Predicting Cases of Cervical Cancer in the 'Hospital Universitario de Caracas'

Ben Herndon-Miller & Nicole Jaiyesimi

Department of Statistics, Rice University

December 12, 2018

1 Introduction

This paper is the product of a term project for the course Generalized Linear Models and Categorical Data Analysis. Our selected project goal was to "obtain an appropriate data set and analyze it". Accordingly, we chose a data set focusing on the risk factors for cervical cancer, specifically at 'Hospital Universitario de Caracas' in Caracas, Venezuela.

In terms of analyzing the data set, we decided to build predictive models for identifying cases of cervical cancer among patients in the 'Hospital Universitario de Caracas', and identify which features are most important for predicting cases of cervical cancer. We will compare multiple families of Generalized Linear Models for classification, as well as more modern machine learning techniques in order to determine our best predictive model. Furthermore, we will use our best performing model to report the relative importance of each predictor to understand what factors are most associated with cervical cancer. Hopefully, our discoveries can be used by medical professionals to inform their research and ability to treat patients.

2 Background & Motivation

With 570,000 new cases in 2018 representing 6.6% of all female cancers, cervical cancer is the most common cancer among women worldwide (World Health Organization, 2018). An even more striking figure is how many deaths from cervical cancer occur in middle and low income countries: 90% (World Health Organization, 2018). This brings us back to our data, focusing on Venezuela. According to The Development Policy and Analysis Division of the Department of Economic and Social Affairs of the United Nations Secretariat Venezuela is classified as a 'developing country'(DPAD, 2013). A 'developing country' is the lowest possible rank in their system designed to reflect countries' economic conditions. Venezuela's 11.5% (Raúl, et al, 2016) cervical cancer mortality rate compared to America's 2.3% (Kaiser Family Foundation, n.d.) one further emphasizes the higher occurrence of death from cervical cancer in developing countries. In fact, in Venezuela, approximately 4,973 women are diagnosed with cervical cancer and 1,789 women die of it every year. These figures make cervical cancer the second most frequent cancer among Venezuela women as a whole, and the first most frequent cancer among Venezuelan women between the ages of 15 and 44 years old (ICO/IARC Information Centre, 2018).

According to the World Health Organization, "The high mortality rate from cervical cancer globally could be reduce through a comprehensive approach that includes prevention, early diagnosis, effective screening, and treatment programmes." This became the basis of our motivation for this project. We believe that predictive modeling can play a role in identifying the most important risk factors for cervical cancer in developing countries.

3 Data Description & Missing Data Methodology

The data was collected at Hospital Universitario (which is located in Caracas Venezuela) in 2017. There are 36 attributes in the dataset, all of which are either integers or booleans (e.g. age, number of sexual partners, whether or not the individual smokes, whether or not the individual has AIDS).

Variable	Variable Type	Missing Values	Percent Missing
Age	int	0	0.0000000
Number.of.sexual.partners	int	26	0.0303030
First.sexual.intercourse	int	7	0.0081585
Num.of.pregnancies	int	56	0.0652681
Smokes	bool	13	0.0151515
Smokes.years.	int	13	0.0151515
Smokes..packs.year.	int	13	0.0151515
Hormonal.Contraceptives	bool	108	0.1258741
Hormonal.Contraceptives..years.	int	108	0.1258741
IUD	bool	117	0.1363636
IUD..years.	int	117	0.1363636
STDs	bool	105	0.1223776
STDs..number.	int	105	0.1223776
STDs.condylomatosis	bool	105	0.1223776
STDs.cervical.condylomatosis	bool	105	0.1223776
STDs.vaginal.condylomatosis	bool	105	0.1223776
STDs.vulvo.perineal.condylomatosis	bool	105	0.1223776
STDs.syphilis	bool	105	0.1223776
STDs.pelvic.inflammatory.disease	bool	105	0.1223776
STDs.genital.herpes	bool	105	0.1223776
STDs.molluscum.contagiosum	bool	105	0.1223776
STDs.AIDS	bool	105	0.1223776
STDs.HIV	bool	105	0.1223776
STDs.Hepatitis.B	bool	105	0.1223776
STDs.HPV	bool	105	0.1223776
STDs..Number.of.diagnosis	int	0	0.0000000
STDs..Time.since.first.diagnosis	int	787	0.9172494
STDs..Time.since.last.diagnosis	int	787	0.9172494
Dx.Cancer	bool	0	0.0000000
Dx.CIN	bool	0	0.0000000
Dx.HPV	bool	0	0.0000000
Dx	bool	0	0.0000000
Biopsy	bool	0	0.0000000

Table 1: Missing Data by Variable

There are 858 instances of the 36 attributes, but there are some missing values. Furthermore, the description of the dataset does not include detailed information on a variety of variables. First and foremost, there are 4 variables that could be considered the outcome variable of interest for a diagnosis of cancer: *Hinselmann*, *Schiller*, *Citology*, or *Biopsy*. These all represent different screening techniques for identifying and diagnosing cervical cancer. Since we need to determine a single outcome variable to predict, we proceed using the binary variable *Biopsy* as it is the gold standard for diagnosing cervical cancer. We will discard the other potential outcome variables from our dataset before we proceed.

We can see from the missing data analysis in Table 1 that the majority of our variables have

missing data points. However, most features are not missing a high-proportion of values, except for Time Since First/Last Diagnosis of STD. These are all missing for the patients who stated that they had not had any STD's, so it would be inaccurate to simply replace them with 0 as that would bias the model towards more patients having a small time period since their STD. Replacing them with the mean or median value would also not make sense since the true value would be infinity since they have never had a diagnosis. For the sake of simplicity, we will move forward by removing these two variables from our analysis.

However, we still must deal with the missing values for the other variables. For the integer variables of *Age of first sexual intercourse*, *Number of sexual partners*, and *Number of pregnancies*, we can replace the missing values with the median values of their respective variables. In contrast, for the variables pertaining to smoking, STD's, and contraceptives, we cannot simply replace the missing values with the median value as they are based on boolean values. Thus, we will then remove the rows with missing values for the missing variables and report. After cleaning the data with the above methodology we are left with a dataset of 726 observations and 31 features (30 predictors). We can now move forward with exploratory analysis and predictive modelling.

4 Exploratory Data Analysis

We begin with looking at the Biopsy variable, as it is our outcome variable.

