Name: Ibrahim Shaglil
Neptune Code: X2MFJZ
Date: 04 October 2024

University of Miskolc
Computer Science Engineering MSc
Data Mining

Data Analysis and Data Mining Assignment

Objective

This report documents the analysis and predictive modeling processes performed on a dataset containing health information of Alzheimer's patients. This assignment analyzes a dataset on Alzheimer’s disease patients to achieve the following objectives as outlined in the guidelines:

1. Data cleaning.
2. Univariate and bivariate analysis.
3. Hypothesis testing.
4. Classification modeling and performance comparison.

Dataset Description

The dataset, sourced from Kaggle, contains health records of 2,149 Alzheimer’s patients. Key attributes include demographics, lifestyle factors, medical history, clinical measurements, cognitive assessments, and behavioral symptoms as documented in the source link.

- **Source**: Kaggle.
- **Description**: Comprehensive patient health information.
- **Purpose**: To study factors associated with Alzheimer’s and develop predictive models.

1. Data Cleaning

- Categorical data (e.g., Gender, Ethnicity, EducationLevel) encoded numerically were mapped to descriptive labels for better interpretability.
- Dropped irrelevant columns (PatientID, DoctorInCharge).
- Most data in the dataset is labeling (0: No, 1:yes) for some columns.

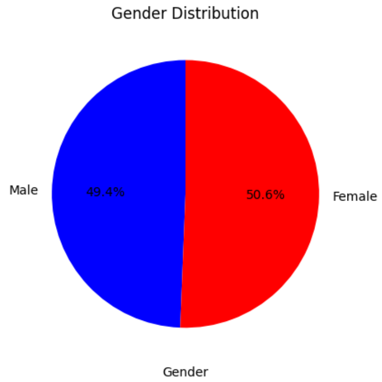
Steps Taken

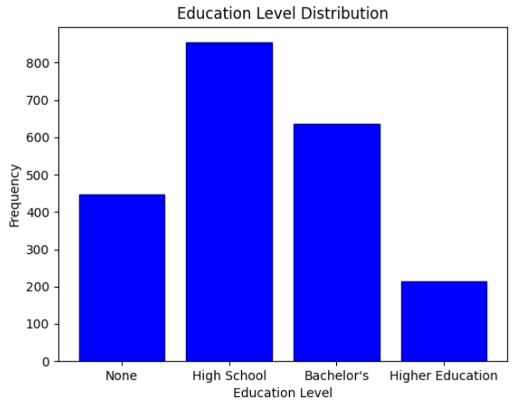
1. Converted Ethnicity (0: Caucasian, 1: African American, 2:Asian, 3:Other), Gender (0: Male, 1: Female), and EducationLevel (0: None, 1: High School, 2:"Bachelor's", 3:"Higher Education") into human-readable labels.
2. Ensured missing values were handled appropriately (not present in the dataset).

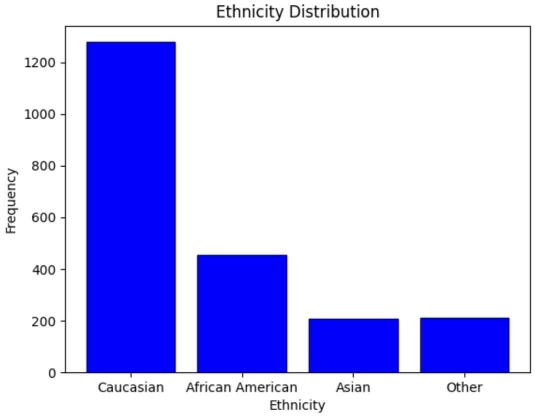
2. Univariate and Bivariate Analysis

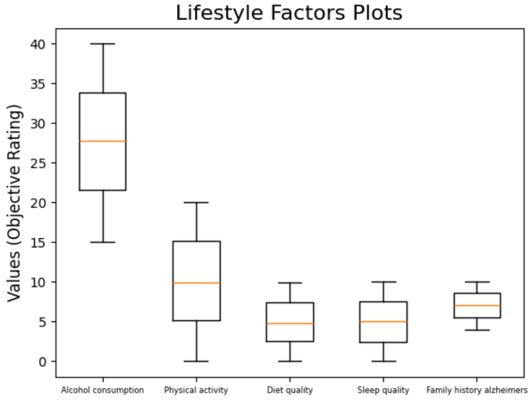
Univariate Analysis

- **Distribution of Attributes**: Visualized Ethnicity, Gender, and Education diversity using bar and pie charts.

