Jasmine Clark
Therapeutic Sciences Graduate Program
DATA1030
December 10, 2023

Cervical Cancer Risk Data Analysis

GitHub Repository

https://github.com/jassiemae2/data1030project

Introduction

Motivation

Cervical cancer is the fourth leading cause of death in women globally with approximately 600,000 cases diagnosed each year[1]. Approximately 50% of cases occur in women between the ages of 35 and 54 and HPV infection is associated with the majority of cervical cancer cases[2]. In adults, the biggest risk factor for HPV is sexual activity with an infected individual. In the U.S. cervical cancer mortality rates have been declining steadily since the development of early screening and detection with the Pap test[3]. Several factors increase the risk of developing cervical cancer such as the age of the first sexual encounter, the number of partners you have had, smoking, HPV infection, prolonged use of oral contraception, etc[2]. Understanding which risk factors have more predictive power on determining your risk for cervical cancer is integral to assisting with early detection and treatment of cervical cancer.

Dataset

The Cervical Cancer Risk Analysis dataset is a classification dataset and was acquired from the UCI machine learning repository[4]. It focuses on identifying different risk factors that lead to the prediction of cervical cancer. The data was collected from the Hospital Universitario de Caracas in Caracas, Venezuela. The target variable is Biopsy with Class 0 meaning no indication of cervical cancer and Class 1 meaning an indication of cervical cancer. There are 858 patients and 36 feature columns. This is an iid dataset and has missing values.

Exploratory Data Analysis

The dataset was loaded and viewed to check for missing values. The fraction of points with missing values was 0.221. A heatmap of all the features was created and it was observed that Schiller had the highest correlation with our target variable, Biopsy (Figures 1 and 2). The Schiller test is a medical test in which an iodine solution is applied to the cervix to diagnose cervical cancer. As another predictive exam for cervical cancer diagnosis, it makes sense as to why it has such a high correlation to the target variable, 'Biopsy'.

