NUTRI QUÉBEC .COM

PREDICTING VEGETABLE AND FRUIT CONSUMPTION IN DISTINCT SOCIOECONOMIC GROUPS USING MACHINE LEARNING MODELS

Mélina Côté, Catherine Laramée, Annie Lapointe, Simone Lemieux, Sophie Desroches, Ariane Bélanger-Gravel, Benoît Lamarche

Purpose: Recruiting vulnerable populations represents a considerable challenge in public health research. Developing and validating predictive models of nutrition-related outcomes using data from broader population samples for use in hard-to-reach low socioeconomic groups would be valuable. This study aimed to assess the applicability in different socioeconomic subgroups of a machine learning (ML) model developed to predict adequate vegetable and fruit consumption (VFC) in a general population.

Methods: Data from a large array of variables (96) potentially associated with dietary habits in a sample of 2836 adults (86% women) from the NutriQuébec project were used. Adequate VFC (≥5 servings/d) was measured by averaging data from two to three web-based 24-h dietary recalls and used as the outcome to predict. The sample was randomly divided into training (60%), testing (20%) and validation (20%) sets. The training set (general population) was used to develop a Random Forest (RF) classification algorithm to predict adequate VFC. The prediction performance of the model was evaluated in the testing set (general population) using the accuracy score (proportion of correct predictions). The validation set was divided into low (household income < 50 000$CAN/y, n=130), middle (50 000$CAN to < 100 000$CAN/y, n=204) and high (≥100 000$CAN/y, n=230) income groups. The accuracy of the model was then evaluated in each income subgroup. The analysis was repeated by dividing the validation set into low (high school or less), middle (pre-university or certificate), and high (bachelor or higher) education groups.

Results: The model developed in the training set predicted adequate VFC with an accuracy of 0.60 (95%CI 0.56-0.64) in the testing set. The model predicted adequate VFC in low, middle and high income subgroups with accuracies of 0.65 (95%CI 0.60-0.70), 0.65 (95%CI 0.62-0.68) and 0.55 (95%CI 0.52-0.58), respectively. Relatively similar results were obtained in the education subgroups.

Conclusion: The RF model predicted adequate VFC in the low income subgroup with similar accuracy as in the total population, in which the model was developed and tested. These results suggest that ML models predicting a nutrition-related outcome in a general population can be used to predict a similar outcome in harder-to-reach socioeconomic groups.