Introduction

Dementia, a neurological disorder, progressively disrupts the memory, reasoning, and ability of a patient to function. Detecting the disorder early is essential so that the right interventions can be made before severe damage and changes occur. The Mini-Mental State Examination (MMSE) is one of the traditional diagnostic tools that assist with early detection, but it is known to be time-consuming and subjective, depending on the doctor conducting the medical assessment.

A recent study by Yassine et al. (2022) highlights the impact of metabolic and nutritional health on cognitive health the findings concluded that malnutrition and poor dietary habits are closely related to fast cognitive decline. This study emphasized the use of nutritional tools, such as the Mini Nutritional Assessment (MNA) [1]. Liang et al. (2022) equivalently showed through their research that a Body Mass Index (BMI) that is either too low or too high, as well as a waist circumference that is either too small or significantly above average, is also associated with changes in cognitive abilities among older adults [2].

This project will be built upon the findings analyzed from the dataset by the author, utilizing machine learning techniques. The two main techniques employed by are logistic regression and decision tree classification, which are used to predict dementia risk by analyzing data such as age, BMI, MNAa_tot, MNAb_tot, and Waist circumference. This project aims to evaluate how physiological and nutritional factors can serve as indicators of cognitive decline in elderly individuals with dementia.

Data Description

The ICT583 S2 2025 CSV dataset was analyzed in this project. The data were derived from a mobile healthcare service conducted between 2008 and 2018 in collaboration with non-governmental organizations that operate elderly care centers in various districts of Hong Kong. The organization provided different services such as physical, nutritional and mental assessment for elders in those districts. The dataset contains data that has individual records for each annotated by the ID feature in the dataset and a mix of continuous and categorical variables that were used to assess and model the risk of dementia in individuals, which is identified by the output variable MMSE-class (0 = not at risk; 1 = at risk). The continuous variables associated with the dataset are as follows:

- Ages (years)
- Body_height (kg)
- Waist (cm)
- BMI (kg/m²)
- MNAa_tot and MNAb_tot variables and scores of the Mini Nutritional Assessment that are related to dietary intake and health.

The categorical variables associated with the dataset, which include MNAa-q3, Hyperlipidaemia, Mobility_ID, and Education_ID, represent the individual's lifestyle that may influence the risk of dementia.

The preliminary investigations were executed in R-Studio and the following functions were used: head(), names(), and str(). The results when the functions were run in R confirmed that both numeric and factor data were present. The missing values were inspected using the is.na() %>% colSums() and sapply(), which indicated that there were limitations in data analysis due to missing data from different variables. When these were removed, the dataset was suitable for case analysis. The further findings of the project indicated that when the data were summarized using statistical measures, the participants in the dataset were elderly adults with a normal BMI and various mobility methods. These features are well aligned with previous studies that were carried out on dementia risk research and provide a good foundation for classification modeling using logistic regression and decision tree analysis.

Data Pre-Processing

Before performing the preliminary investigations, the author pre-processed the dataset by analyzing it and cleaning the contained data to ensure that the values were suitable for analytical purposes. Firstly, the missing values were first analyzed using the data %>% is.na() %>% colSums() functions and operators. Secondly, the is.na() function was used to mark all the missing cells as TRUE and the colSums() function was used to essentially add up these TRUE values for all the variables in the dataset giving the total number of the missing values per column.

The author then created a small function to double check the results f <- function(x) sum(is.na(x)). The following functions were then used in different ways repeatedly:

- apply() R applies this function in different columns of the data frame.
- lapply() then returns the list of the number of values that are missing.
- sapply() then simplifies the list summarized by the lapply() function into a clear numeric summary.

The pre-processing procedures indicated that several variables, such as body weight, body height, Mobility, Age, and others, contained missing values; however, the MMSE class was complete. The number of values that were missing was minimal, therefore, the author used na.omit(data) to remove the rows that contained missing values, and executing this function resulted in the complete observations being kept. The variables were then prepared using the mutate() function, which updated and transformed the columns within the dataframe. The structure of the dataframe was then confirmed by the author using str(data) and verifying that there were no missing values that remained using colSums(is.na()). The author then confirmed that the dataset was clean, consistent and ready for exploratory data analysis.

Exploratory Data Analysis

The author carried out exploratory analysis once the dataset was cleaned to understand how the key variables are related to each other before predictive modeling was executed. The visualization approach of exploratory data analysis was used to analyze the categorical and continuous variables using ggplot2. To compare the BMI between two dementia risk groups which were classified as “at risk = 1” and those “not at risk = 0”, a box plot was executed using the ggplot(data, aes(x = factor(MMSE_class), y = BMI)) + geom_boxplot(). To compare and explore the relationship between Age and BMI a scatterplot was used (geom_point()). The weak results from the scatterplot, which were portrayed in a pattern that was almost horizontal, indicated that the correlation matrix (r = -0.018) was practically negligible. A histogram of Age (geom_histogram(binwidth = 1)) was used and indicated that most of the individuals in the dataset were elderly people ranging between the ages of 60 and 90 years old, which is closely related to the ages that are usually preferred for dementia screening.