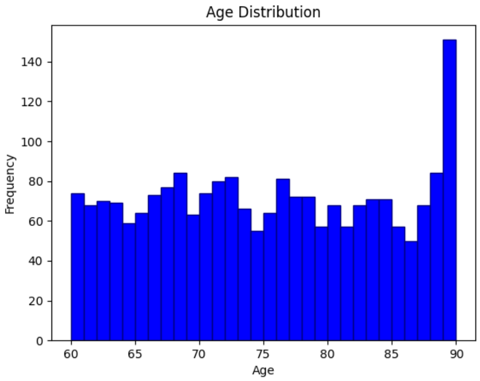






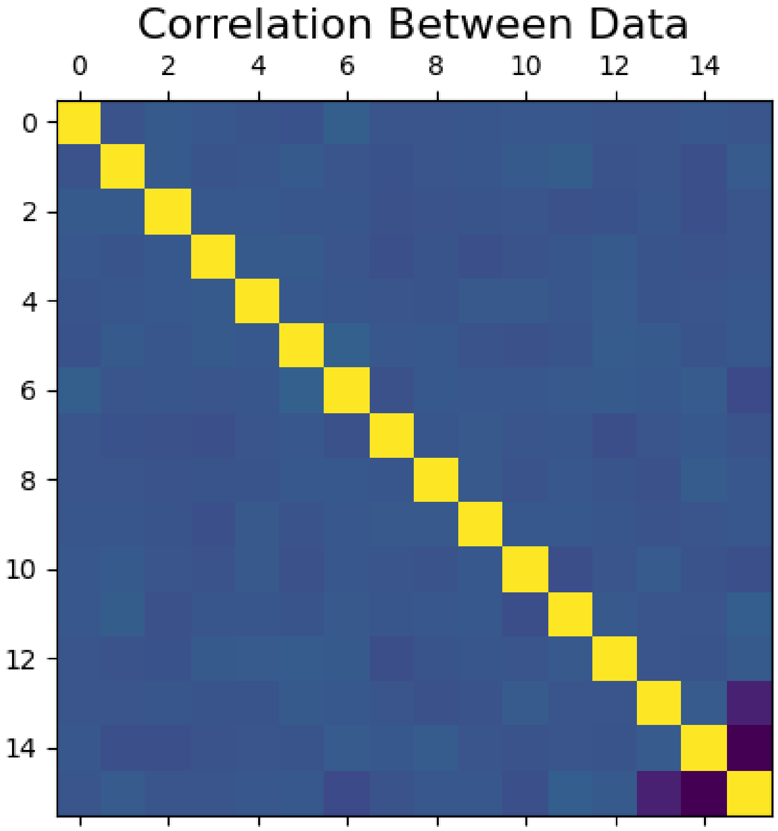


- **Age Analysis**: Histogram revealed the age distribution, focusing on key age groups affected.



Bivariate Analysis

- Explored relationships between lifestyle factors (e.g., DietQuality, PhysicalActivity) and behavioral symptoms (e.g., MemoryComplaints).
- Correlation heatmap identified significant relationships among numerical attributes.
- The Correlation is max (Yellow) and min (Blue), here there is no correlation between numerical values



3. Hypothesis Testing

Hypothesis 1:

Does depression affect diet quality for people in the 60 to 90 age group?

- **Hypothesis**:
  - H0: Diet Quality is the same for people with and without depression.
  - H1: Diet Quality is different for people with and without depression.
- **Methodology**:
  - Kolmogorov-Smirnov tests assessed normality for both groups (with/without depression).
  - A T-test evaluated differences in Diet Quality.

- **Results**:
  - The sample size is 30 people, and the values are normally distributed.
  - Conclusion: diet quality did not change for both groups, P-value = 0.927.

Hypothesis 2:

Does education level affect Mini-Mental State Examination (MMSE) scores?

