Cervical cancer risk classification

JiEun Song

Overview

Cervical cancer is the fourth most common type of cancer among women in the world. More than 10,000 new cases are diagnosed each year in the US. Although each year cervical cancer kills about 4,000 women in the US, it is the most preventable type of cancer. Cervical cancer is highly preventable because many risk factors are known for initiation of the cancer. My aim is to examine the dataset, and use a classification model to predict whether a patient is diagnosed with cervical cancer or not.

I found the dataset on Kaggle, which has information of 850 patients with 37 columns with risk factors. Each patients are classified as cancer positive or negative.

Features

Number of partners
First sexual intercourse
Number of pregnancies
Smokes
Smokes (years)
Smokes (packs/year)
Hormonal contraceptives
Hormonal contraceptives (years)
IUD
IUD (years)
STDs
STDs (number)
STDs:condylomatosis
STDs:cervical condylomatosis
STDs:vaginal condylomatosis
STDs:vulvo-perineal condylomatosis
STDs:syphilis
STDs:pelvic inflammatory disease
STDs:genical herpes
STDs:molluscum contagiosum
STDs:AIDS
STDs:HIV
STDs:Hepatitis B
STDs:HPV
STDs:number of diagnosis
STDs:time since first diagnosis
STDs:time since last diagnosis

Known unknowns

I have studied cancer biology, especially brain cancer, and brain cancer is distinct from other cancers, so I am not familiar with cervical cancer. I know that HPV infection is a major cause of cervical cancer, but am not sure prediction power of each features.