A correlation analysis was then performed by the author using pairs() and cor(), which resulted in the findings of body size measurements being strongly and positively related such as BMI and waist (r = 0 0.787), BMI and body weight (r = 0.770) and waist and body weight (r = 0.751). Nutritional Assessment score were also closely and positively related with these measures; MNAa_tot and BMI (r = 0.640) and MNAa_tot and waist (r = 0.611) showing that higher nutritional status in elderly people is related to a higher body weight. On the other hand, Age indicated mildly negative correlations related to height (r = −0.536) and MNAb_tot (r = −0.526) which shows that activity ability and nutritional requirements decline with ageing. The nutritional assessment scores were mildly related with a score of MNAa_tot ↔ MNAb_tot r = 0.616.

Prediction Modeling

The author made use of two prediction models; Logistic Regression and a Decision Tree Classifier. The dependent variables used in this project were used MMSE_class (0 = not at risk, 1 = at risk) and the predictors were Age, BMI, waist, MNAa_tot and MNAb_tot. The logistic regression (glm(formula = MMSE_class ~ Age + BMI + waist + MNAa_tot + MNAb_tot, family = binomial(logit))) showed that Age (p < 0.001), BMI (p < 0.001), waist (p = 0.021) and MNAb_tot (p < 0.001) were the main dementia risk predictors. Other factors such as old age, higher BMI and a large waist circumference also increased the odds of an elderly person being at risk MMSE_class (1 = at risk) but a higher MNAb_tot (better nutritional assessment score) reduced the risk of developing dementia.

The model (MMSE_class ~ Age + MNAa_tot) proved to be signification in this project and had an overall of (p < 0.001) and an AIC of 1299.8. The accuracy of this model was 84.5%, the sensitivity was 95% and the specificity was 34% proving that the model was able to identify most of the cases that were low risk but was not able to identify the risk cases. The ROC plotted curve that was executed using the ROCR package identified these discrepancies and moderate discrimination.

The author then used the rpart() to execute a Decision Tree with the same predictors. This resulted in the model having 91.2 % accuracy, 94 % sensitivity and 76 % specificity as well as a balanced accuracy of 0.85. This indicates that the tree was able to identify both of the classes fairly and had higher agreement (k = 0.70) when compared to the logistics model (k = 0.35). The decision tree performed better than the logistics regression model on this dataset and offered higher accuracy levels.

Results and Discussion

The results from both models offer important and contributory data about dementia risk classification. The logistics regression model proved that the variables Age, BMI, waist circumference, and MNAb_tot were important predictors with a p score of ($p < 0.05$). The positive coefficients for Age, BMI, and waist suggested that elderly people in this study with a higher body weight and larger waist circumference were more than likely to be classified as at risk but on the other hand the negative coefficient for MNAb_tot showed that higher nutritional assessment scores reduced the overall risk of dementia. The authors' findings were closely related to the studies completed by Yassine et al. (2022) and Liang et al. (2022) that found that age and nutritional assessment scores were strongly related to cognitive dementia decline later in life.

The logistic regression model yielded an overall accuracy of 84.5%, with high sensitivity (95%) and low specificity (34%). This means that the model was capable in identifying the non risk cases and not the risk cases. The ROC curve confirmed that there was mild discrimination between the classes. On the other hand, the decision tree classifier that was executed using rpart() had higher predication levels with a 91.2 % accuracy, 94 % sensitivity, and 76 % specificity and a balanced accuracy of 0.85. The higher K matrixes of 0.70 vs 0.35 for logistic regression show that there was a stronger relation between predicted and actual classes. The structure of the decision tree show that Age was the main value in the splitting analysis followed by BMI and MNAb_tot, proving that age and nutrition are the most important features. Both of these models prove that age and nutritional health are major factors in dementia risk.

Conclusion

This project was completed using predictive modelling; logistic regression and decision tree classifiers to help predict the risk of dementia using clinical and nutritional variables. The dataset used for this project was cleaned and transformed for the evaluations and analyses demonstrated. The models used identified and confirmed that Age, BMI, and MNAb_tot are key predictors, providing strong evidence that links nutrition and cognitive health. The decision tree was able to produce more accurate results and had a balanced performance, making it more suitable for practical risk assessment. To conclude, this project demonstrates how statistics and machine learning techniques can aid in early dementia detection for elderly individuals and identify the associated risk factors. These findings can help improve clinical decision-making.

References

1. Yassine, H. N., Samieri, C., Livingston, G., Glass, K., Wagner, M., Tangney, C., Plassman, B. L., Ikram, M. A., Voigt, R. M., Gu, Y., O'Bryant, S., Minihane, A. M., Craft, S., Fink, H. A., Judd, S., Andrieu, S., Bowman, G. L., Richard, E., Albensi, B., & Meyers, E. (2022). Nutrition state of science and dementia prevention: recommendations of the Nutrition for Dementia Prevention Working Group. The Lancet Healthy Longevity, 3(7), e501–e512. https://doi.org/10.1016/s2666-7568(22)00120-9

2. Liang, F., Fu, J., Moore, J. B., Zhang, X., Xu, Y., Qiu, N., Wang, Y., & Li, R. (2022). Body Mass Index, Waist Circumference, and Cognitive Decline Among Chinese Older Adults: A Nationwide Retrospective Cohort Study. Frontiers in Aging Neuroscience, 14. https://doi.org/10.3389/fnagi.2022.737532

3. Murdoch University, ICT583 Data Science Application Project Tutorials 04-08, School of Information Technology, Semester 2, 2025. (Internal teaching material)