Capstone Project 1 Final 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.

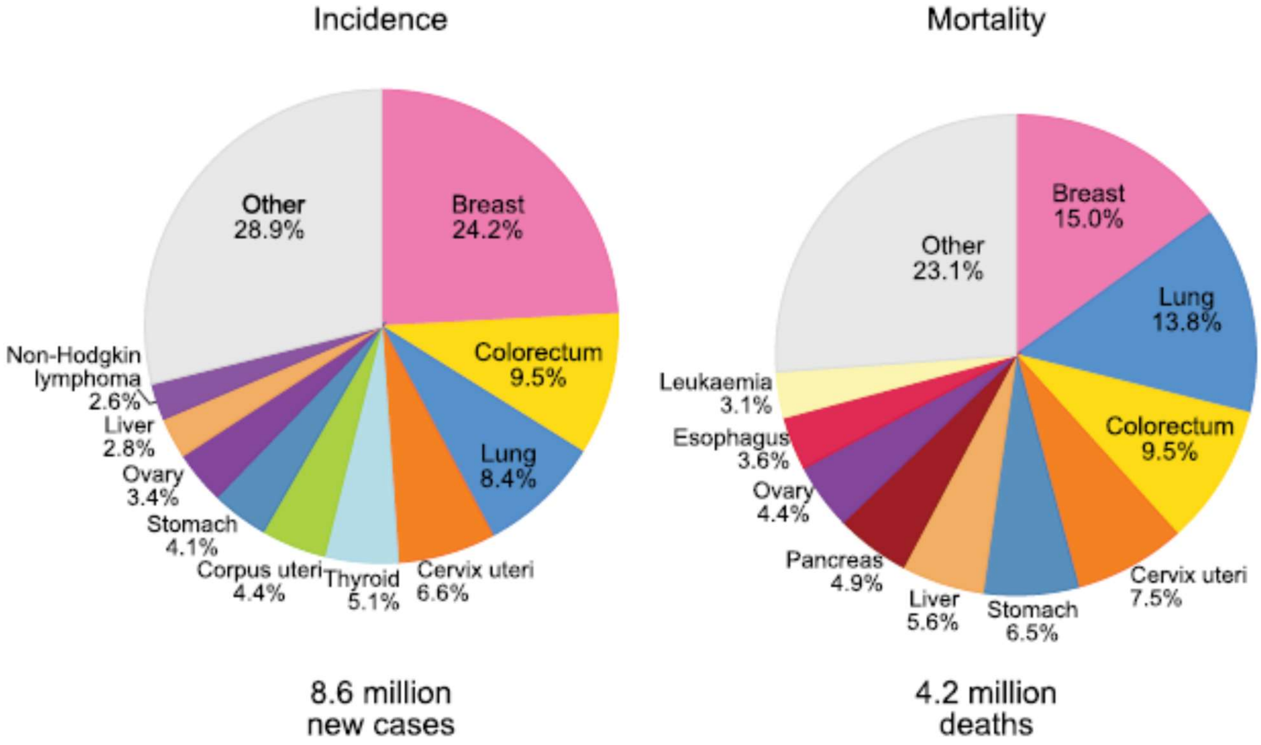


FIGURE 1. Pie Charts Present the Distribution of Cases and Deaths for the 10 Most Common Cancers in 2018 for Females. The area of the pie chart reflects the proportion of the total number of cases or deaths. (CAAC paper)

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”).

