Capstone Project 1 Milestone Report
Predicting Cervical Cancer with Machine Learning

Yi Li

Introduction

Cervical cancer is both the fourth commonly diagnosed cancer and the fourth most common cause of death from cancer in women (Fig. 1). In 2018, approximately 311 000 women died from cervical cancer, more than 85% of these deaths occurring in low- and middle-income countries.

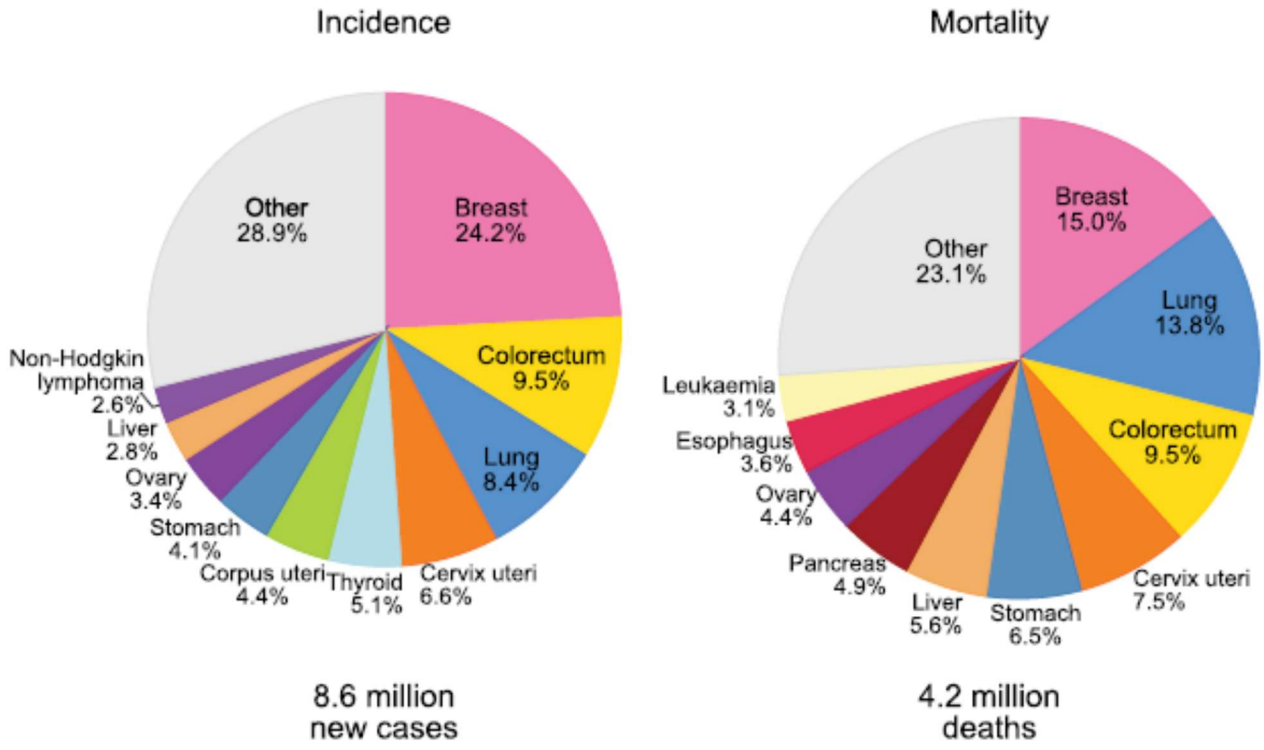

FIGURE 1. Pie Charts Present the Distribution of Cases and Deaths for the 10 Most Common Cancers in 2018 for Females. The area of the pie chart reflects the proportion of the total number of cases or deaths. (CAAC paper)

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 of 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.

In this project, I am going to use the cervical cancer (Risk Factor) dataset 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.

Data Wrangling

After importing the cervical cancer (Risk Factor) dataset as a pandas DataFrame, I found all the missing values are represented by "?". So I converted all the values with type of object to numeric values, so the missing values are represented by "NA".

The columns with numeric data are as follows, "Number of sexual partners", "First sexual intercourse", "Num of pregnancies", "Smokes (years)", "Smokes (packs/year)", "Hormonal Contraceptives (years)", "IUD (years)" and 'STDs (number)". For these columns, I filled the missing values with the median of the columns.

The columns with categorical data are as follows, "Smokes", "Hormonal Contraceptives", "IUD", "STD", "STDs: condylomatosis", "STDs: cervical condylomatosis", "STDs: vaginal condylomatosis", "STDs: vulvo -perineal condylomatosis", "STDs: syphilis", "STDs: pelvic inflammatory disease", "STDs: genital herpes", "STDs: molluscum contagiosum", "STDs: AIDS", "STDs: HIV", "STDs: Hepatitis B" and "STDs: HPV". For these columns, I filled missing values with the mode of the columns.

This DataFrame contains two columns regarding to HPV infection, “STDs: HPV” and “Dx: HPV”. “Dx: HPV has more precise diagnosis result of HPV infection with no missing data, so I excluded the “STDs: HPV” column.

