Bonsang, E. (2009). Does informal care from children to their ol

d people parents substitute for formal care in Europe? *Journ*

*al of Health Economics*, 28, 143–154. https://doi.org/10.101

6/j.jhealeco.2008.09.002.

Bookman, A., Harrington, M., Pass, L., & Reisner, E.

(2007). *Family Caregiver Handbook*. Cambridge, MA:

Massachusetts Institute of Technology.

Boult, C., Kane, R. L., Louis, T. A., & McCaffrey, D. (1994). C

hronic conditions that lead to functional limitation in the eld

erly. *Journal of Gerontology, 49*: M28-M36.

Breiman, L. (2001). *Random Forests. Machine Learning 45*(1),

5-32.