**Introduction**

Dementia is a general term used to refer to the impaired ability to remember, think and make decisions hence tampering with persons daily activities. Those with dementia normally have problems with; communication, attention, memory, reasoning, judgment and solving problems. Some of the factors associated with dementia are age, race, family history, poor health, body mass index etc.

Effective treatment and care of dementia rely on early detection and risk prediction. Data driven models are becoming increasingly crucial in dementia research. This is because they can be used for prediction of risks using large and complex datasets. The major aim of this project is to use machine learning prediction models to identify patterns and relationships regarding dementia and risk factors.

**Data description**

The data set used in this study was obtained from a mobile health care service offered in collaboration with non-governmental organizations that run elderly care centers. This service was provided to elderly people residing in various districts of Hong Kong for free from 2008 to 2018. The data set consists of 2299 cases, each of which includes eleven variables. These variables include age, body height, body weight, education level, financial support, geriatric depression scale score, out-of-pocket financial source (whether they were independent or dependent on family), marital status, Mini Nutritional Assessment Part A score, Mini Nutritional Assessment Part B score. The outcome labels were based on the categories of the Mini Mental State Exam.

**Data pre-processing**

Some data transformation techniques were applied on the data to make it appropriate and easily analyzable. The missing values in the dataset was dealt with by imputing the categorical variables with their mode and the numeric variables with mean. Body mass index is one of the risk factors associated with dementia. The body mass index (BMI) was calculated by dividing the weight of the respondent by height squared. The marital status of the respondent was also recoded into two categories (Married and Not married).

**Exploratory data analysis**

|  |  |  |
| --- | --- | --- |
| Variable | Category | Frequency |
| Gender | Male  Female | 1149  566 |
| Education | No education  Primary school  Secondary school  Tertiary | 513  710  418  74 |
| Marital status | Single  Married  Widowed  Separate  Divorced  Married but not live spouse | 90  969  562  12  81  1 |
| Mini mental state exam | Negative  Positive | 1428  287 |
| Financial status | Independent  Comprehensive social security assistance  Disability allowance  Old age allowance  Depend on family with Old Age Allowance  Depend on family with Disability Allowance  Depend on family with CSSA | 980  291  125  365  492  14  11  21 |

Table 1: Frequency table

Frequency distribution table in table 1 above illustrate the number of respondents in each category. The results shows that there were more males (1149) in the study than females (566). It can also be evident that most participants had primary level of education (N=710) while the least had up to tertiary level of education (N=74). It is also shown that most participants were married (N=969) compared to those who were divorced, single, separated and widowed combined (N=746). Financial status of the respondents was also explored and the results shows that most of them depended on family with old age assistance (N=492) Most participants in the sample tested negative for dementia (N=1428) while only 287 tested positive.

|  |  |
| --- | --- |
| Average age | 77.73 |
| Average Bmi | 24.04 |
| Average\_GDS | 2.16 |
| Average\_MNA | 12.72 |
| Average weight | 57.19 |
| Average height | 154.24 |

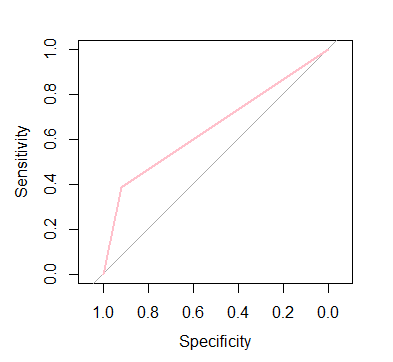
Table 2: Summary statistics

Summary statistics in table 2 above was used to describe the numeric variables used in the study. The average age of the participants in our sample was 77 years. The mean height of the participants in the study is 154.24cm while the mean weight was 57.19. The average BMI was also found to be 24.04 which is ideal for any human being. The averageGeriatric Depression scale was found to be 2.16 which means that the population was normal in average. Average mini nutritional assessment was also calculated and found to be 12.72 implying the sample was protein-calorie malnourished.

**Prediction modelling**

Two prediction models were used to predict the mini mental state exam outcome which determines the presence of dementia in an individual. These models used were logistic regression model and classification decision tree model. Logistic regression model is a classification algorithm used predict the category of individual based one or more variables. This model is appropriate when the outcome variable is binary. Decision trees are machine learning algorithms used for both classification and regression tasks. Classification decision trees model are appropriate when the outcome variable is categorical. Both of these models were used to predict dementia based on the risk factors available. The dependent variable is mini mental state exam outcome (Negative=0, positive=1). The accuracy of these prediction models was determined. Both of the models had an accuracy of 83%. This implies that 83% of the participants were correctly classified. The ROC curves for both models were also drawn to further compare their performances.