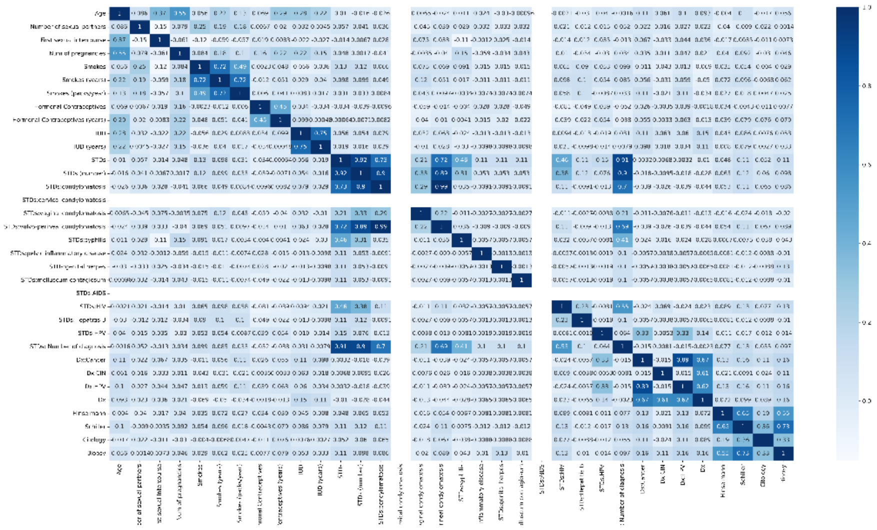

Figure 1: Heatmap of Dataset features

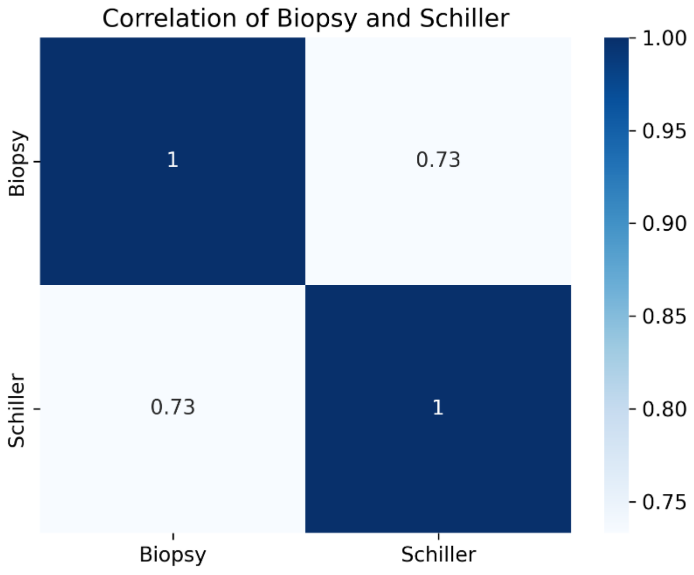

Figure 2: Heatmap of Biopsy and Schiller Correlation

It was also observed that age was an interesting factor in the distribution of biopsy results in Figure 3. It was shown that the incidence of a positive biopsy was highest between the ages of 38 and 52 which is consistent with the age range for most cases of a cervical cancer diagnosis.

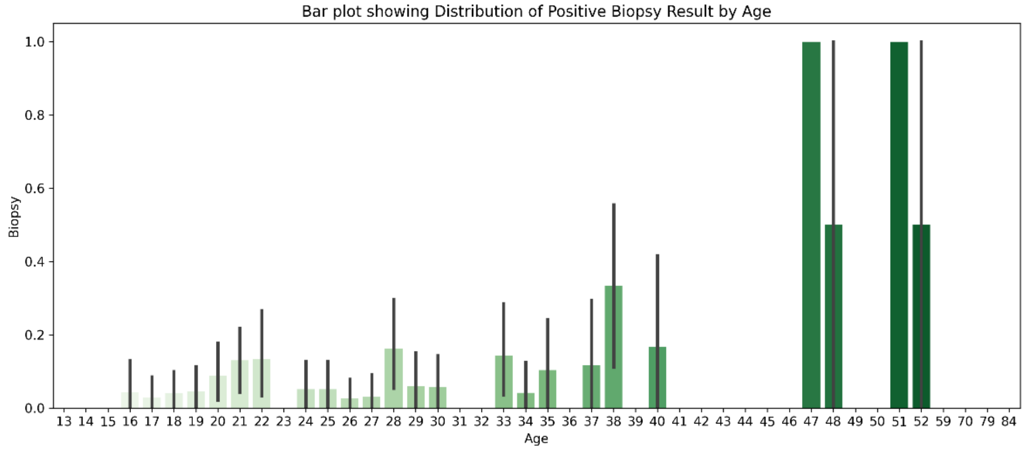

Figure 3: Bar plot of distribution of biopsy results by age

The baseline recall score of the dataset was calculated to be 0.064. In Figure 4, the target variable, ‘Biopsy’, was observed to be heavily imbalanced with a count of 803 points in Class 0 and 55 points in Class 1.