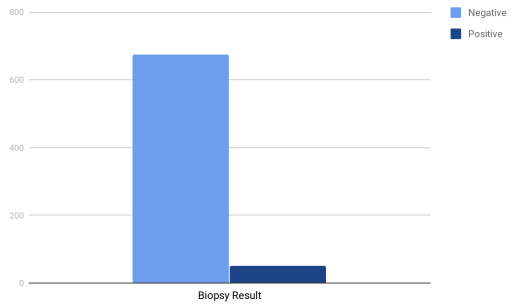


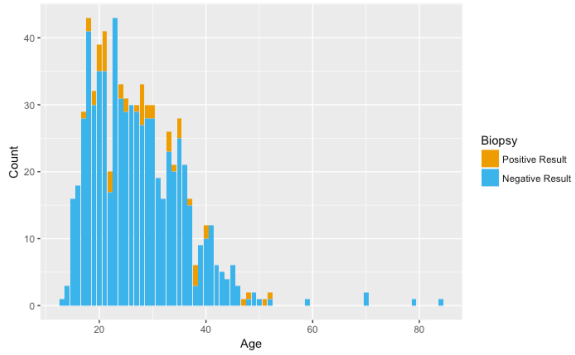
Figure 1: The Unbalanced Outcome Variable

Of the 726 observations remaining after cleaning our data, only 50 of them (6.9%) have a positive biopsy outcome. We will address this imbalance in the predictive modeling portion of our presentation.

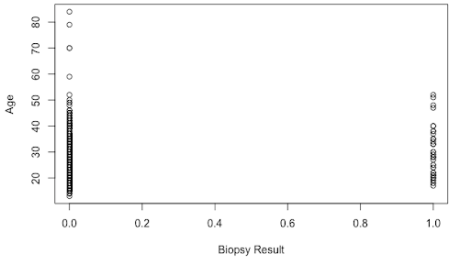
Considering our goal of building a highly predictive model for identifying cases of cervical cancer, we believe that exploring our 30 predictors is a crucial aspect of identifying relationships that may exist between them and Biopsy outcome variable. The imbalance of the *Biopsy* variable begs the question of whether or not specific values of certain predictors can be tied to the positive Biopsy cases. None of our predictors are continuous, so we can divide our discrete variables into two categories for further exploration: integer and boolean.

4.1 Integer Predictors

We constructed simple scatter plots and stacked bar graphs for our integer predictors. Plotting each integer predictor versus Biopsy outcome with the scatter plots allows us to compare the integer values of the points with a positive cancer biopsy to those with a negative one. The stacked bar graphs allow us to see relative frequencies of positive and negative biopsies for each value of an integer predictor.



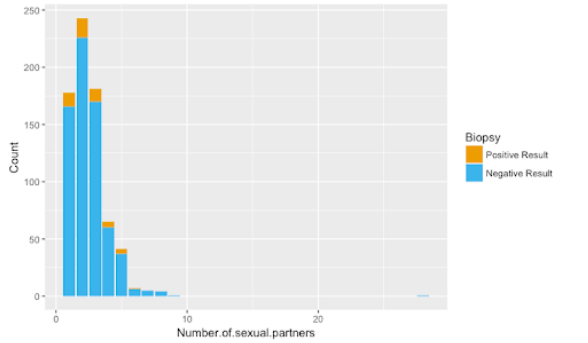
(a) Stacked Bar Chart



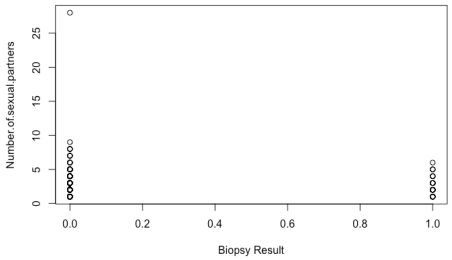
(b) Scatter Plot

Figure 2: Graphics for Integer Predictor: Age

The majority of patients with a positive biopsy were between 18 and 52 years old and no woman over 60 years old had a positive biopsy result.



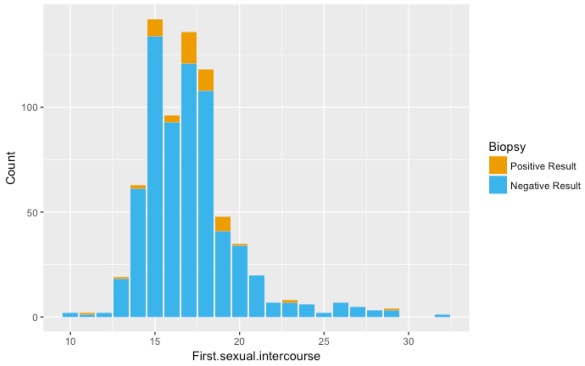
(a) Stacked Bar Chart



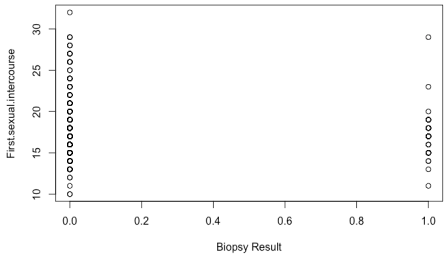
(b) Scatter Plot

Figure 3: Graphics for Integer Predictor: Number of Sexual Partners

Most patients in this data set had 5 or less sexual partners.



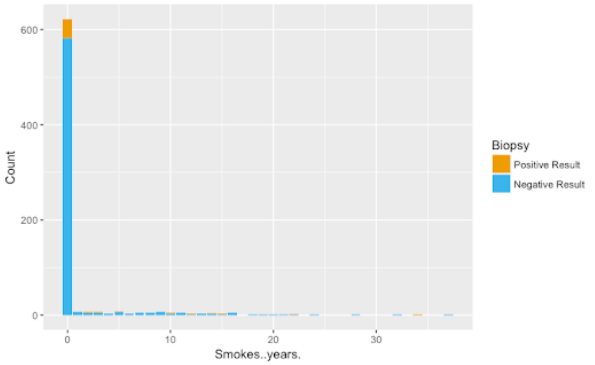
(a) Stacked Bar Chart



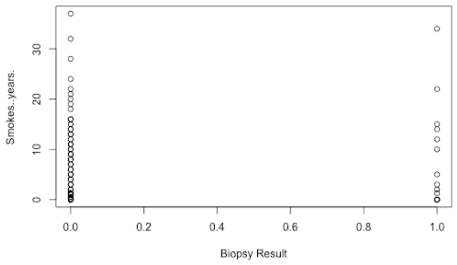
(b) Scatter Plot

Figure 4: Graphics for Integer Predictor: Age of First Sexual Intercourse

Most people in this data set had sexual intercourse for the first time between the ages of 13 and 21. For those with a positive biopsy result, the age of first sexual intercourse was clustered between 15 and 19.