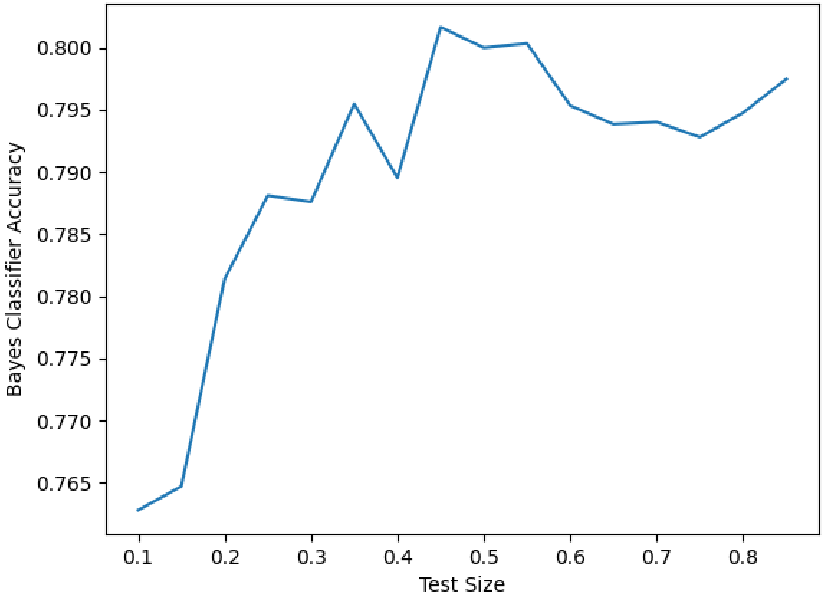
- **Hypothesis**:
  - H0: MMSE is the same for people with and without higher education.
  - H1: MMSE is different for people with and without higher education.
- **Methodology**:
  - Compared MMSE scores between patients with higher education (Bachelor's, Higher Education) and those with none or high school education.
  - Normality tested via Kolmogorov-Smirnov tests.
  - T-test evaluated group differences.
- **Results**:
  - The sample size is 30 people, and the values are normally distributed.
  - Conclusion: Education Level does not affect MMSE, P-Value= 0.293.

4. Classification Modeling

Two classification models were implemented using lifestyle factors to predict memory complaints:

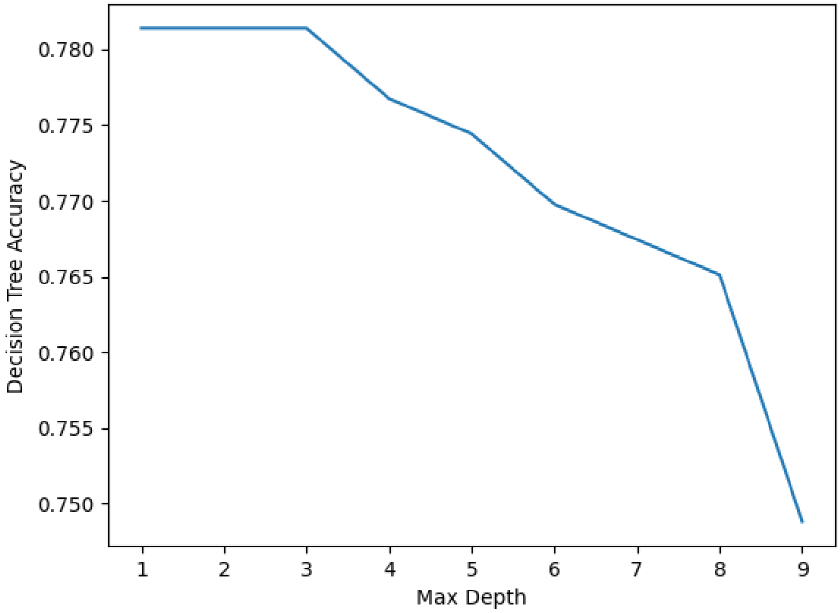
Model 1: Naive Bayes Classifier

- **Features**: BMI, Smoking, AlcoholConsumption, PhysicalActivity, DietQuality, SleepQuality.
- **Optimization**: Tuned optimum test data ratio for maximum accuracy.
- **Performance**:
  - Accuracy: The accuracy of the Naive Bayes is 80% at 45% in test data.



Model 2: Decision Tree Classifier

- **Features**: Same as above.
- **Optimization**: Tuned maximum tree depth for best accuracy.
- **Performance**:
  - Accuracy: The accuracy of the Decision Tree is 78% at 1 tree depth.



5. Prediction

Using both models, a prediction was made with the following sample inputs:

- **Inputs**: BMI = 25, Smoking = 7, Alcohol Consumption = 5, Physical Activity = 1, Diet Quality = 1, Sleep Quality = 7.
- **Prediction**:
  - Naive Bayes: The patient is likely to have memory complaints.
  - Decision Tree: The patient is not likely to have memory complaints.

Conclusion

This study provided insights into Alzheimer’s-related health factors and demonstrated the feasibility of predictive modeling for memory complaints. Future work could explore deep learning methods or ensemble models for improved accuracy.