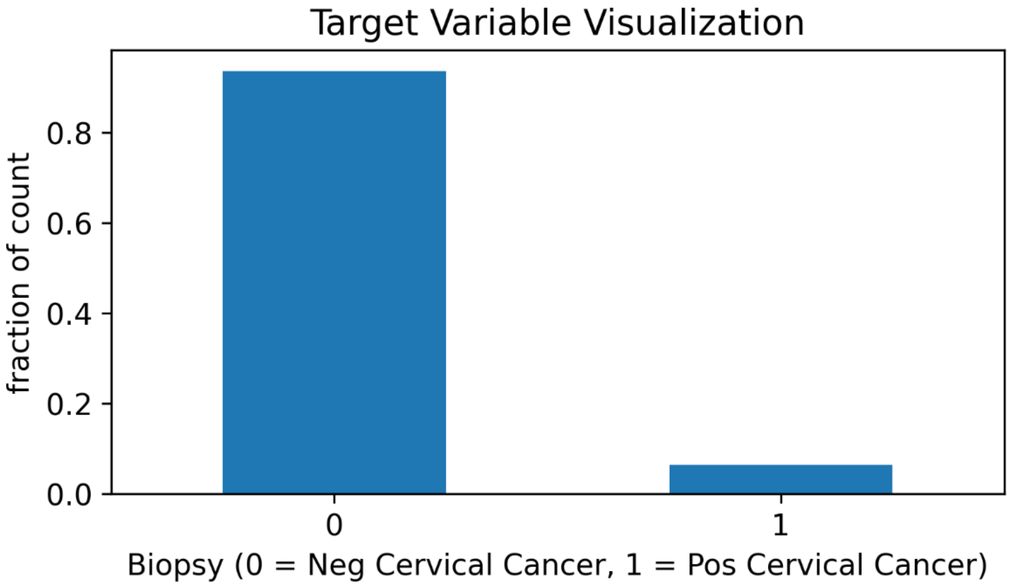

Figure 4: Target variable, 'Biopsy' visualization

Methods

The data was split using sklearn’s train_test_split package and was split into 80% training and 20% testing. Once that was completed, the training set was further split using a StratifiedKFold with four folds split into 60% training and 20% validation finalizing the split to a 60-20-20 train, test, and validation split. A StratifiedKFold split was utilized instead of a regular KFold split because it takes into consideration class imbalance when dividing the data into the folds. This was important because the target variable was heavily imbalanced. There was a mix of continuous and categorical features within the dataset so for preprocessing, a pipeline for the continuous features was created with StandardScalar. For the categorical features, a pipeline with IterativeImputer to impute any missing values within the data and OneHotEncoder was used. 4 models were executed on the data: Logistic Regression with L1 and L2 penalty, Random Forest Classifier, and Support Vector Classifier. These were all executed with GridSearchCV to tune hyperparameters as a means for cross-validation.

Table 1: Machine Learning Models and their hyperparameters

Machine Learning Model	Hyperparameters	Values used
Logistic Regression(L2 Regularization)	C penalty	[0.0001, 0.001, 0.01, 0.1, 1, 10, 100,1000] ‘l2’
Logistic Regression(L1 Regularization)	C penalty	[0.0001, 0.001, 0.01, 0.1, 1, 10, 100,1000] ‘l1’
Random Forest Classifier	n_estimators max_depth	[3, 10] [4, 5, 6, 7, 8, 9, 10]
Support Vector Classification	gamma C	[0.001, 0.1, 10, 1000, 100000], [0.1, 1, 10]

The models were trained over 10 different random states. This helps to ensure proper hyperparameter tuning and to avoid overfitting. The evaluation metric used for this analysis was recall. Recall was chosen because of the small target class and because of the class imbalance in the target variable. Recall was also chosen because it was important to correctly detect positive samples aka positive biopsy (indication of cervical cancer).

Results

To determine the best model, mean recall was calculated. Based on that score, the LogisticRegression with L1 penalty model was the best performing at 0.679 +/- 0.170. All models except the untuned SVC model performed better than the baseline of 0.064 and overall, the worst-performing model was the SVC model. This is shown in Figure 5. The baseline recall was calculated by dividing the number of Class 1 points by the number of all data points.

