Capstone project 1 proposal

Predicting Cervical Cancer with Machine Learning

Yi Li

Cervical cancer is both the fourth commonly diagnosed cancer and the fourth most common cause of death from cancer in women. In 2018, approximately 311 000 women died from cervical cancer, more than 85% of these deaths occurring in low- and middle-income countries.

Molecular and epidemiological studies have demonstrated that infection by human papillomavirus (HPV) is the most important risk factor for cervical cancer. Smoking, earlier sexual debut, younger age at first pregnancy, high parity, long-term use of oral contraceptives have also been confirmed as risk factors cervical cancer.

Preventive strategies state that women at high risk should be screened earlier and frequently. Hence building a risk prediction model to identify women most likely to develop cervical cancer will facilitate the screening. This will help the Centers for Disease Control and Prevention (CDC) to manage the screening better, by setting up a better screening strategy and routine according to the risk, especially for low resource countries.

I am going to use the cervical cancer (Risk Factor) data set from The UCI Machine Learning Repository to do the prediction. This dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information (such as age), habits, and historic medical records of 858 patients, including some indicators, risk factors of cervical cancer as well as the diagnosis.

Link: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

The deliverables will be code and a PowerPoint slide.