Assignment 2

**Introduction:**

This study aims to investigate the relationship between the variables in the dataset and the cognitive status of elderly participants, as determined by the MMSE categories. By analyzing this dataset, valuable insights can be gained into the factors that impact cognitive health in the elderly population. The findings of this study have the potential to contribute to the development of effective interventions and support strategies. Our goal is to enhance the understanding of factors influencing cognitive health among the elderly in Hong Kong and similar contexts, thereby providing practical implications for improving their well-being and quality of life.

**Dataset Description:**

This study utilizes a dataset obtained from a mobile health care service in collaboration with non-governmental organizations operating elderly care centers in various districts of Hong Kong. The service was provided free of charge from 2008 to 2018, resulting in a dataset containing 2,299 cases with eleven variables.

These variables provide valuable information about the elderly participants, covering demographic and health-related factors such as age, body height, body weight, education level, financial support, geriatric depression scale score, out-of-pocket financial source (indicating independence or dependence on family), marital status, Mini Nutritional Assessment (MNA) part A score, and MNA part B score.

The outcome labels in the dataset are derived from the categories of the widely used Mini Mental State Exam (MMSE), a cognitive screening tool that assesses multiple cognitive domains.

In the upcoming sections of this study, prediction models will be applied to the dataset, and their performance will be evaluated. The selected models will be suitable for handling the dataset's characteristics and complexity, enabling accurate predictions and insights. The evaluation will be based on appropriate performance metrics, allowing for a comparison of the models' predictive capabilities.

**Data pre-processing:**The dataset underwent several pre-processing steps to ensure its quality and suitability for prediction modeling:

1. Records with missing values were removed since incomplete data cannot be used for prediction modeling.
2. The arguments in the dataset were converted to appropriate data types to enhance the prediction capabilities of the models.
3. Outliers were identified and removed from the dataset to prevent them from skewing the results and negatively impacting the performance of the models.
4. The dataset was normalized to ensure that variables with larger values do not dominate over other variables, promoting a balanced representation in the modeling process.

**Exploratory data analysis:**

During the exploratory data analysis, we employed various visualization techniques to gain insights into the dataset. Histograms were used to examine the distributions of age, body height, and body weight, while bar plots were utilized to analyze gender, education, marital status, and the MMSE binary class.

The histogram of age revealed that the majority of participants are clustered around the age of 70, indicating a concentration of individuals within this age range.

In terms of body height, the histogram illustrated that the highest number of participants had a height around 151 cm, suggesting that this height category was most prevalent among the surveyed individuals.

The histogram of body weight indicated that a weight of approximately 53 kg was the most common among the participants, highlighting a concentration of individuals around this weight value.

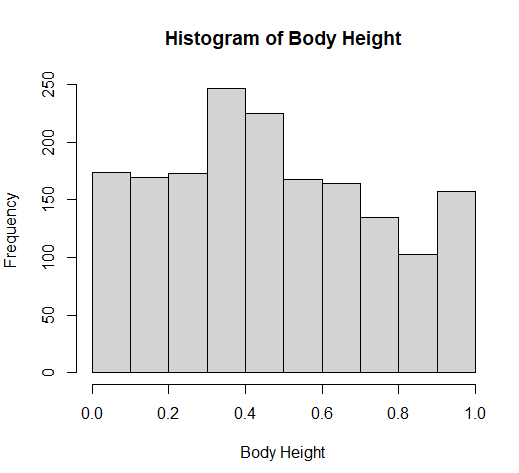
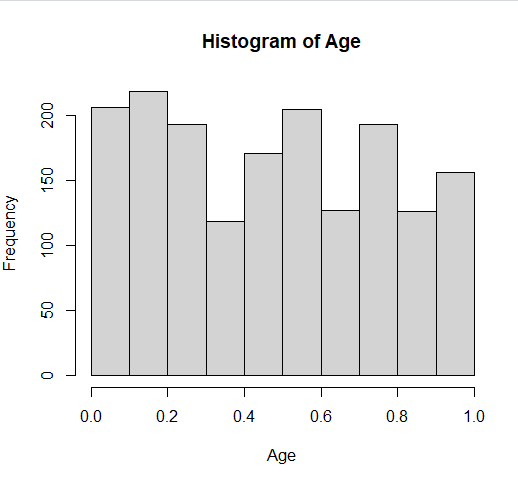
The bar plot of gender displayed an imbalance, with one gender being more prevalent than the other. This imbalance in gender representation could potentially impact the performance of the models and should be taken into consideration during the analysis.

