BIOPSY PREDICTION OF CERVICAL CANCER

Samuel Aboagye[1], Armand Heydarian[1] and Swetha Kalla[1]

[1]Department of Data Science, The George Washington University, Washington, D.C., USA

Abstract

Cervical Cancer is a leading cause of death in women.In the United States, there are over eleven thousand new diagnoses of cervical cancer each year. . The stated purpose of this project is to ultimately find a link between certain medical conditions that we use as variables, and the increased risk of being diagnosed with cervical cancer.This project aims to evaluate and compare different classifiers on the Cervical Cancer dataset. There are 4 targets, but for the purposes of this project only one of them which is the Biopsy will be classified. The target classes are boolean but have been encoded. Any target with a value of (1) is encoded for True which represents Malignant cervical cancer and any target with a value of (0) represents benign cervical cancer. A Logistic Regression, SVM, Decision Tree and Random Forest were trained on the Data. Logistic Regression was our choice of classifier because it had the best combination of Classification Metrics.

Keywords: Cervical Cancer, human papillomavirus, Biopsy, Age,Random Forest, Logistic Regression

---

1. Introduction

Cervical cancer is normally diagnosed through the performance of Biopsy by a trained physician. Biopsy is known to be one of the best ways to confirm the presence of cervical cancer as being benign or malignant. Automating the process of diagnosing the disease could be helpful in the early screening and detection of cervical cancer. From the literature, certain features are stronger predictors of the presence of cervical cancer. Through feature selection using Random Forest, the most important features were confirmed by the research. The 10 most important features were Age, Number of Sexual partners, IUD, Number of Pregnancies, Smoking, first Sexual Intercourse, Hormonal Contraceptives.

2. Background

The dataset was collected from the Hospital Universitario de Caracas in Caracas, Venezuela. The data set we utilized had compiled data based on the medical records and demographics of 858 patients, although there is some missing data points since some patients refused to disclosure parts of their medical history [3]. The data set covers a range of information from the patient's age, whether or not they smoke, to the number of pregnancies and sexual partners they've had in the past. There's also an extensive list covering a range of sexually transmitted diseases that a patient may have had been diagnosed with in the past. All these attributes are believed to be possible factors in increasing a person's risk of having invasive cervical cancer.

3. Machine Learning Approaches

In the previous publications, people have predicted the risk assessment of cervical cancer using various supervised machine learning models. Our project was more oriented towards addressing the potential issues that could rise without proper pre-processing. To learn the effect of pre-processing, we had to run various models on the raw data, and see as to what preprocessing methods could be used to address the underlying issues.

3.1 Method

Various data mining methods were employed in this project. The data was first collected, processed, modeled and then evaluated.

3.2 Data Preprocessing

Data cleaning was done by removing columns with more than 50% of missing data, as the inference from this data will be less efficient. These features were 'STDs: Time since first diagnosis' and 'STDs: Time since last diagnosis'. For the observations with barely missing values, the imputation was done using the mean of the respective column. There were 4 targets in total and due to time constraints we decided to focus our modelling on just one of them which was Biopsy. The choice of this Target was due to the fact that several research papers pointed to it being one of the most reliable ways of screening for the presence of cervical cancer.

We had to upsample the observations for the 'with cancer' class to increase the sampling rate to keep up with the heaviness of the 'without cancer' class. After the upsampling, the data has risen to 623 observations in each class.

The following graph shows the importance of all the features present within the dataset.

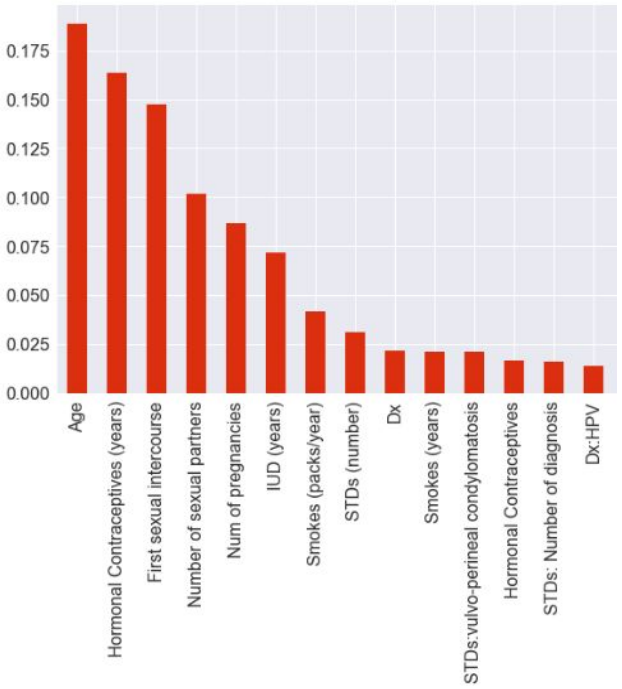

figure 3.2 (a)

The following curve shows the explained variance ratio vs number of features from the Principal Component Analysis.

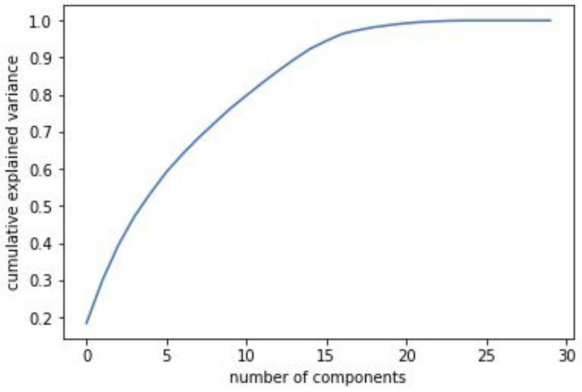

figure 3.2(b)

As we can see from the figure above, nearly 18 features are able to narrate the entire dataset. This reduction in dimensionality prevents overfitting and removes all the redundant and correlated features.