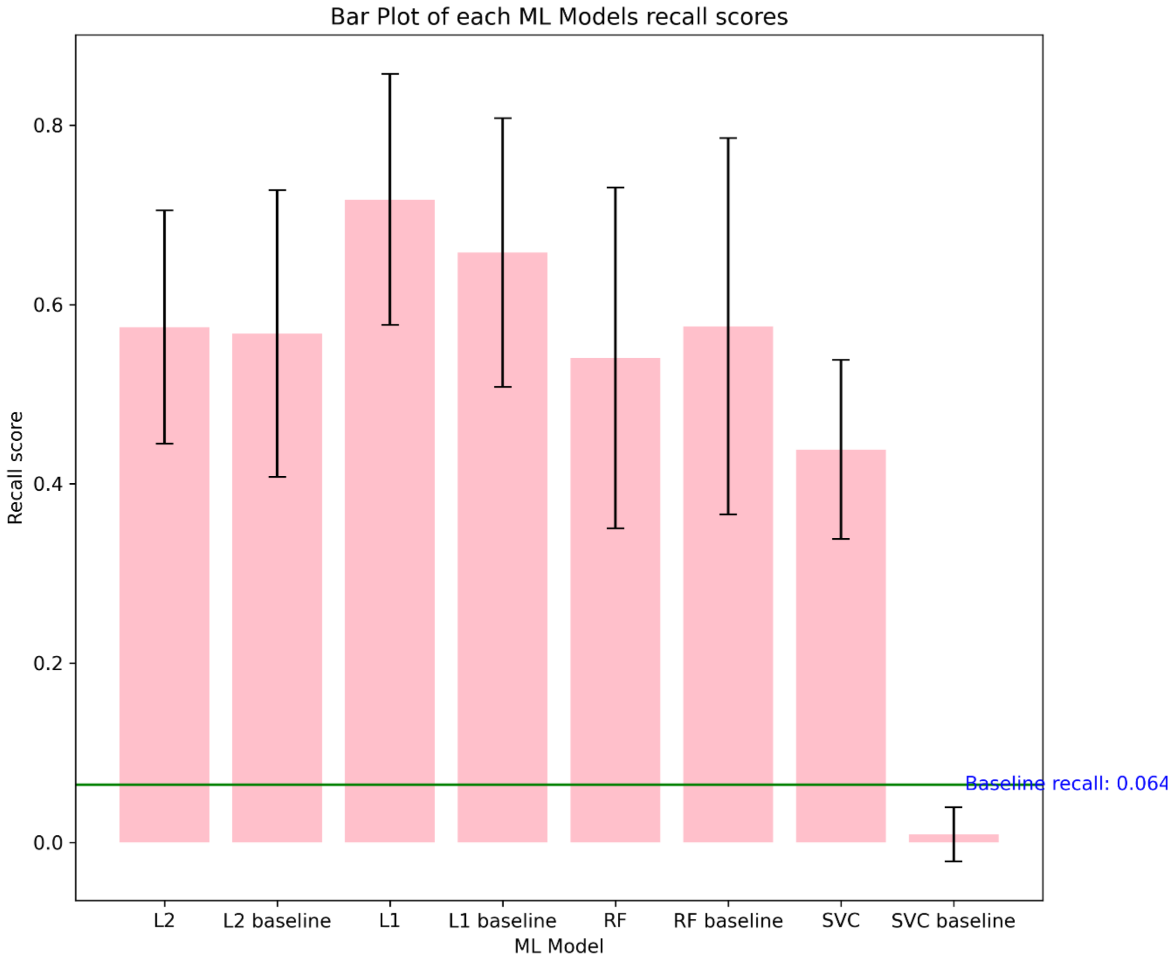

Figure 5: Bar plot of ML models recall scores

The confusion matrixes for all the models were generated and it was observed that they all did well in predicting the instances of Class 0, but poorly in properly predicting Class 1. This could be due to the class imbalance of the target variable.