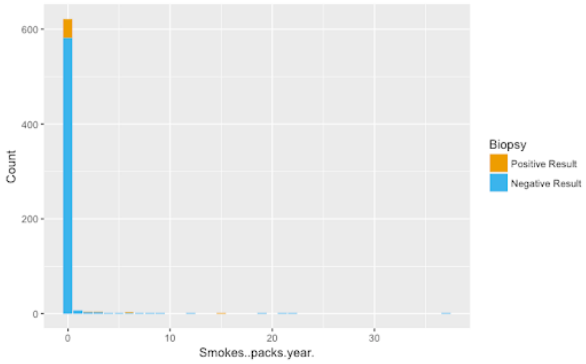
(a) Stacked Bar Chart



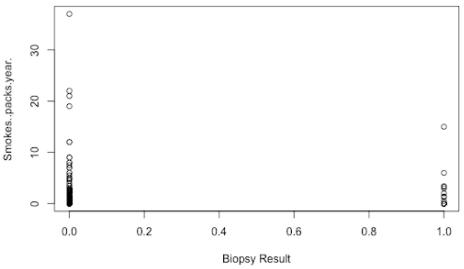
(b) Scatter Plot

Figure 5: Graphics for Integer Predictor: Number of Years Smoking

The majority of patients in this data set did not smoke.



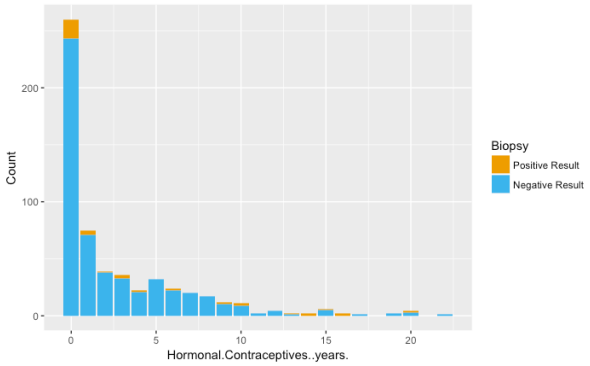
(a) Stacked Bar Chart



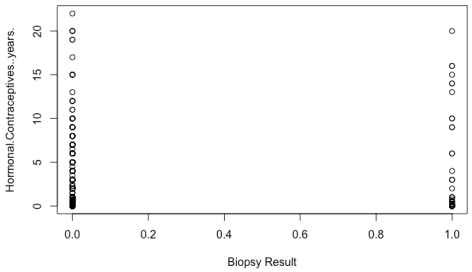
(b) Scatter Plot

Figure 6: Graphics for Integer Predictor: Number of Packs Smoked per Year

Again, the majority of patients in this data set did not smoke. For those with a positive biopsy, smoking was clustered between 0-10 packs/year.



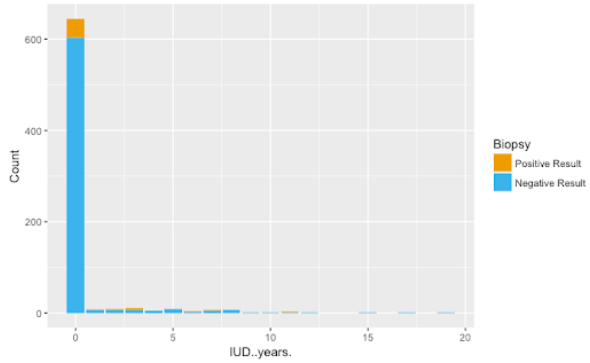
(a) Stacked Bar Chart



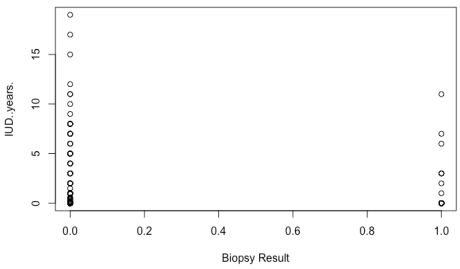
(b) Scatter Plot

Figure 7: Graphics for Integer Predictor: Years of Hormonal Birth Control Use

It is important to note the scale on this graph. Although there is a peak at 0, the y-axis goes up to 300 instead of 200. So, the majority of patients in this data set had used hormonal birth control for somewhere between 1 and 25 years.



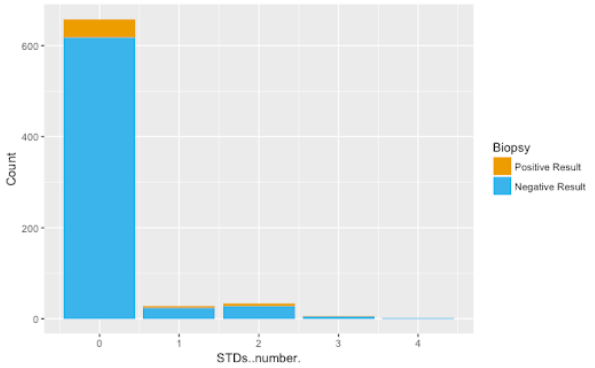
(a) Stacked Bar Chart



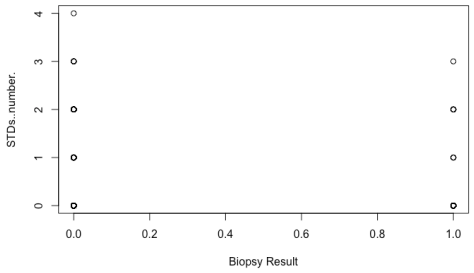
(b) Scatter Plot

Figure 8: Graphics for Integer Predictor: Years of IUD Use

The majority of patients in this study had never used an IUD. All of those who have a positive biopsy had used an IUD for less than 15 years, most being between 0 and 5 years.



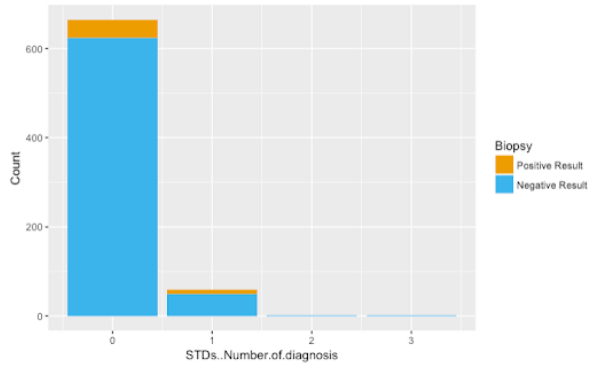
(a) Stacked Bar Chart



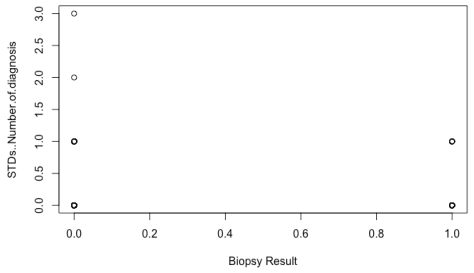
(b) Scatter Plot

Figure 9: Graphics for Integer Predictor: Number of STDs

The majority of patients in this study have never had an STD, but we also see a non-negligible amount of patients who have had one or two. No patients with a positive biopsy had more than 3 STDs.



