Abstract

Type 2 diabetes is a rising problem among the California population, with nearly half of all adults in California being pre-diabetic or undiagnosed. The purpose of this study is to determine what factors may contribute to the presence of type 2 diabetes, and whether the risk of type 2 diabetes can be predicted using these factors. To achieve this, we analyzed public survey data from the California Health Interview Survey (CHIS) - a statewide health survey conducted yearly by the UCLA Center for Health Policy Research. Participants were aged 18 or older, and important demographics and health topics, such as marital status, education, and exercise tendencies, were recorded. The data analyzed spans over the course of 9 years, from 2012 to 2020 and the combined dataset contains about 190,000 surveyees and 785 covariates. After performing feature selection using random forest classifiers, the number of relevant variables for modeling was reduced to 245 covariates. The chosen variables were fitted using KNN, stochastic gradient descent, decision trees, random forests, and XGBoost. To account for the class imbalance of the dependent variable, we implemented cost-sensitive learning and SMOTE to each model. The accuracy of each model was evaluated using its F1 and AUC scores. The results indicate the XGBoost model had the highest scores. This model performed with a F1 score of 0.897 and an AUC score of 0.874. Given the results, the model demonstrated a strong capability of accurately predicting the risk of type 2 diabetes in California adults and the fitted variables may be significantly tied to affecting the risk. Therefore, it can be valuable for an individual to monitor these factors. In conclusion, the most significant factors for predicting the risk of type 2 diabetes is BMI, high blood pressure, and if they have prediabetes.