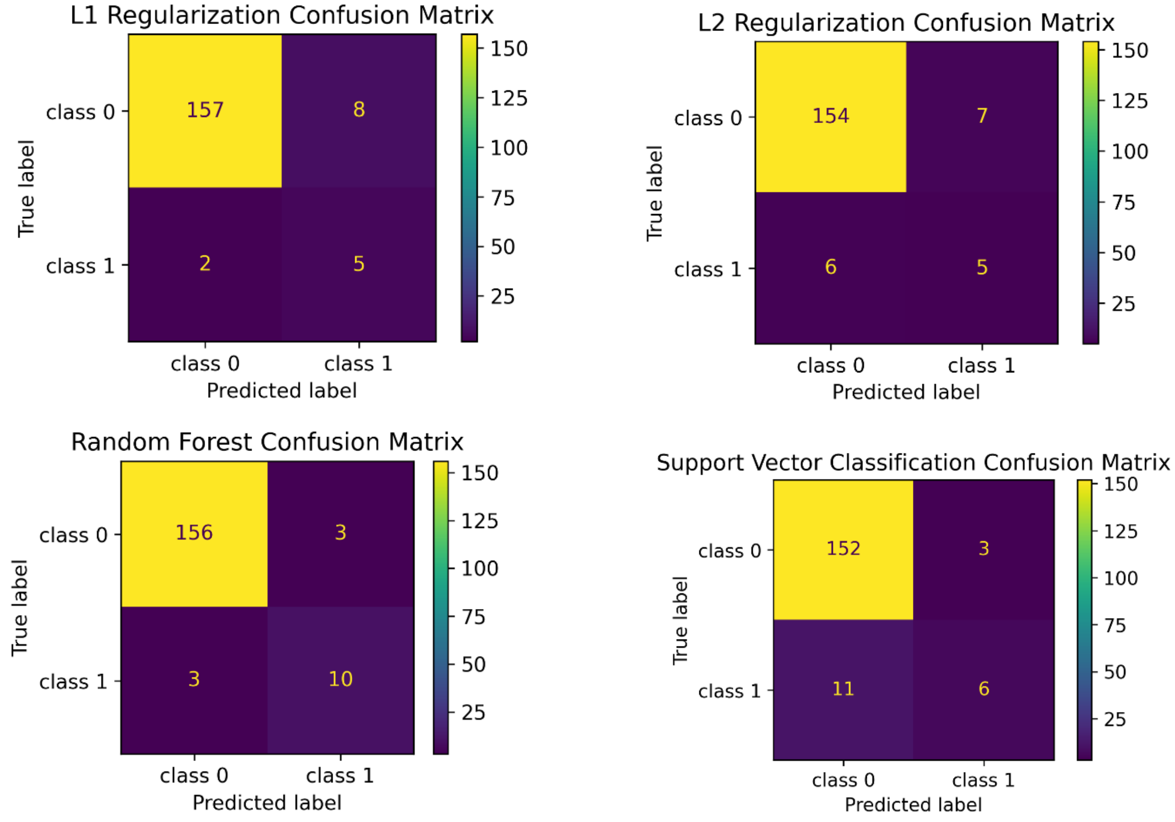

Figure 6: ML Models Confusion Matrixes

To determine which features contributed the most to the models' prediction, coefficient values were used. The top 10 most important features were determined. The most important feature for the logistic regression model with L1 penalty was cat_Schiller_0 which has a negative correlation with the prediction. For the LogisticRegression model with L2 penalty, the most important feature was the cat_Schiller_1 and it has a positive correlation with the prediction.

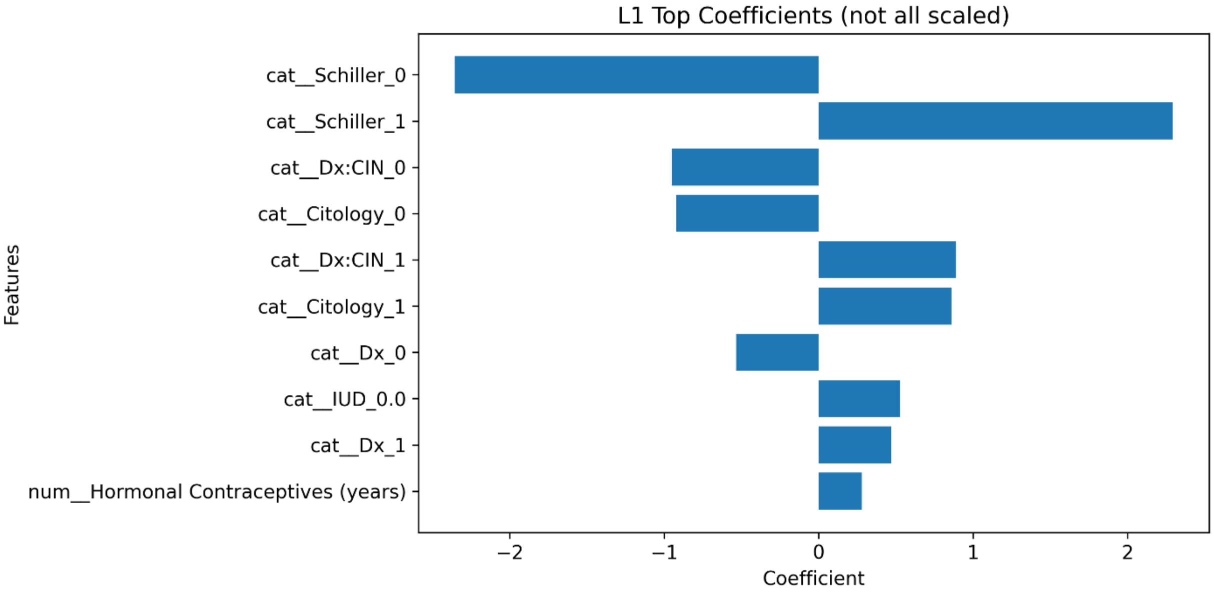

Figure 7: LogisticRegression with L1 penalty top 10 most important features based on coefficient