(a) Stacked Bar Chart



(b) Scatter Plot

Figure 10: Graphics for Integer Predictor: Number of STD Diagnoses

The majority of patients in this study have never been diagnosed with an STD, and that none of the patients who have a positive biopsy result have more than two STD diagnoses. It is interesting that there is such a low count for two diagnoses, considering that above we saw that there was a higher count for having had two STDs than one.

4.2 Boolean Predictors

Plotting the boolean predictors against the Biopsy outcome would not give us a useful result. Chart data is more valuable in this case. For each boolean predictor we list the count and percentage of patients with positive biopsies that also have a corresponding positive value for that predictor.

Predictor	Count Positive	Percent Poitive
Smokes	10	(20%)
Hormonal.Contraceptives	33	(66%)
IUD	9	(18%)
STDs	11	(22%)
STDs.condylomatosis	7	(14%)
STDs.cervical.condylomatosis	0	(0%)
STDs.vaginal.condylomatosis	0	(0%)
STDs.vulvo.perineal.condylomatosis	7	(14%)
STDs.syphilis	0	(0%)
STDs.pelvic.inflammatory.disease	0	(0%)
STDs.genital.herpes	1	(2%)
STDs.molluscum.contagiosum	0	(0%)
STDs.AIDS	0	(0%)
STDs.HIV	4	(8%)
STDs.Hepatitis.B	0	(0%)
STDs.HPV	0	(0%)
Dx.Cancer	6	(12%)
Dx.CIN	2	(4%)
Dx.HPV	6	(12%)
Dx	6	(12%)

Table 2: Positive Boolean Variables with Positive Biopsy Results

Here we see that a large percentage of women with positive biopsies from this data set also have used hormonal contraceptives before. Other notable predictors are smoking, IUD use, and having had an STD.

5 Predictive Modelling

5.1 Methodology

5.1.1 Data Splitting & Cross-Validation

An important aspect of predictive modelling is the tuning of hyper-parameters of models in order to achieve the maximum predictive accuracy. In order to do this, we must be able to work with unseen data to provide unbiased estimates of prediction accuracy. This can be achieved through data splitting and cross validation. We first separate the data into a training (70%) and testing (30%) set. We then use the training set in order to achieve the optimal parameters for each model using 5-fold cross-validation. Once we have determined the optimal combination of hyper-parameter values, we can compare predictive accuracy between models by predicting on the hidden test set allowing us to identify our best performing model overall.

5.1.2 Loss Function

In order to measure and compare predictive accuracy in a consistent manner between both different versions of models and different families of models, we must choose a loss function, or metric for predictive accuracy. For this project we elect to use the Area Under the Curve (AUC), with the curve in reference being the Receiver Operating Characteristic (ROC) curve. This choice of loss function was especially important due to the unbalanced nature of our data. Since we have one class that is

only 6.9% of our data, using a metric such as Misclassification Accuracy would lead us to optimize our models to predict only the majority class.

As previously stated, the AUC loss function measures the area under the ROC curve. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), or in other words, the power of the model versus its Type I error rate. We will report the AUC on both the Training and Test sets in order to get a sense of how each model predicts on seen and unseen data, the latter of which being far more important. We define the aforementioned equations below as well as provide an example ROC curve that shows what a good, random, and bad classifier look like in terms of their ROC curve.

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

$$FPR = 1 - Specificity = \frac{FP}{TN + FP}$$

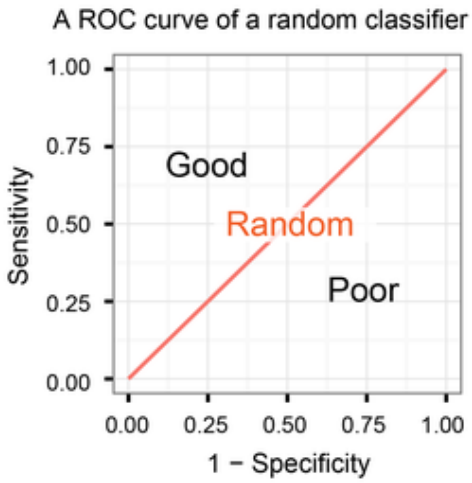


Figure 11: ROC Curve Example (Saito, Takaya, and Marc Rehmsmeier, 2015)

5.2 Binomial Regression with Different Link Functions

We begin by examining how different link functions for binomial regression perform on our data set. We elect to use the logit, probit, cauchit, and complementary log-log link functions. These different link functions use different equations to keep the model output between 0 and 1, which is quite important for our task of predicting a binary variable. We report the AUC on the training set and the test set for these four models below as well as visualizing the ROC curves.

Link Function	Train AUC	Test AUC
Logit	0.772	0.534
Probit	0.783	0.531
Cauchit	0.712	0.511
Complementary Log-Log	0.764	0.469