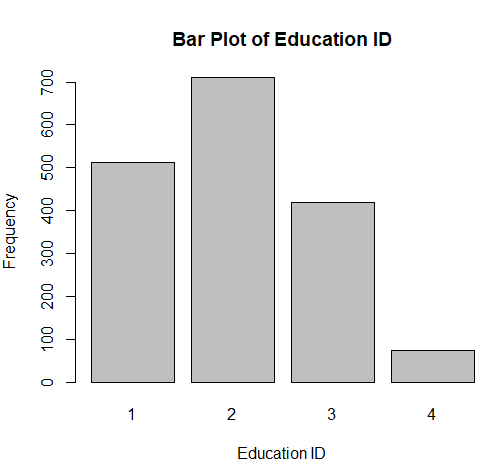
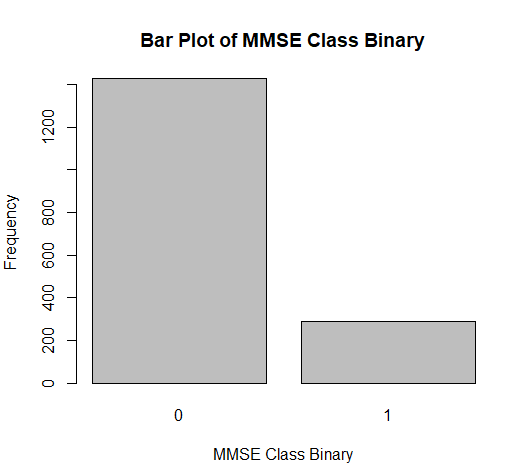
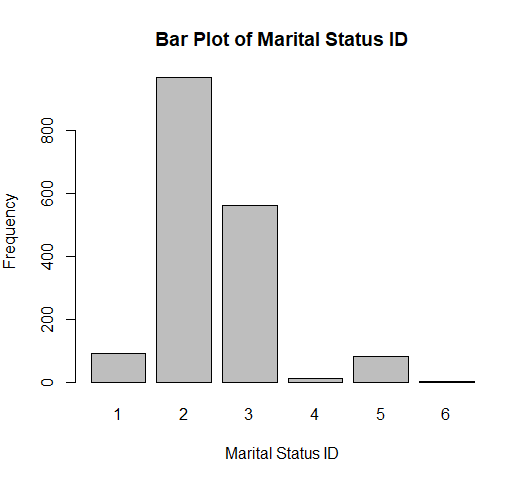
Analyzing the bar plot of education ID, it was observed that education ID 2 had the highest frequency among the participants, while education ID 4 had the lowest frequency. This information provides insights into the educational distribution within the dataset.

The bar plot of marital status ID revealed that marital status ID 2 was the most common category among the participants, indicating a higher representation of individuals with this marital status.

Finally, the bar plot of the MMSE binary class demonstrated significant skewness within the dataset, suggesting an imbalance in the distribution of the two classes. This imbalance should be considered during the modeling process.

Overall, these exploratory data analysis visualizations provided valuable information about the dataset, shedding light on the distributions and characteristics of the variables under investigation.

Following are the plots:

   
**Prediction modelling:**

**Logistic Regression:**

Logistic Regression is a commonly used statistical model for binary classification problems. It models the relationship between the independent variables and the binary outcome variable by estimating the probabilities of belonging to each class. In this case, it can be used to predict the MMSE binary class based on the other variables in the dataset.

Reason for choosing Logistic Regression:

* Logistic Regression is suitable when the outcome variable is binary, which aligns with the MMSE binary class in the dataset.
* It can handle both continuous and categorical independent variables, making it appropriate for this dataset with a mix of different variable types.
* Logistic Regression provides interpretable coefficients, allowing us to understand the impact of each variable on the outcome.

**Random Forest:**

Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to make a final prediction. It is effective for both classification and regression tasks and can handle complex interactions between variables.

Reason for choosing Random Forest:

* Random Forest can handle a mix of continuous and categorical variables, making it suitable for this dataset.
* It can capture nonlinear relationships and interactions between variables, which can be important in predicting the MMSE binary class.
* Random Forest is robust to outliers and can handle missing values.

**Comparison:**

Following tables shows the comparison between the two:

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression** | **Random Forest** |
| **Accuracy** | 0.842105263157895 | 0.83625730994152 |
| **Precision** | 0.891525423728814 | 0.841791044776119 |
| **Recall** | 0.92280701754386 | 0.989473684210526 |
| **F1 Score** | 0.906896551724138 | 0.909677419354839 |

From the above table, it can be seen that, “**Logistic Regression**” is better for such dataset.