The columns “STDs: Time since first diagnosis” and “STDs: Time since last diagnosis” are also excluded, since there are too many missing data in these two columns.

Exploratory Data Analysis and Statistical Inference

The dataset provide the results of four diagnostic methods, “Hinselmann”, “Schiller”, “Cytology” and “Biopsy”. Since the gold standard for cervical cancer diagnosis is usually biopsy, I choose “Biopsy” as the target, and exclude the results of the other three methods (columns “Hinselmann”, “Schiller” and “Cytology”).

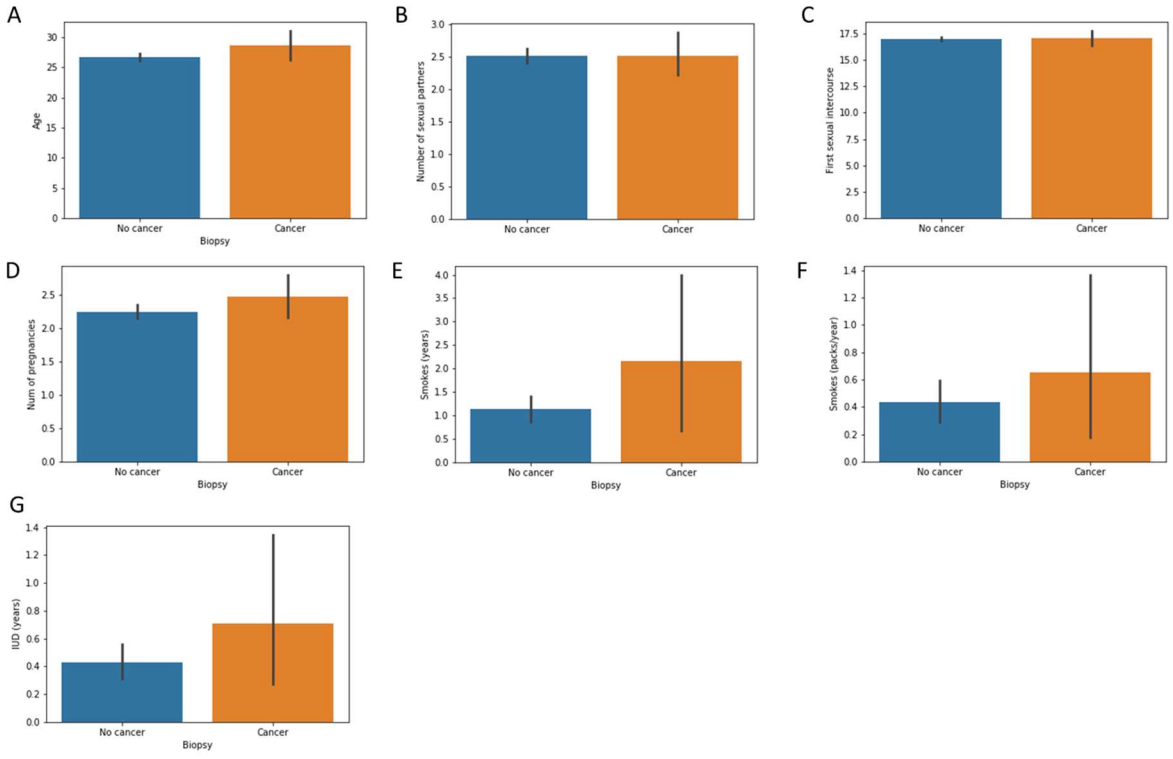

FIGURE 2. Bar plots present the difference in mean of numeric features for people with or without cervical cancer. (A) Age. (B) Number of sexual partners. (C) First sexual intercourse. (D) Number of pregnancies. (E) Smokes (years). (F) Smokes (packs/year). (G) IUD (years).

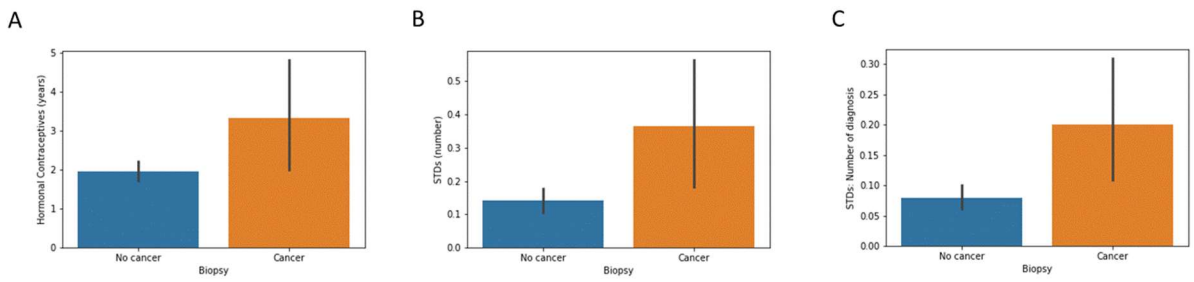

FIGURE 3. Bar plots present the difference in mean of numeric features for people with or without cervical cancer. (A) Hormonal Contraceptives (years). (B) STDs (numbers). (C) STDs: Number of diagnosis.