Table 3: AUC for Different Link Functions

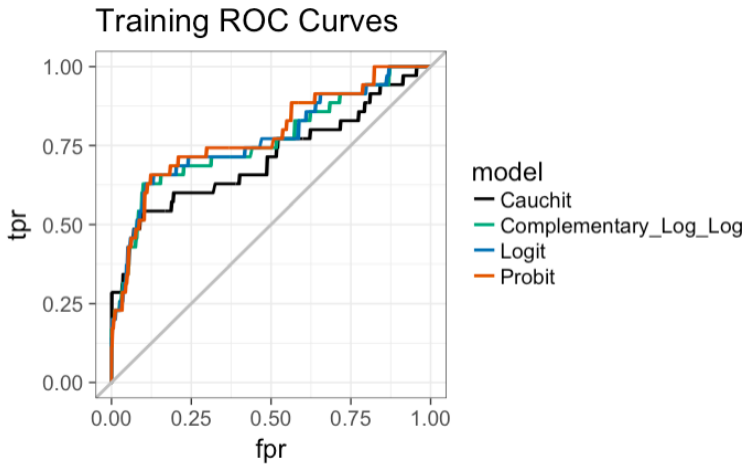


Figure 12: Training ROC Curves for Link Functions

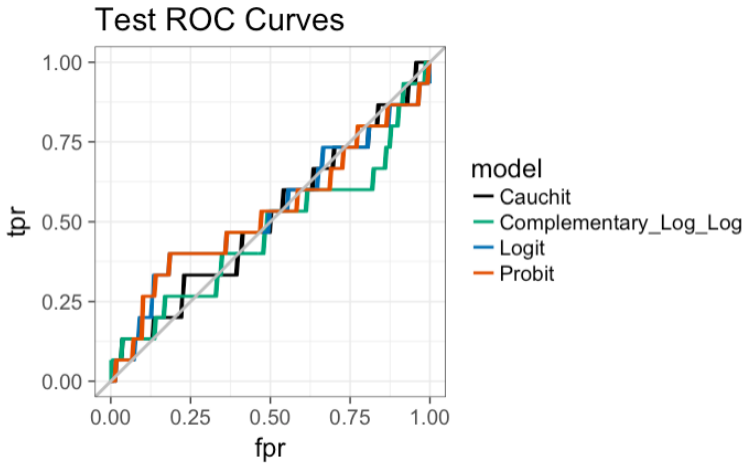


Figure 13: Test ROC Curves for Link Functions

We can see that these models predict pretty well on the training data, but clearly perform quite poorly on the unseen test data. This shows that all of these basic binomial regression models are over-fitting to the training data and not performing much better than a random guess on new data.

5.3 Penalized Logistic Regression & Feature Selection Algorithms

After trying different link functions, we decided to take another approach using binomial regression with the logit link, or logistic regression. We implement a penalty term on the coefficients in order to reduce irrelevant predictors influence on the outcome variable. Furthermore, we compare this with a stepwise selection algorithm that reduces the number of features in the model by comparing the AIC of each model. We report the AUC on the training set and the test set for these four models below as well as visualizing the ROC curves. We include the un-penalized logistic regression model as our baseline for further comparison.

Method	Train AUC	Test AUC
Elastic Net Logistic Regression	0.766	0.589
L2 Logistic Regression	0.770	0.585
Stepwise Logistic Regression	0.771	0.557
L1 Logistic Regression	0.770	0.548
Logistic Regression	0.772	0.534

Table 4: AUC for Different Penalties and Feature Selection Models

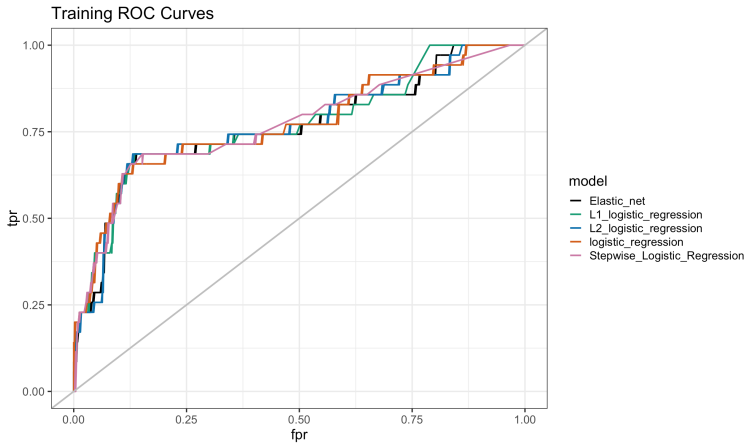


Figure 14: Training ROC Curves for Different Penalties and Feature Selection Models

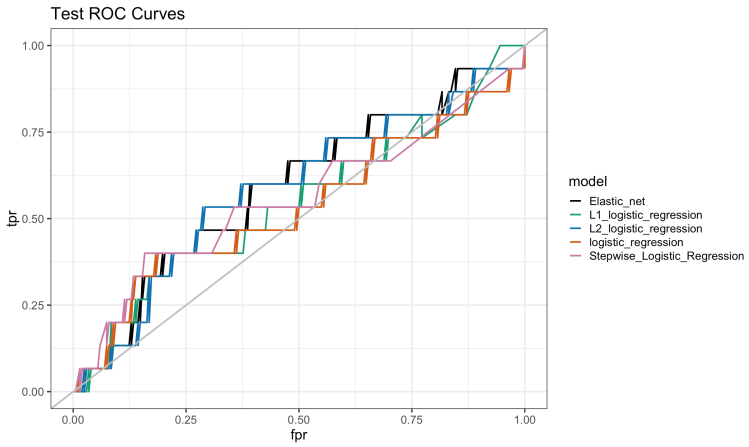


Figure 15: Test ROC Curves for Different Penalties and Feature Selection Models

We can see that these models certainly perform better on the test set than the un-penalized binomial regression models we tried originally. Furthermore, we see that the Elastic Net model which combines the L1 and L2 penalty terms performs the best our of all penalization or feature selection algorithms. Most of these models perform better than a random guess, but they still over-fit to the training data and are pretty weak predictive models.

5.4 Penalized Logistic Regression with Re-sampling Techniques

A common method for dealing with unbalanced data is to use different re-sampling techniques during the cross-validation process. We use a variety including using class-weights, up-sampling, down-sampling, and Synthetic Minority Over-sampling Technique (SMOTE). These methods try to use different weighting and re-sampling schemes to increase the importance of the minority class or produce more balanced data sets. We report the AUC on the training set and the test set for these five models below as well as visualizing the ROC curves. We include the un-penalized logistic regression model and the Elastic Net without re-sampling as our baselines.

Method	Train AUC	Test AUC
Up-sampling	0.784	0.590
No Re-sampling	0.766	0.589
Class Weights	0.788	0.587
Down-sampling	0.745	0.573
SMOTE	0.783	0.557
Logistic Regression	0.772	0.534

Table 5: AUC for Penalized Logistic Regression with Re-sampling Techniques

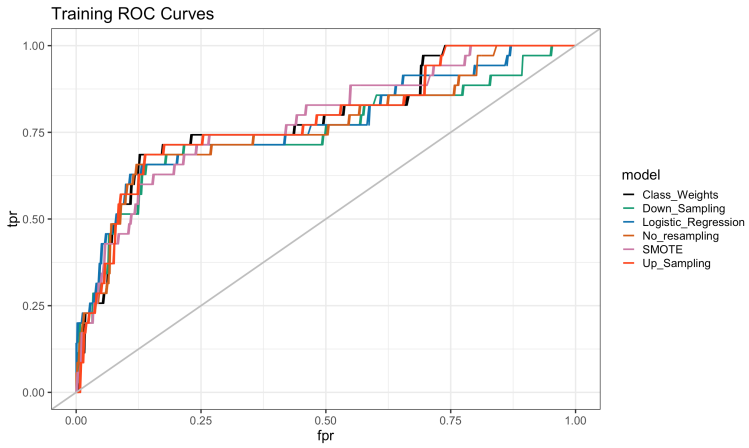


Figure 16: Training ROC Curves for Penalized Logistic Regression with Re-sampling Techniques