3.3 Models:

The performance of different machine learning algorithms on the cervical cancer dataset before and after the preprocessing is explored and compared, to study the effect of

preprocessing the data, and to use the best machine learning algorithm for the risk assessment of cervical cancer. Performance measurement of the supervised algorithms. Here, five supervised learning classifiers are trained on the dataset and their prediction accuracy for the test data were measured. This process is mainly divided into two stages where the algorithms were run:

3.3.1 Prior to the preprocessing of the data

3.3.2 After the preprocessing of the data

3.3.1 Models before preprocessing

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is used to curb these imperfections. But prior to that, we ran several machine learning algorithms on the raw data to study and compare the effects of preprocessing, and as to how much information the data is providing, with regards to the biopsy results, in its entirety.

Given that over 90% of the data had one or more missing values, we had to get rid of observations that are in short supply of data, since running the models with the missing values will be very difficult to deal with. As a result of this, the data has vastly shrunk from over 850 observations initially to 59 in total, of which only 9 patients are diagnosed with cervical cancer. The training test and the test set were split in such a way that the test set consisted of 30 percent of the observations (=17.7 ≈ 18), of which only 3 women were diagnosed with cancer. The various models that were run on this set of data were mentioned below.

The accuracy of the models run below are dependent on the quality of the raw data, and how much useful information can be gained directly from it. The table below demonstrates the comparison of accuracies of all the models run prior to any cleaning methodology:

Model	Accuracy
Decision Tree	94.44
Random Forest	100
Support Vector Machine	88.88
*K-Nearest Neighbors	83.33
Naïve Bayes	83.33

*The accuracy mentioned for the KNN model was the accuracy for the model when n = $\sqrt{59} \approx 7$

The high efficiency observed in the above models can be attributed to the lower number of observations in the test model and also the fact that most of the data consisted of women who weren't diagnosed with cervical cancer. Given this class imbalance, stating the risk of cervical cancer for every observation as zero would itself yield an accuracy of 88 percent. This tells us that accuracy is not the best measure to assess these models.

So, now we look at the other metrics from which we might be able to derive some useful information. The tables below show us the metrics for all the models.

Models	Precision	Recall	F1 Score
Decision Tree	1	0.93	0.97
Random Forest	0.83	1.00	0.91
Support Vector Machine	0.88	1.00	0.94
K-Nearest Neighbors	0.83	1	0.91
Naïve Bayes	0.88	0.93	0.90

For women without cervical cancer

Models	Precision	Recall	F1 Score
Decision Tree	0.75	1	0.86
Random Forest	0	0	0
Support Vector Machine	1	0.33	0.50
K-Nearest Neighbors	0	0	0
Naïve Bayes	0.50	0.33	0.40

For women with cervical cancer

From the above tables, we can see that although random forest and k nearest neighbors have an accuracy of 83 percent, they only seem to have accurate predictions when the patient is not diagnosed with the cervical cancer.

3.3.2 Models after preprocessing

After the pre-processing of the data, we had to re-run various supervised learning models to check for the impact of preprocessing, and also to see how well each model was able to learn from the data. The following models were run in this stage:

1. Random Forest
2. Support Vector Machine
3. Logistic Regression

3.4 Model Evaluation

Given the rareness of the disease, the healthcare data is deemed to be imbalanced. So simply considering the accuracy would not produce anything informative. So we had to look at the f1 score of the overall model, for both the classes, with and without the cancer. So we had to choose the support vector machine model as it could potentially reduce the false negative rate.

4. Results

Even though the overall efficiency of the model has decreased before and after the preprocessing, we can deduce that there was an overall increase in other metrics.

The following table gives a summary of the metrics of the machine learning models we've run after the preprocessing:

```
The accuracy of SVC is: 0.7661691542288557


[0 1]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.81      0.87       187
           1       0.05      0.14      0.08        14

   micro avg       0.77      0.77      0.77       201
   macro avg       0.49      0.48      0.47       201
weighted avg       0.87      0.77      0.81       201



array([[152,  35],
       [ 12,   2]], dtype=int64)
```

SVM was chosen as our final model because it had the best measure of accuracy and performance.

5. Conclusion

Given that we've only used one target variable (Biopsy), we would like to further delve into this project and see how other targets would perform on similar machine learning models, and if there is a way we can produce an ensembling model that could much lower the false negative rate.

References

[1]Center for disease Control and Prevention (CDC).

https://www.cdc.gov/cancer/cervical/basic_info/risk_factors.htm

[2]Source: Kelwin Fernandes (kafc at inINESC TEC & FEUP, Porto, Portugal. Jaime S. Cardoso - INESC TEC & FEUP, Porto, Portugal. Jessica Fernandes - Universidad Central de Venezuela, Caracas, Venezuela.

[3] American Cancer Society:

https://www.cancer.org/content/dam/CRC/PDF/Public/8599.00.pdf

[4]Analytics Vidhya:https://www.analyticsvidhya.com/blog/category/machine-learning/

https://www.analyticsvidhya.com/blog/category/machine-learning/