MAKERERE UNIVERSITY

Master of Science in Computer Science

Department of Computer Science

College of Computing and Information Sciences, Makerere University

MACHINE LEARNING

ASSIGNMENT 001

By

RHODA DIANA NDIBALEKERA

Reg No: 2024/HD05/21951U

Student No: 2400721951

17th -September-2024

Dataset Title

National Health and Nutrition Health Survey 2013-2014 (NHANES) Age prediction Subset

Summary

NHANES was carried out by the Centers for Disease Control and Prevention (CDC), which collects extensive health and nutritional information from diverse U.S. population. The dataset focused on predicting the respondents' age by extracting a subset of features from the larger NHANES dataset. The selected features included are listed in the variable's table below with their descriptions

Variables Table

Variable Name	**Role**	**Type**	**Demographic**	**Description**	**Units**	**Missing Values**
SEQN	ID	Continuous		Respondent Sequence Number		no
age_group	Target	Categorical	Age	Respondent's Age Group (senior/non-senior)		no
RIDAGEYR	Other	Continuous	Age	Respondent's Age		no
RIAGENDR	Feature	Continuous	Gender	Respondent's Gender		no
PAQ605	Feature	Continuous		If the respondent engages in moderate or vigorous-intensity sports, fitness, or recreational activities in the typical week		no
BMXBMI	Feature	Continuous		Respondent's Body Mass Index		no
LBXGLU	Feature	Continuous		Respondent's Blood Glucose after fasting		no
DIQ010	Feature	Continuous		If the Respondent is diabetic		no
LBXGLT	Feature	Continuous		Respondent's Oral		no
LBXIN	Feature	Continuous		Respondent's Blood Insulin Levels		no

Note: The target of the dataset was age_group

Questions

1. Can age be predicted using health and Nutrition data from the dataset?
2. What is the relationship between lifestyle choices and age?
3. How does respondents' gender and body mass index correlate with age_group
4. Can LBXIN, body mass index and PAQ605 be strong predictors of age?

Data Wrangling

The dataset was loaded in Google Colab for analysis and the initial steps involved checking for duplicates, missing values and converting data types where necessary. i.e. there was only one categorical data which also served as our dataset target data and thus, there was need for its conversion to numerical data to facilitate the analysis.

The dataset was found to be clean with zero missing values and all features were deemed relevant for the analysis.

Exploratory Data Analysis (EDA)

Visualizations were created to explore the dataset. Histograms were used to understand the distribution of variables, scatter plots to examine relationships between features and the target variable (age), and box plots to identify outliers. A correlation matrix was also generated to identify strong correlations between features and age.

Conclusion

The EDA revealed several interesting patterns. For instance, higher body mass index (BMI) and blood glucose levels were found to be positively correlated with age. Additionally, lifestyle choices such as engaging in physical activities showed a negative correlation with age. These insights suggest that physiological measurements and lifestyle choices are significant predictors of age.

Findings

In summary, the EDA provided valuable insights into the relationships between health and nutrition variables and age. The analysis confirmed that certain physiological measurements and lifestyle choices are strong predictors of age. These findings can inform future research and help in developing predictive models for age based on health data. However, the analysis is limited by the scope of the dataset and the need for further validation.

The code for this analysis is available on GitHub: https://github.com/DianaNdibalekera1/Machine-Learning.git