Prediction of Diabetes-Related Factors through NHANES Data Analysis

Faculty of Sport Sciences
5022A044-3 Misuzu Shimoyama
Research supervisor：Katsuhiko Suzuki

【Introduction】

Millions around the world are impacted by diabetes, a major health issue on a global scale. Its importance lies in its widespread impact and the serious complications it can lead to, such as heart disease, stroke, kidney failure, etc. Understanding the features and behaviors that lead to the onset of diabetes is crucial for several reasons, such as awareness of these risk factors can help in early detection and prevention. Recently, artificial intelligence (AI) techniques have become very popular in the analysis of data sets and predictive modeling in several domains, especially health industries.

【Method】

To study the diabetes-related factors, we made use of the national health and nutrition examination survey (NHANES) dataset. The NHANES dataset is a comprehensive dataset that covers a broad cross-section of the United States population. This survey dataset has been available online since 1999 and includes individuals from diverse ethnic, racial, and age groups, providing a more generalized overview of health, disease, and nutritional states, including diabetes. We began by collecting the dataset from the NHANES website and pre-processing it to prepare structured data for machine learning algorithms. In this step, we considered various survey data, such as demographic data, clinical laboratory data, questionnaires, etc. The NHANES dataset, itself, does not have details if a person is diabetic or not. Therefore, using previous survey data on diabetes and lab blood glucose level data, we classified the respondents into diabetic or non-diabetic. We prepared well-structured comma separated values (CSV) data from the original SAS transport format (XPT) survey data by merging various survey data, such as demographic data, clinical laboratory data, etc. of the respondents. In total, the training and validation data consisted of 1,347 and 337 respondents, respectively.

For inputting the data to the models, such as decision trees, or neural networks, we did pre-processing, such as filling in missing data using mean imputation for continuous variables. For categorical variables, we appropriately imputed the data, such as null values to 0. We also encoded categorical variables using one-hot-encoding so that machine learning models were unbiased towards any category. Finally, we used the processed dataset to fit a decision tree model. Using the decision tree model, we studied its formation based on the Gini coefficient values to understand which factors are most important that contribute to diabetes in respondents. After considering the important factors using decision trees, we built a diabetes predictive model using a neural network consisting of 32 neurons in the hidden layer for the calculation of features from the input data and one neural with sigmoid activation for the estimation of probability if the respondent in the NHANES dataset is diabetic or not. We finally evaluated the performance of the model using various metrics, such as accuracy, recall, and precision.

【Results】

Analyzing the construction of the decision tree and observing important features based on the Ginni value and feature importance, we found variables such as age, body mass index (BMI), gender, waist size, diabetes in a family member, and education level to be of prime importance compared to others, such as household income, ethnicity, etc. Lifestyle factors, notably diet and physical activity, were identified as crucial in predicting the likelihood of developing the condition.

We trained the model using the 80% of the original processed data consisting of 1,347 respondents and then evaluated its performance on the holdout 20% validation set consisting of 337 respondents. On the validation set, we obtained an overall accuracy of up to 84.60%. We noticed that the accuracy of the deep learning model to classify diabetes increased as we increased the number of input features in the network.

【Discussion】

In the classification of diabetes using a decision tree, age, waist size, BMI, weight, and height were key factors due to their strong associations with insulin resistance and metabolic health. Age was linked to decreased insulin sensitivity, while waist size indicated central obesity, a major risk factor. BMI and weight are direct measures of obesity, which greatly influence the risk of developing type 2 diabetes. Height, though a less direct factor, can be associated with early-life nutritional factors affecting metabolic health. These variables collectively reflected the physiological and lifestyle factors critical in assessing the likelihood of diabetes.

For the predictive neural network model, the high accuracy is due to a limited number of diabetic samples in the dataset because of which the model easily classifies non-diabetic cases leading to a high accuracy. For improvement of classification of diabetic respondents, it could be worth looking into sampling techniques and ensemble models.

【Conclusion】

In conclusion, this study highlighted the critical role of AI in understanding and predicting diabetes. It demonstrated the significance of using a comprehensive dataset like NHANES to identify key factors associated with diabetes and provided a deep learning model that achieved promising accuracy in diabetes classification. This research contributes to the field of predictive healthcare and opens doors for further research to enhance diabetes prevention and management strategies.