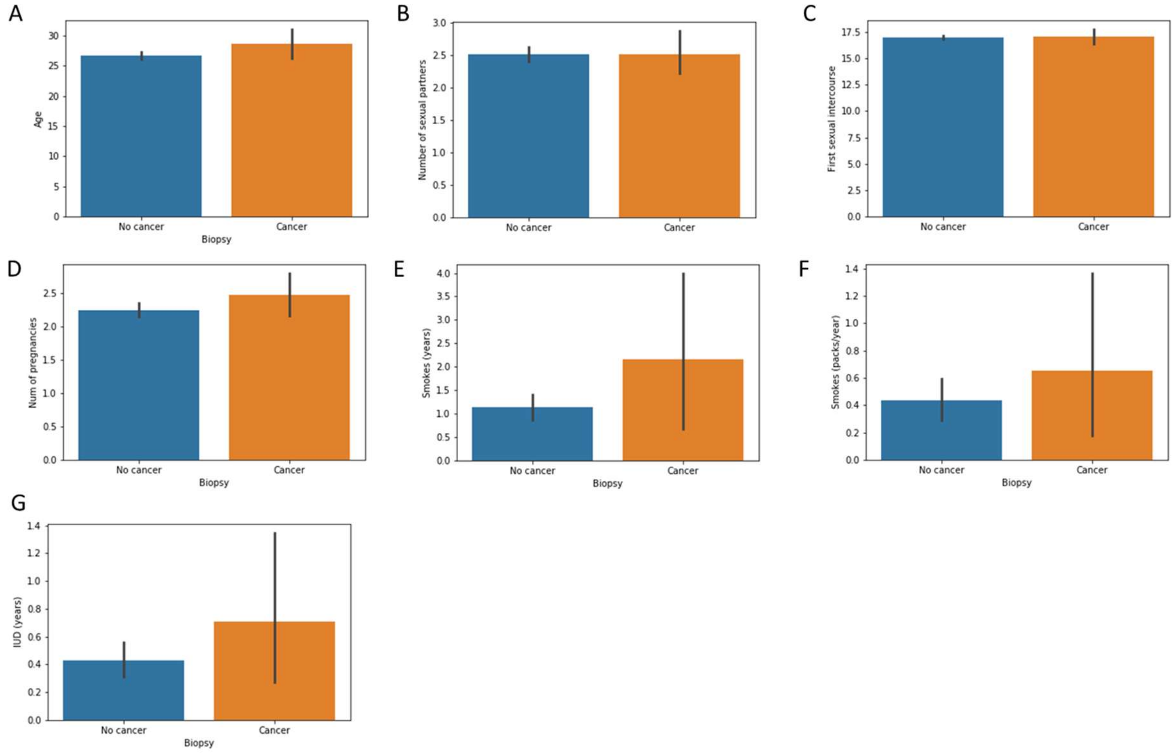


FIGURE 2. Bar plots present the difference in mean of numeric features for people with or without cervical cancer. (A) Age. (B) Number of sexual partners. (C) First sexual intercourse. (D) Number of pregnancies. (E) Smokes (years). (F) Smokes (packs/year). (G) IUD (years).

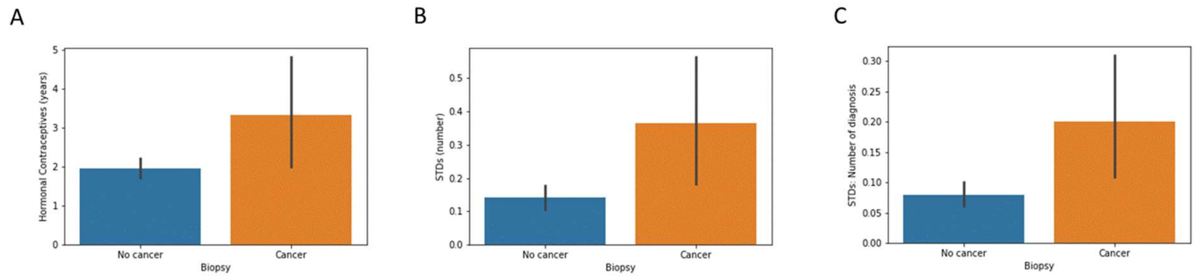


FIGURE 3. Bar plots present the difference in mean of numeric features for people with or without cervical cancer. (A) Hormonal Contraceptives (years). (B) STDs (numbers). (C) STDs: Number of diagnosis.

