ABSTRACT

THESIS: Subgroup Identification with Virtual Twins and GUIDE Algorithms - an Application to Adult Fitness Data

STUDENT: Mst Sharmin Akter Sumy

DEGREE: Master of Science

COLLEGE: Science and Humanities

DATE: July 2020

PAGES: 54

Cardiorespiratory fitness (CRF) is often used as an indicator of physical fitness that has an inverse association with mortality. It is practical to identify subgroups where CRF exerts positive or negative association with mortality status than on the entire population. We considered Classification and Regression tree-based algorithms: GUIDE and Virtual Twins (VT) to find subgroups of participants where CRF exerts positive or negative association with mortality status from the Ball State Adult Fitness Program Longitudinal Lifestyle Study data. Results are compared with traditional penalized Logistic regression with LASSO penalty. The results indicate that VT for classification and regression tree identified subgroups by considering BMI (Body Mass Index) and age as important predictors for different thresholds. We observed that when the threshold was larger than some specific value, VT couldn't identify any subgroups. GUIDE selected subgroups with BMI, age, and sex as important predictors. Finally, Logistic regression with LASSO penalty found the interaction with fitness rank and other predictors: age, sex, smoking status, and hypertension. Thus, our results suggest that tree-based methods identified fewer predictors compared to the non-tree based logistic regression method.