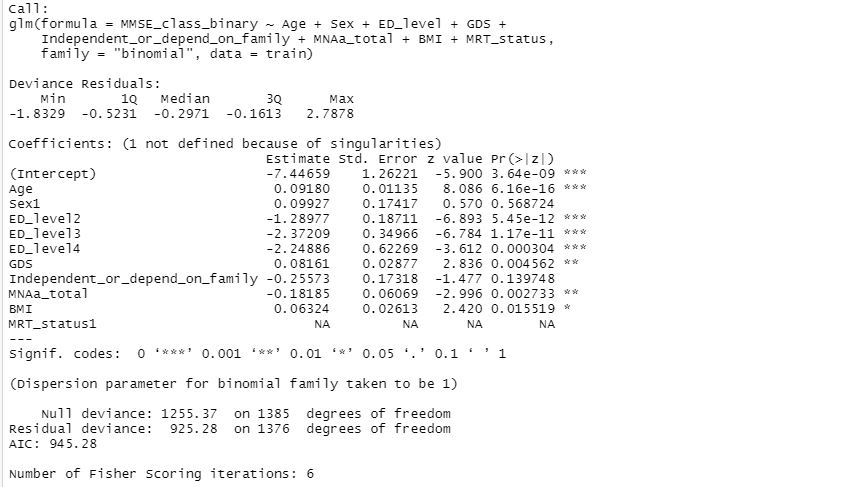


AUC= 0.652

Figure 2: ROC curve for classification tree model.

Logistic regression model performs better than classification tree model. This is because the area under the curve (AUC=0.852) for logistic regression model is greater than that of classification tree model (AUC=0.652).

**Results and discussion**



**Table 3: Logistic regression model**

Logistic regression results in table 3 above shows that only Sex and whether respondents depend on family or not did not have a significant effect on dementia. Age, depression and BMI had a significant positive impact on Dementia. This means an increase in these factors has a significant increase in probability of being a dementia positive. Education level and malnutrition had a significant negative effect on dementia. This imply that an increase in these factors is associated with a decrease in probabilities of being dementia positive. The accuracy of the model was calculated and the misclassification error was found to be 16.97%. The accuracy of the model is calculated as (1-16.97%) =83.03%.

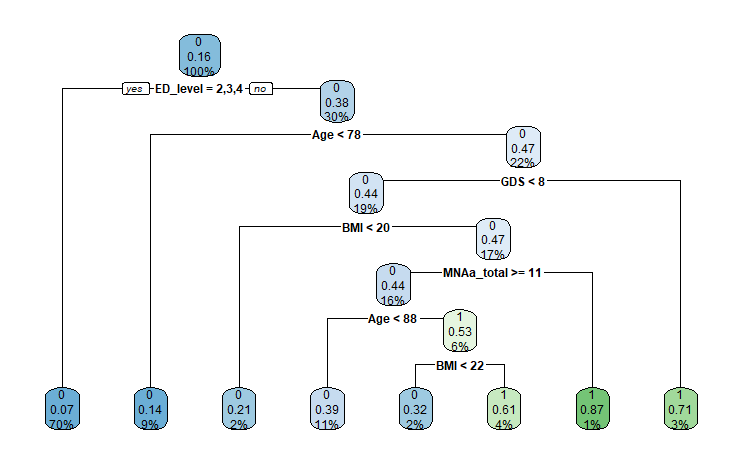


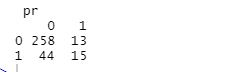
Figure 2: Decision tree model.   


Table 4: Confusion matrix

Decision tree model results shows that 16% of the respondents were categorised as being dementia positive. Only 1.4% of those aged less than 78 were classified as being dementia positive. 17% of the participants had BMI greater than 20 and 4.7% of them were classified as dementia positive. Participants with malnutrition less than 11 were 16% and among them 4.4 % were dementia positive. The confusion matrix in table 4 was used to calculate the accuracy of the tree model (13+44/ (258+13+44+15)). The misclassification error of the model is 17.27%. The accuracy of the classification tree model is (1-17.27%) = 82.73%.

**Conclusion**

Machine learning models are important in determining the risk factors associated with dementia. Both decision tree model and logistic regression model shows that Education level, age, Body mass index and malnutrition significantly predicts dementia. Sex and dependence on family do not have an impact on dementia. Both models used in the analysis were had the same accuracy.

**References**

Chen, Jen-Hau, Kun-Pei Lin, and Yen-Ching Chen. "Risk factors for dementia." Journal of the Formosan Medical Association 108.10 (2009): 754-764

Geldmacher DS, Whitehouse PJ. Evaluation of dementia. New England Journal of Medicine. 1996 Aug 1;335(5):330-6.

Kitwood, Tom. "The experience of dementia." Aging & mental health 1.1 (1997): 13-22.

Kleinbaum, David G., et al. Logistic regression. New York: Springer-Verlag, 2002.

LaValley, Michael P. "Logistic regression." Circulation 117.18 (2008): 2395-2399.

Leys, Didier, et al. "Poststroke dementia." The Lancet Neurology 4.11 (2005): 752-759.

Loh, Wei‐Yin. "Classification and regression trees." Wiley interdisciplinary reviews: data mining and knowledge discovery 1.1 (2011): 14-23.

Rabins, Peter V., Nancy L. Mace, and Mary Jane Lucas. "The impact of dementia on the family." Jama 248.3 (1982): 333-335.