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)

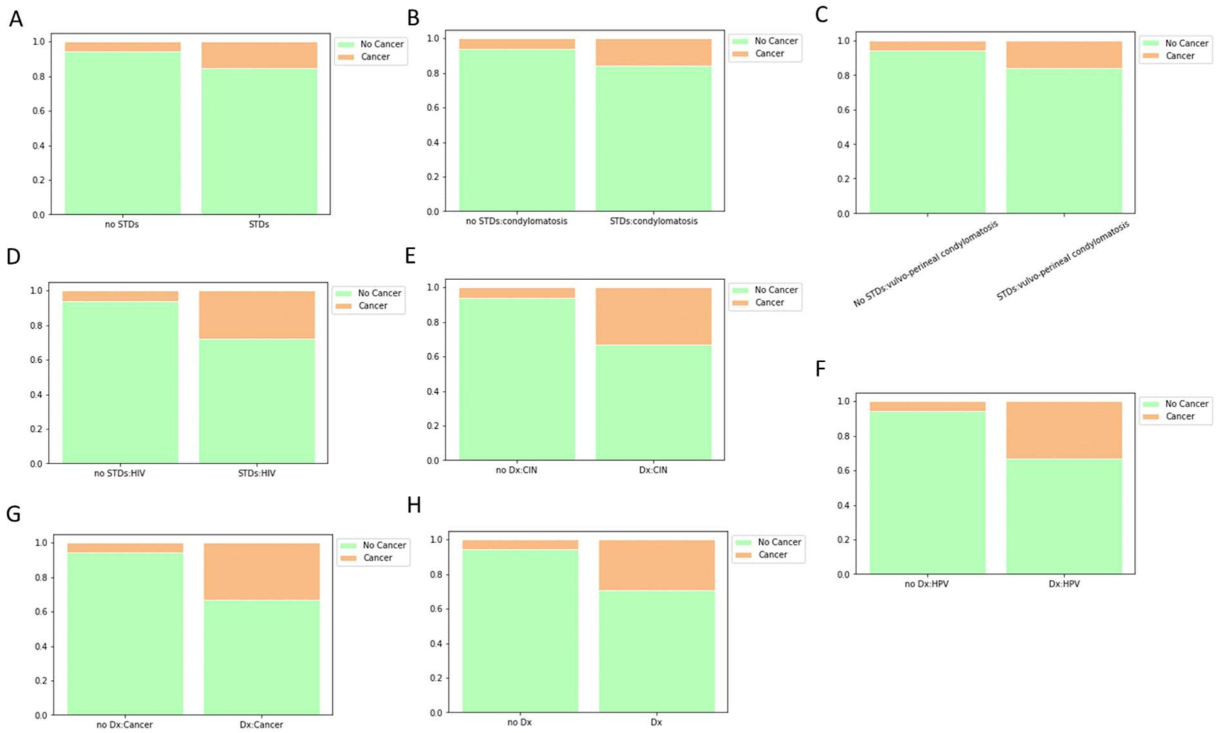


FIGURE. 4 Stacked bar plots compare the proportion of people with or without cancer for different categorical features. (A) STDs. (B) STDs: condylomatosis. (C) STDs: vulvo-perineal condylomatosis. (D) STDs: HIV. (E) Dx: CIN. (F)"Dx: HPV. (G) Dx: Cancer. (H) Dx.

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.

Machine Learning

Supervised classification algorithms were used to predict whether a person has cervical cancer or not. Biopsy was considered as target.

1. Preprocessing data

Before analyzing, standardization is performed to the datasets, since it is a common requirement for many machine learning estimators. After standardization, all the features with numeric data are with zero mean and unit variance.

2. Classification metrics

For medical data like this cervical cancer dataset, we should consider the correct diagnosis, only the total accuracy is not enough for evaluating an algorithm. So I employed the following classification metrics.