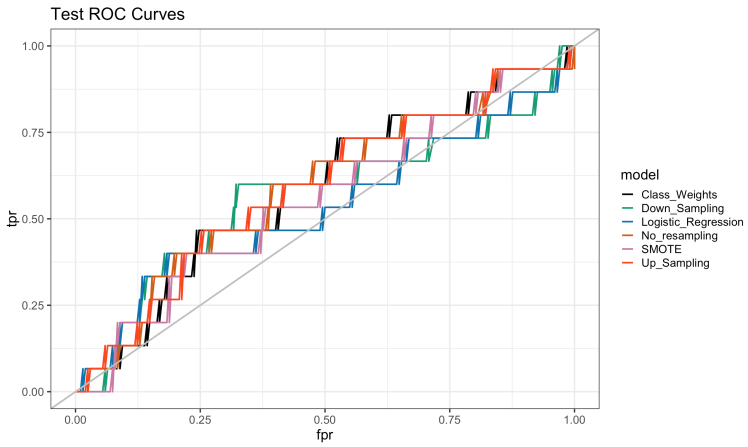


Figure 17: Test ROC Curves for Penalized Logistic Regression with Re-sampling Techniques

We can see that using these re-sampling methods do not increase our predictive performance on the test set by any significant margin. Up-sampling performs the best, but only by .001 compared to the Elastic Net model without re-sampling. This suggests that re-sampling is not particularly effective on penalized logistic regression models for this data. However, this does not mean that it may not be effective for other non-linear models we will investigate later.

5.5 Other GLM Frameworks & Support Vector Machines

In addition to the varieties of binomial regression, and more specifically, varieties of penalized logistic regression, we try two other Generalized Linear Model frameworks as well as Support Vector Machines (SVM) with different kernels. We report the AUC on the training set and the test set for these four models below as well as visualizing the ROC curves. We include the un-penalized logistic regression model as our baseline for further comparison.

Method	**Train AUC**	**Test AUC**
Generalized Additive Model	0.795	0.605
Radial Basis SVM	0.957	0.591
Linear SVM	0.557	0.559
Logistic Regression	0.772	0.534
Boosted Logistic Regression	0.768	0.526

Table 6: AUC for Other GLM Frameworks & Support Vector Machines

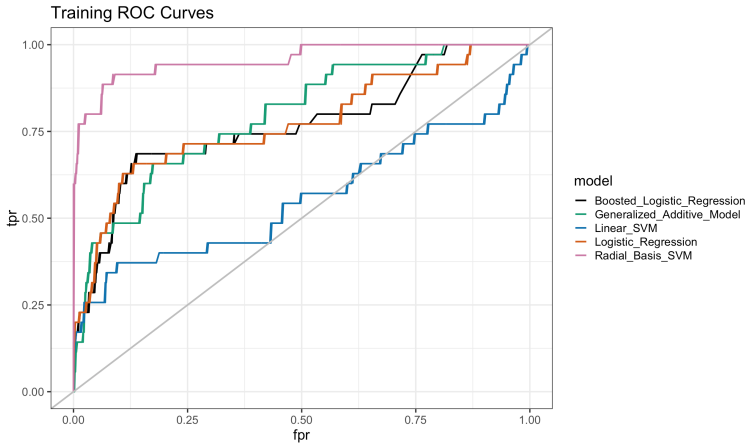


Figure 18: Training ROC Curves for Other GLM Frameworks & Support Vector Machines

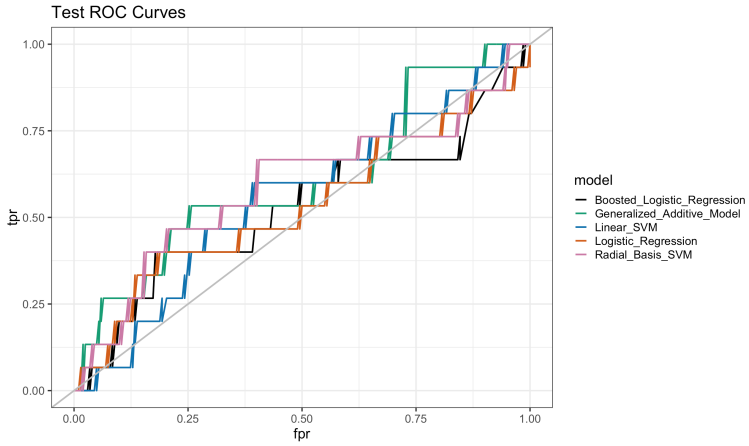


Figure 19: Test ROC Curves for Other GLM Frameworks & Support Vector Machines

We notice that the Generalized Additive Model (GAM) performs the best out of all the models in this group on the test set. Interestingly, the Radial Basis SVM has a pretty amazing predictive accuracy on the training set, but is out-performed on the unseen data by the GAM. This suggests that the Radial Basis function causes the SVM to overfit to the data. Even though we get some slight improvements with these models, we still have relatively weak predictive models and are only performing slightly better than making a random guess on new data.

5.6 Modern Machine Learning Classifiers

After trying different versions of GLM's and other classifiers that use a linear framework, we decided it would valuable to try some non-linear predictive models that are commonly used in machine learning today. We try a variety of these methods and report the AUC on the training set and the test set as well as visualizing the ROC curves. We include the un-penalized logistic regression model as our baseline for further comparison.

