Title: Prediction Model for Obesity Classification

Date: 2024-04-26

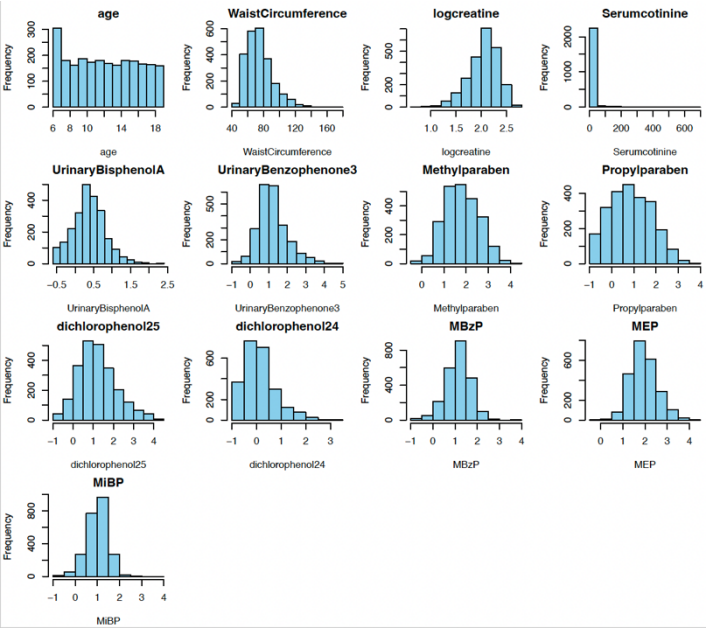
Author: Seowoo Kim

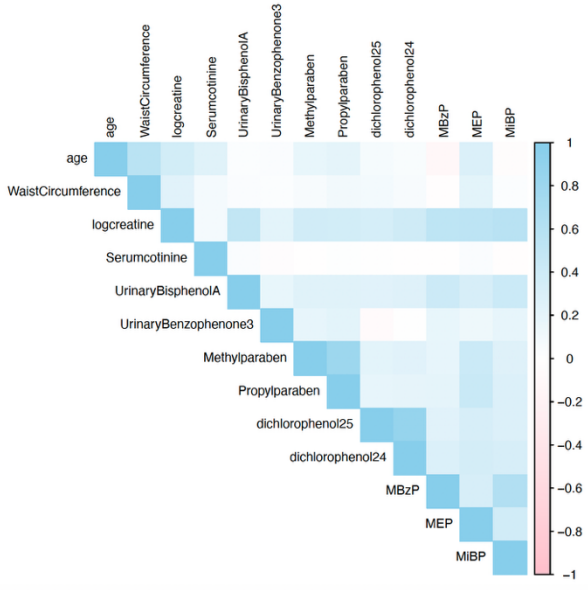
Introduction:

Obesity is a multifactorial condition influenced not only by biological determinants such as body mass index (BMI) or cholesterol levels but also by behavioral, socioeconomic, and environmental factors. Identifying predictors of obesity beyond traditional biomarkers can provide valuable insights into preventive strategies and targeted interventions. In this study, I developed and compared predictive models for obesity classification using a dataset that integrated demographic, anthropometric, clinical, lifestyle, and family history information.

Data:

The dataset consisted of 2,372 samples with a mixture of continuous and categorical variables reflecting a broad range of obesity-related risk factors. Continuous predictors included demographic information such as age, anthropometric measures like BMI z-scores and waist circumference, and several urinary biomarkers (e.g., Bisphenol A, Benzophenone-3, Parabens, and Phthalate metabolites). Most anthropometric measures followed approximately normal distributions, whereas the biomarker concentrations displayed right-skewed patterns, with certain variables such as serum cotinine exhibiting extreme outliers. Categorical predictors captured sociodemographic characteristics (gender, race, education, income), lifestyle behaviors (smoking status, dietary energy intake, exercise, sleep duration), and family history (parental obesity and diabetes). The outcome variable, obesity, was coded as a binary yes/no response, with an overall prevalence of about 20%.





Exploratory data analysis revealed that while most predictors had weak to moderate pairwise correlations, some expected relationships were present. To prepare for modeling, categorical predictors were dummy encoded, resulting in 38 total features. Finally, the dataset was randomly divided into training (80%) and testing (20%) subsets to enable unbiased evaluation of predictive performance.

Method:

Before model training, we performed preprocessing and exploratory data analysis (EDA) to ensure data quality and to better understand the characteristics of the dataset. The outcome variable, obesity, had a prevalence of about 20%, indicating moderate class imbalance but not severe enough to require resampling. Continuous predictors were inspected for skewness, with cholesterol and glucose showing right-skewed distributions, while variables such as blood pressure were more symmetric. To place variables on a comparable scale, continuous predictors were standardized, and categorical predictors were dummy encoded, which expanded the dataset to a total of 38 predictors. Missing values were minimal and addressed during preprocessing.