	Diagnosis	
	positive	negative
Test outcome positive	TP (True positive)	FP (False positive)
Test outcome negative	FN (False negative)	TN (True negative)

Table 1. Basic terminology and derivations

Term	Formula
Accuracy	(TP + TN)/(P+N)
Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
F-measure	(2 x recall x precision) / (recall+precision)

Table 2. Basic terminology and formula

Confusion Matrix were used to present the TP, FP, FN and TN of the prediction. Classification Report were used to show the precision, recall, F1-score and support for each class.

3. Compare Machine Learning Models

The following classification algorithms were used to analyze this dataset:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Decision Tree (DR)
- Random Forest (RF)

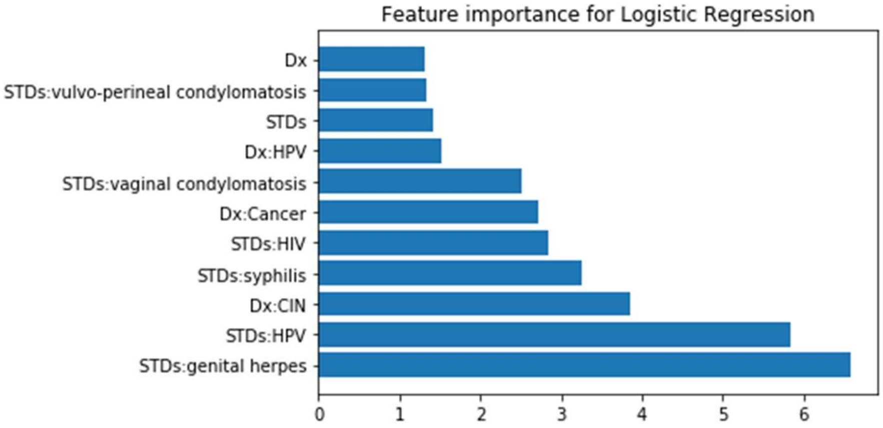
After tuning the parameters using cross validation (Recall was used as the scoring parameter, because correctly detecting more people with cancer is the most important. I also used f1 as the scoring parameter, similar result has gotten). Although the accuracy seems ok, all four classifiers suffered with low Recall, precision and F1 scores. Decision tree and random forest perform better than the other two models in terms of recall, precision and F1.

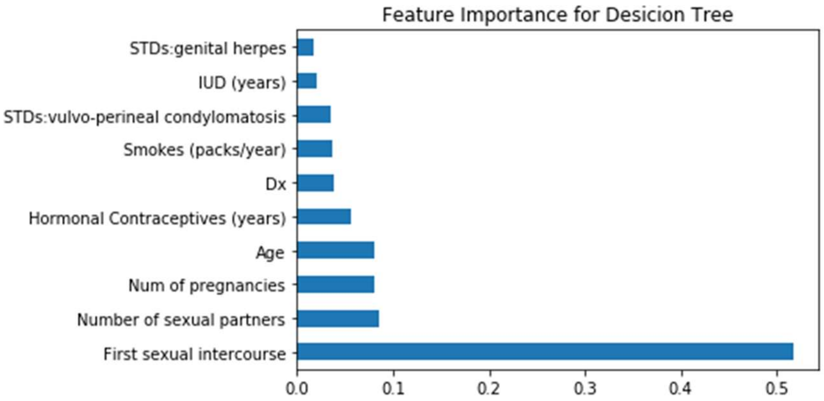
Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.942	0.913	0.900	0.953
Recall	0.111	0.222	0.444	0.333
Precision	0.333	0.200	0.222	0.6
F1	0.167	0.211	0.296	0.429

Table 3 Result of different classifiers (using original data)

3.1 Feature Selection

Keeping irrelevant features can result in overfitting and decreasing accuracy. Feature selection could reduce overfitting, improve accuracy and reduce training time. So for each classifier (except for the SVM with rbf kernel), the most important features for the prediction were identified. The models were trained with the most important 10 features to improve the performance of the classifiers.