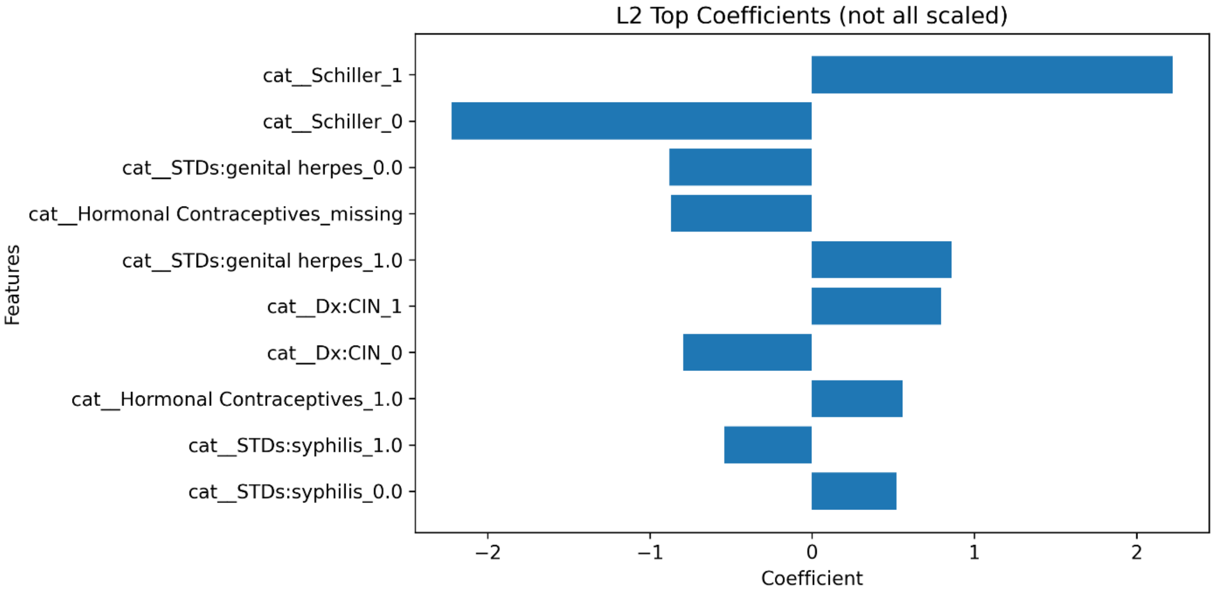

Figure 8: LogisticRegression with L2 penalty top 10 most important features based on coefficient

Outlook

To improve the models I would deal with the missing data differently. Because this is a medical dataset using random permutation instead of imputation works to help avoid bias in the model training, reduce overfitting, and ensure class balancing. If I had more time to work with this dataset I would also explore other target variables. There were four target variables identified and reported: Biopsy, Schiller, Hinselman, and Citology. I would use a combination of Biopsy and Schiller or a combination of all four as the target variable 'Cancer' to determine model performance. For diagnosing different diseases several tests are taken to confirm a diagnosis and it may be necessary for all the preliminary examinations to be considered in the target variable for the models to make better predictions on the target class 1.

References

1. Shrestha, A. D., Neupane, D., Vedsted, P., & Kallestrup, P. (2018). Cervical Cancer Prevalence, Incidence, and Mortality in Low and Middle Income Countries: A Systematic Review. *Asian Pacific journal of cancer prevention: APJCP*, *19*(2), 319–324. https://doi.org/10.22034/APJCP.2018.19.2.319
2. Johnson CA, James D, Marzan A, Armaos M. Cervical Cancer: An Overview of Pathophysiology and Management. Semin Oncol Nurs. 2019 Apr;35(2):166-174. doi: 10.1016/j.soncn.2019.02.003. Epub 2019 Mar 14. PMID: 30878194.
3. Buskwofie A, David-West G, Clare CA. A Review of Cervical Cancer: Incidence and Disparities. J Natl Med Assoc. 2020 Apr;112(2):229-232. doi: 10.1016/j.jnma.2020.03.002. Epub 2020 Apr 8. PMID: 32278478.
4. Fernandes,Kelwin, Cardoso,Jaime, and Fernandes,Jessica. (2017). Cervical cancer (Risk Factors). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z310.