Creation of a Machine Learning App to Facilitate Pancreatic Cancer Prediction

Fleishman, Reid (School: William A. Shine Great Neck South High School)

While pancreatic cancer is a deadly disease that is difficult to diagnose at an early stage, early diagnosis yields a much higher survival rate, making it key for treatment. With an increase in data, machine learning has shown promise in facilitating the early diagnosis of such diseases. The goals of this study were to develop an effective machine learning model to predict one's risk of developing pancreatic cancer and to create a mobile app to facilitate prediction by an end-user. The training dataset was derived from the Integrated Public Use Microdata Series (IPUMS) health surveys and the National Cancer Institute's Prostate, Lung, Colorectal and Ovarian (PLCO) study, and consisted of 17 pancreatic cancer risk factors encompassing 752,527 patients (983 of whom had pancreatic cancer). Data were re-coded, normalized, and split into training, validation, and testing sets. Measures were taken to balance the classes and handle missing values. Decision Tree (DT), Random Forest (RF), Boosted Trees (BT), Logistic Regression (LR), and Support Vector Machine (SVM) models were created. The BT model achieved a sensitivity of 0.788, a specificity of 0.791, and the highest Area Under the Receiver Operating Characteristic Curve (AUC) of 0.870 out of the five algorithms and thus was embedded within the mobile app. Moreover, age, physical activity, smoking, and race were important predictors of pancreatic cancer, whereas depression, among others, were not. In this study, machine learning models were created that could be the basis for an important diagnostic tool to help identify one's risk of developing pancreatic cancer.