Method	Train AUC	Test AUC
Extreme Gradient Boosting	0.817	0.719
Adaboost	0.999	0.709
Gradient Boosting Machine	0.840	0.559
Random Forest	0.944	0.645
Logistic Regression	0.772	0.534

Table 7: AUC for Modern Machine Learning Classifiers

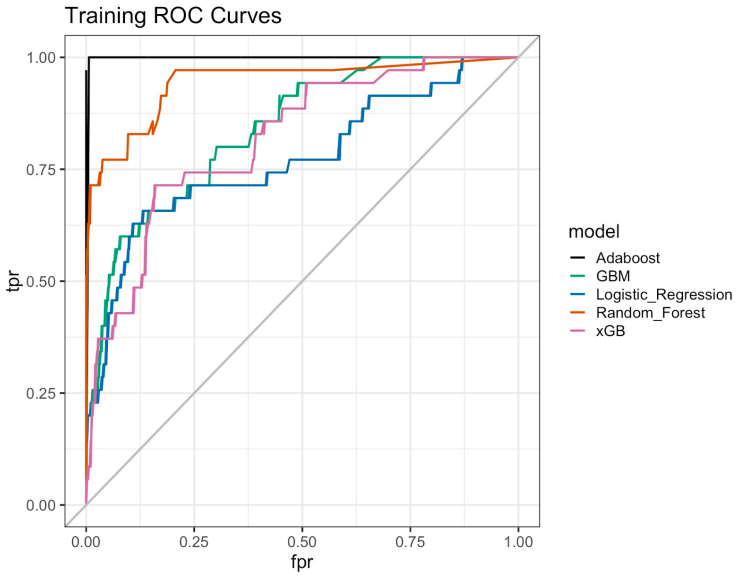


Figure 20: Training ROC Curves for Modern Machine Learning Classifiers

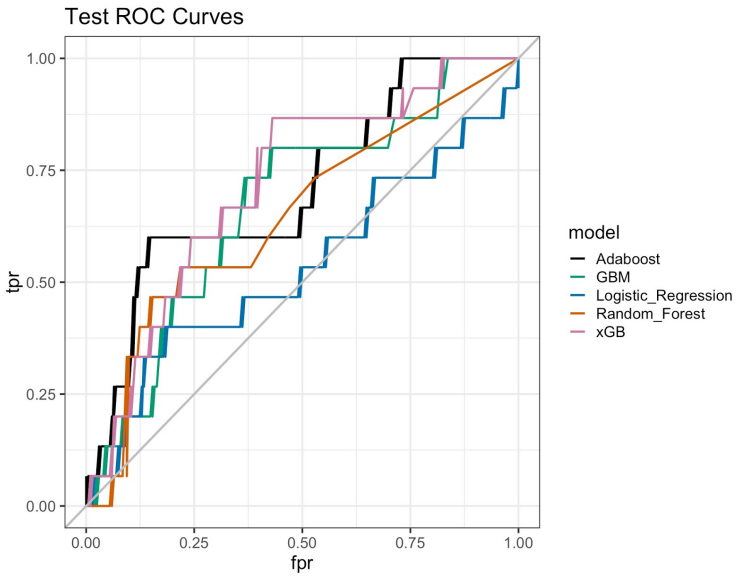


Figure 21: Test ROC Curves for Modern Machine Learning Classifiers

Not surprisingly, these more modern machine learning techniques perform *significantly* better than any of the linear models that we tried previously. Extreme Gradient Boosting (XGB) improves the AUC by almost 0.2. We see that the Adaboost model predicted the training data almost perfectly, but did not perform as well on the test set and thus it was clear that it over-fit. Furthermore, every single modern method performed better than the best performing linear method.

5.7 Extreme Gradient Boosting with Re-sampling Techniques

Now that we have determined our best performing model, we can re-visit the re-sampling techniques we tried earlier with the Elastic Net. We use these methods in the cross-validation process for Extreme Gradient Boosting to see if we can improve our predictive performance further. We report the AUC on the training set and the test set for these four models below as well as visualizing the ROC curves. We include the un-penalized logistic regression model and the Extreme Gradient Boosting model without re-sampling as our baselines.

Method	Train AUC	Test AUC
XGB Down-sampling	0.771	0.725
XGB No Re-sampling	0.817	0.719
XGB Up-sampling	0.854	0.682
XGB SMOTE	0.711	0.604
Logistic Regression	0.772	0.534

Table 8: AUC for Extreme Gradient Boosting with Re-sampling Techniques

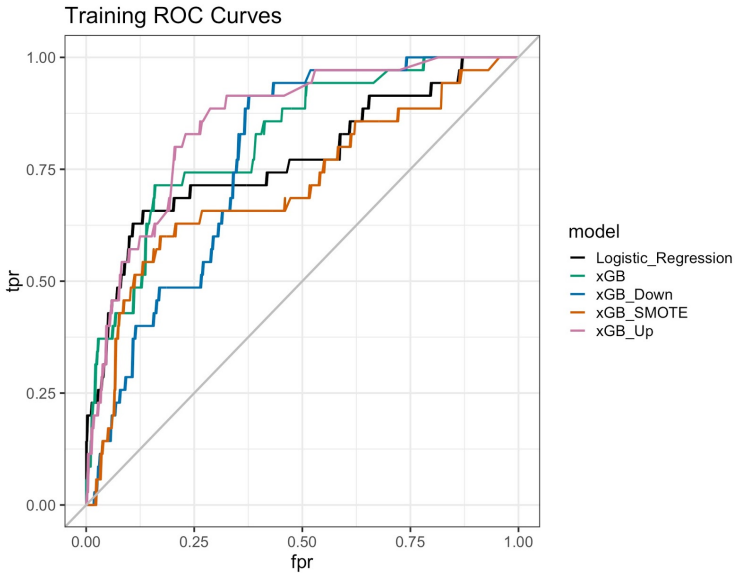


Figure 22: Training ROC Curves for Extreme Gradient Boosting with Re-sampling Techniques

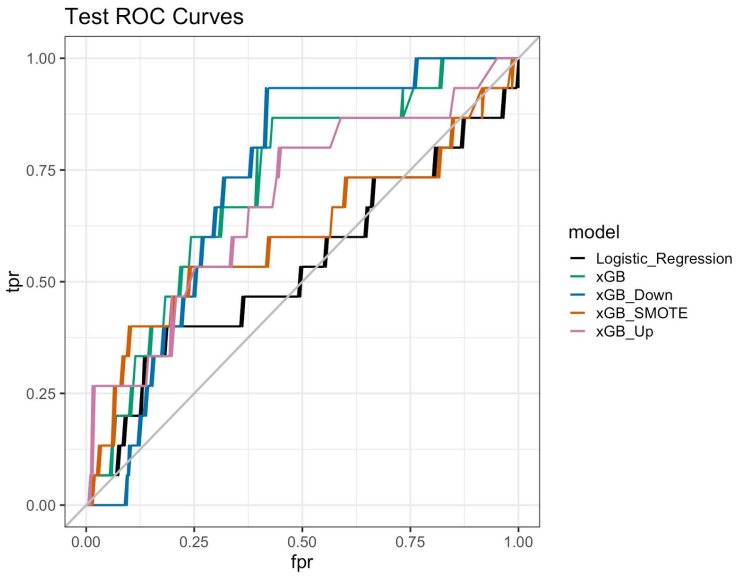


Figure 23: Test ROC Curves for Extreme Gradient Boosting with Re-sampling Techniques

We can observe that down-sampling allowed us to marginally increase our predictive performance from the basic Extreme Gradient Boosting algorithm. However, we again notice that the re-sampling techniques do not improve our ability to generate a predictive model by a significant amount. Regardless, we have still found that Extreme Gradient Boosting with down-sampling performs the best out of every single model we tried throughout this entire project.