To test which features are correlated to the target variable "Biopsy", I employ t-test to compare the mean between people with ("Biopsy" = 1) and without ("Biopsy" = 0) cancer for the features with numeric data. The results showed that there is no significant difference in "Age" ($p = 0.1$), "Number of sexual partners" ($p=0.99$), "First sexual intercourse" ($p=0.83$) and "Num of pregnancies" ($p=0.24$) between people with and without cancer. (Fig. 2) The above results indicated that these features are not factors for cervical cancer for this dataset.

For some features, such as "Smokes (years)" ($p=0.073$), "Smokes (packs/year)" ($p=0.47$) and "IUD (years)" (IUD, intrauterine device) ($p=0.26$), the results do not show statistical significant difference, since the variance is too big. (Fig. 2)

The statistical analysis results show significant difference ($p < 0.05$) between people with and without cancer for features "Hormonal Contraceptives (years)", "STDs (number)" (STDs, sexually transmitted diseases) and "STDs: Number of diagnosis". These suggest that these features are correlated with the target variable. (Fig. 3)

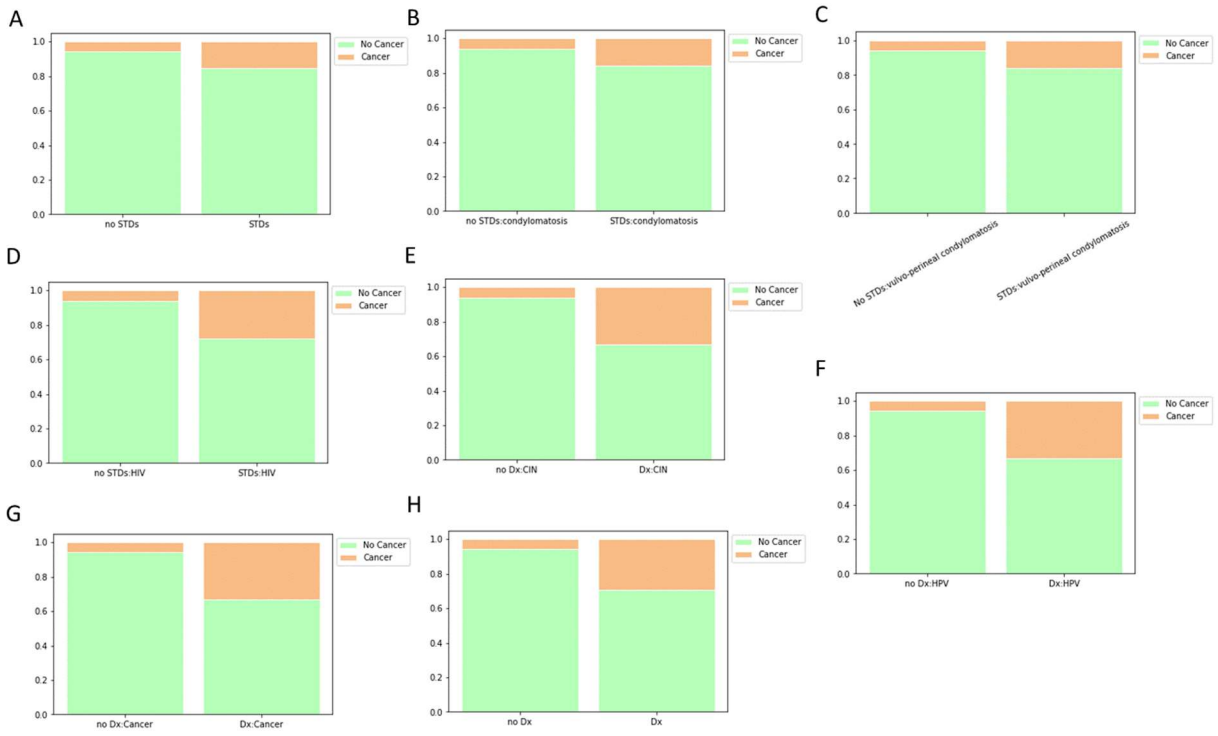

FIGURE. 4 Stacked bar plots compare the proportion of people with or without cancer for different categorical features. (A) STDs. (B) STDs: condylomatosis. (C) STDs: vulvo-perineal condylomatosis. (D) STDs: HIV. (E) Dx: CIN. (F)"Dx: HPV. (G) Dx: Cancer. (H) Dx.

I use chi-squared test to test to whether there is a significant relationship between features with categorical data and the target variable. A contingency table is made for each feature and the target variable. And chi-squared test is performed based on the contingency table. The results show that features "STDs", "STDs: condylomatosis"，"STDs: vulvo-perineal condylomatosis"，"STDs: HIV", "Dx: Cancer", "Dx: CIN", "Dx: HPV" and "Dx" are correlated with the target variable ($P<0.05$). For feature “STDs: genital herpes”, there is only one case with genital herpes, and this person also got cervical cancer.