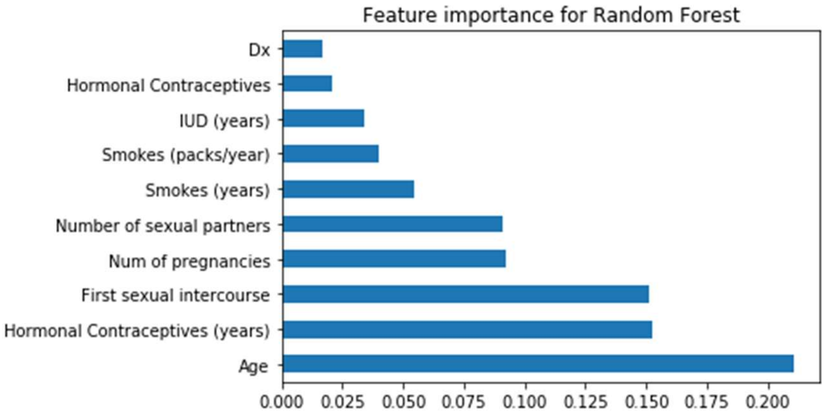


FIGURE. 5 10 most important features for logistic regreesion, decision tree and random forest classifiers.

Classifier /Performance	Logistic regression	Decision Tree	Random Forest
Accuracy	0.948	0.907	0.948
Recall	0.111	0.667	0.222
Precision	0.5	0.316	0.5
F1	0.182	0.429	0.308

Table 4 Result of different classifiers (using original data with feature selection)

After feature selection, decision tree classifier was improved with higher recall, precision and F1. The accuracy of logistic regression also increased. The performance of random forest did not change much.

4. Class Imbalance

To address the imbalanced classes' problem, the following three techniques were used.

4.1 Oversample minority class (people with cancer)

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.744	0.779	0.907	0.959
Recall	0.556	0.333	0.222	0.222
Precision	0.111	0.0857	0.182	1
F1	0.185	0.136	0.200	0.363

Table 5 Result of different classifiers (using upsampling technique)

The recall improved for logistic regression and SVM, but the accuracy decreased for these two classifiers. Decision tree and random forest did not perform better than using the original data.

4.2 Undersample majority class (people without cancer)

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.599	0.738	0.517	0.639
Recall	0.444	0.555	0.667	0.777
Precision	0.0588	0.109	0.0698	0.104
F1	0.104	0.182	0.126	0.184

Table 6 Result of different classifiers (using downsampling technique)

This technique did not perform well, because the sample size is too small.

4.3 Generate synthetic samples

Imblearn's SMOTE (Synthetic Minority Oversampling Technique) was used to generate synthetic samples. This technique uses a nearest neighbors algorithm to generate new and synthetic data.

Classifier /Performance	Logistic regression	SVM	Decision Tree	Random Forest
Accuracy	0.779	0.756	0.860	0.936
Recall	0.556	0.333	0.222	0.222
Precision	0.128	0.0769	0.105	0.333
F1	0.208	0.125	0.143	0.267

Table 7 Result of different classifiers (using SMOTE)

Similar results has gotten to the oversampling technique. The recall improved for logistic regression and SVM, but the accuracy decreased for these two models. Decision tree and random forest did not perform better than using the original data.

Summary

In this project, I performed cervical cancer prediction use the cervical cancer (Risk Factor) dataset from The UCI Machine Learning Repository. Different machine learning algorithms has been used, including logistic regression, support vector machine, decision tree, and random forest. Feature selection and resample techniques has been used to improve the performance of the classifiers. Based on the results, almost all classifiers suffered with low recall, precision and F1 scores. It appears for this particular dataset, decision tree with feature selection has the best overall performance.

Ongoing Works

I will perform feature selection for the upsampled and synthetic samples. I will employ ROC curves and Precision-Recall curves to further evaluate the classifiers.

Other algorithms, such as XGboost will be tried.

This dataset provides three other targets Hinselmann, Schiller and Cytology. I will combine three targets together with Biopsy, and train the classifier with this new target.

Acknowledgement

The author would like to thank Springboard and especially her mentors for the advice and support.