6 Conclusion

6.1 Feature Importance from Best Performing Model

The following table includes measures of feature importance from our selected best model: Extreme Gradient Boosting with down-sampling.

Variable	Importance
Age of First Sexual Intercourse	100.000
Years of Hormonal Contraceptive Use	80.522
Age	60.392
Number of Sexual Partners	42.861
Number of Pregnancies	39.922
Smokes	18.271
STDs	10.379
Packs Smoked Per Year	8.541
Number of STDs	6.073
Years of IUD Use	.830
Years Smoking	0.000
Hormonal Contraceptive Use	0.000
IUD Use	0.000
STDs Condylomatosis	0.000
STDs Syphilis	0.000
STDs HIV	0.000
Number of STD Diagnoses	0.000
Dx. Cancer	0.000
Dx. HPV	0.000
Dx.	0.000

Table 9: Measures of Variable Importance

The importance values represent the percentage of predictive accuracy that's lost when the corresponding variable is excluded from the model. Tying this back into our preliminary research, we found that that our results concurred with research findings.

The following is true according to the Cancer Treatment centers of America: Sex before age 18, sex with multiple partners, and sex with someone who has had multiple partners are all known risk factors for cervical cancer and HPV. Five or more years of hormonal contraceptive use increases the risk of cervical cancer. Three or more full-term pregnancies and having a full term pregnancy before age 17 results in being twice as likely to get cervical cancer. And, finally, a woman who smokes doubles her risk of cervical cancer. These findings all support our non-zero values for feature importance of the first sexual intercourse, number of sexual partners, number of STDs, years of hormonal contraceptive use, number of pregnancies, smoking, packs smoked per year variables.

We also consulted the American Cancer Society's website while researching for this paper. The following information was gathered there: 35 to 44 years old is the most frequent age range for cervical cancer, and it rarely develops in women under 20. STDs impact the immune system negatively, and the immune system is an important part of destroying cancer cells and slowing their growth and spread. Finally, IUD used had a low association with developing cervical cancer compared to hormonal birth control pills. All of these findings help to support our non-zero values for feature importance of age and STDs. However, there is conflict with our non-zero value for feature importance of years of IUD use. That being said, IUD use is the smallest non-zero value and the findings could be due to intervening variables like IUD use being common among women who are in the 35-44 year old age range.

6.2 Implications & Further Studies

We believe that the results of this project can be useful for medical officials interested in understanding cervical cancer and its associated risks. By building a predictive model that allows for the interpretation of feature importance, we can inform both researchers and practitioners about the risk factors that drive cervical cancer outcomes. However, this project also has a lot of potential for growth. A

possible future course of study would be to see how our model performs on patient data from other countries and demographic groups as well (specifically developing nations since this is where mortality rates are the highest).

Furthermore, a more rigorous approach to calculating feature importance would be increase our confidence in our reported results. This could be accomplished by bootstrapping the data sample and calculating feature importance over many iterations. This would allow medical officials to make decisions based on our research with more certainty, which is vital in the world of practicing medicine.

7 References

"2014wesp_country_classification." Development Policy and Analysis Division (DPAD) of the Department of Economic and Social Affairs of the United Nations Secretariat (UN/DESA), 2013.

"Cervical Cancer." World Health Organization, World Health Organization, 12 Sept. 2018, www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/.

"Cervical Cancer Risk Factors: Pregnancy, HPV, Others — CTCA." CancerCenter.com, Cancer Treatment Centers of America, 1 Jan. 1AD, www.cancercenter.com/cervical-cancer/risk-factors/?invsrc=non$_b$$randed_p$$aid_s$$earch_g$$oogle_t$$p$$ur$ = $prospectingt_m$$ed$ = $onlinet_c$$h$ = $paid_s$$earcht_a$$dg$ = $58244574422t_c$$tv$ = $305604148558t_m$$tp$ = $et_p$$os$ = $1t1t_p$$lc$ = $kwd-342637086234t_s$$i$ = $googlet_t$$ac$ = $nonet_c$$on$ = $non_b$$randt_b$$ud$ = $corpt_d$ = $ct_t$$ar$ = $non_t$$argetedt_a$$ud$ = $anykxconfid$ = $s8ymtai82dsk$ id = $trackeridgclid$ = $EAIaIQobChMIzsONq7z83gIVj4dpCh3Epg1iEAAYASAAEgL3FvD$ $_B$$wEgclsrc$ = $aw.ds$.

Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.

Murillo, Raúl, et al. "Cervical Cancer in Central and South America: Burden of Disease and Status of Disease Control." Cancer Epidemiology, vol. 44, 30 Sept. 2016, pp. S121–S130., doi:10.1016/j.canep.2016.07.015.

"NCI Dictionary of Cancer Terms - Schiller Test ." National Cancer Institute, National Cancer Institute, www.cancer.gov/publications/dictionaries/cancer-terms/def/schiller-test.

"Pap and HPV Testing." National Cancer Institute, National Cancer Institute, www.cancer.gov/types/cervical/pap-hpv-testing-fact-sheet.

"Cervical Cancer Deaths per 100,000 Women." Henry J Kaiser Family Foundation, Kaiser Family Foundation, https://www.kff.org/other/state-indicator/cervical-cancer-death-rate/?currentTimeframe=0sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D.

"Tests for Cervical Cancer — Diagnosed With Cervical Cancer." American Cancer Society, American Cancer Society, 5 Dec. 2016, www.cancer.org/cancer/cervical-cancer/detection-diagnosis-staging/how-diagnosed.html.

Saito, Takaya, and Marc Rehmsmeier. "A ROC Curve of a Random Classifier." Wordpress, Wordpress, 2015, classeval.files.wordpress.com/2015/06/a-roc-curve-of-a-random-classifier.png?w=280h=273.

Tom Fawcett. "Learning from Imbalanced Classes." Silicon Valley Data Science, 25 Aug. 2016, www.svds.com/learning-imbalanced-classes/.

"Venezuela Human Papillomavirus and Related Cancers, Fact Sheet 2018." ICO/IARC Information Centre on HPV and Cancer, 10 Oct. 2018.

"What Are the Risk Factors for Cervical Cancer?" American Cancer Society, The American Cancer Society Medical and Editorial Content Team, 1 Nov. 2017, www.cancer.org/cancer/cervical-cancer/causes-risks-prevention/risk-factors.htmlwritten$_b$$y$.