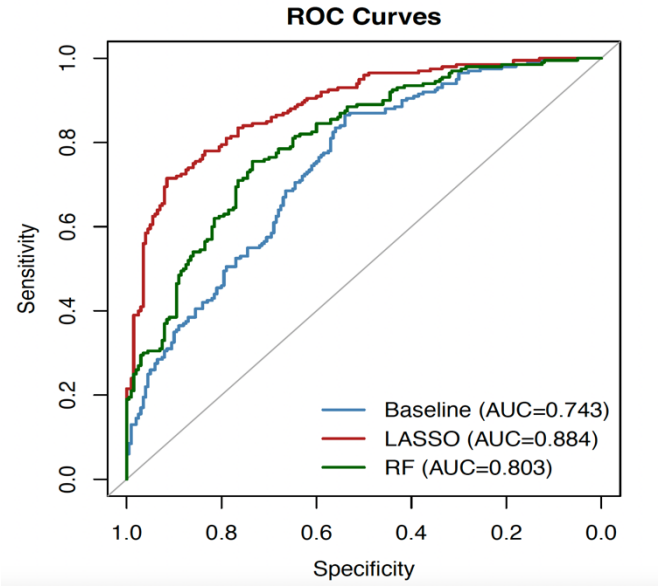
For the baseline logistic regression, we conducted diagnostic checks to ensure model stability. Specifically, we calculated the Variance Inflation Factor (VIF) to identify potential multicollinearity among predictors, as high collinearity can distort coefficient estimates. We also examined influential observations using residual diagnostics, since logistic regression can be sensitive to outliers or high-leverage points. These steps helped establish a reliable benchmark model.

Next, we implemented Lasso logistic regression, which applied an L1 penalty to shrink coefficients and automatically eliminate irrelevant or redundant predictors. Unlike the baseline model, Lasso is less vulnerable to multicollinearity, as the penalty encourages sparsity and improves generalization. This approach reduced the predictor set from 38 to 18 variables, improving interpretability without sacrificing predictive performance.

In addition, we trained Random Forest models, an ensemble method capable of capturing nonlinear relationships and complex interactions between predictors. The Random Forest was first trained using all predictors. We then calculated feature importance scores and retrained the model with the top 10, 15, 20, and 25 predictors to evaluate the impact of dimensionality reduction on predictive accuracy.

Model performance across all approaches was evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metric. ROC-AUC was chosen because it is robust to moderate class imbalance and provides an interpretable summary of classification performance across thresholds.

Results:



Model evaluation was carried out using ROC curves and the AUC metric, which provided a clear comparison of predictive performance across different approaches. Diagnostic checks confirmed that no predictors exhibited severe multicollinearity, suggesting that the dataset was stable enough for regression-based modeling. The baseline logistic regression model, trained with all 37 predictors, achieved an ROC-AUC of 0.743. While this indicated a reasonable ability to distinguish between obese and non-obese individuals, further inspection revealed that many predictors contributed little to overall accuracy. This suggested that although the full model provided a starting point, it lacked efficiency and interpretability. The Lasso logistic regression model substantially improved performance, reducing the predictor set to 18 variables while achieving the highest ROC-AUC of 0.884. The strong performance of Lasso highlights the benefits of penalization and automatic feature selection: by shrinking or eliminating irrelevant predictors, the model avoided overfitting and focused on the variables most strongly associated with obesity. This not only enhanced classification accuracy but also simplified interpretation by providing a more concise set of risk factors. The Random Forest model also outperformed the baseline logistic regression, with an ROC-AUC of 0.803. Its ability to capture nonlinear relationships and interactions between variables proved advantageous, especially when trained with all predictors. Further experiments showed that accuracy improved when the model was retrained with the top ~20 most important features, reflecting the importance of dimensionality reduction in ensemble methods as well. Beyond this point, additional pruning of predictors did not yield further gains. The ROC curve presented corresponds to the Random Forest trained on these top 20 features.

Taken together, the results demonstrate that while both logistic regression and Random Forest models produced reasonable performance with the full feature set, Lasso logistic regression consistently delivered the best results, balancing predictive accuracy with model simplicity**.** This underscores the value of incorporating regularization techniques in predictive modeling, particularly when dealing with datasets that combine a large number of demographic, clinical, and behavioral predictors.

Discussion:

These results demonstrate that feature selection plays a critical role in building predictive models for obesity. Lasso regression delivered the strongest performance, reducing the predictor set to 18 variables while achieving the highest ROC-AUC, and Random Forest models also benefited from dimensionality reduction, with accuracy peaking around the top 20 predictors. Taken together, these findings highlight that removing noisy or redundant variables improves both model accuracy and interpretability. From a practical perspective, the results suggest that focusing on a smaller subset of influential predictors can help design cost-efficient and interpretable screening tools to identify individuals at higher risk of obesity. This enables clinicians and policymakers to prioritize key lifestyle and biomarker factors for early intervention and targeted prevention strategies.