Predicting Cervical Cancer Using Machine Learning[1]

Matthew Falcione [1]

[1] School of Biomedical Engineering, Drexel University, USA

Course : BMES547 Machine Learning in Biomedical Applications
Instructor: Ahmet Sacan
Date : 2021-02-28

ABSTRACT

Cervical cancer is one of the most common causes of death from cancer in the world and especially in developing countries. The objective of this project is to understand medical history information can be used to predict a cervical cancer diagnosis. This is useful because in areas where resources are scarce, this will allow diagnostic resources to be allocated to women at high risk. The dataset used in this study was obtained from 856 patients in Hospital Universitario de Caracas in Caracas, Venezuela. Following data post-processing, methods include t-SNE visualization, tree-based and univariate feature selection, and neural network, random forest, and XGBoost classification. t-SNE visualization showed no discernable patterns or splits in the data for any of the screening tests or the class target value, meaning there are likely multiple factors needed to make an accurate prediction of the target value. The random forest, XGBoost, and neural network classification methods all resulted in similar true negative values while the XGBoost method had the highest true positive value. The true negative rate was chosen as the primary evaluation metric because Type II error or predicting that an individual will not be diagnosed with cervical cancer when in fact they will, is far more dangerous than its Type I counterpart. In the future, collecting data on a wider range of patients would help generalize results to a larger population. Building models on subsets of the data would also be useful so that it can be applied to individuals for whom some screening options are not available.

1 INTRODUCTION

Cervical cancer is a type of cancer that occurs in the cells of the cervix. The cervix is the lower part of the uterus that connects to the vagina. Cervical cancer is the fourth most common cause of cancer and the fourth most common cause of death from cancer in women worldwide. Every year, approximately 570,000 women are diagnosed, and 311,000 women die from cervical cancer (Cervical Cancer Overview | Guide To Cervical Cancer, 2021).

There are several factors that have been observed to increase a woman's risk of contracting cervical cancer. These risk factors include Human Papillomavirus (HPV) infection, sexual history including number of sexual partners, smoking history, and long-term use of oral contraceptives (Basic Information about Cervical Cancer, 2021). These risk factors and others will be studied for predictive capability of cervical cancer in this study.

The early stages of cervical cancer are often symptom-free (Cervical Cancer Research, 2020). Furthermore, diagnosis relies on early detecting using Pap tests. These diagnostic tests are often not possible to do regularly in third world countries where the standard of care is lower. While wealthy patients in first world countries have the resources to obtain annual diagnostic tests, such opportunities are often not present for those of lower socioeconomic status. This is supported by the fact that 80% of cervical cancer cases occur in the developing world, and that cervical cancer is the leading cause of death from cancer among women in developing countries (Sherris, 2001).

The objective of this project is to understand which of a woman's demographic and medical history information can be used to predict a cervical cancer diagnosis. The intended target users are doctors, in order to help aid their understanding of which patients are at high risk of developing cervical cancer, and women, in order to help them understand whether they are at high risk. If successful, and if results can be corroborated using larger and more extensive datasets, this will allow more powerful diagnostic resources to be allocated to high-risk women in locations where resources are

[1] Avoid using identical title to any other publication. Scholarly articles are supposed to be unique identifiers and you do not want your report to appear as a version of the original appear on search engines. If your project is based on a paper, use a title that reflects what you did.

scarce. Focusing resources on high-risk populations will allow for more effective diagnosis. In the long-term, this will hopefully decrease the mortality rate of cervical cancer in developing countries.

2 DATASET

This dataset was published on Kaggle. The data was collected at Hospital Universitario de Caracas in Caracas, Venezuela. There was no specific experiment that took place. Rather, demographic information and historic medical records were obtained from the hospital and patient interviews. The dataset has 856 samples, 35 input attributes, and 1 target attribute. The input attributes, which consist of boolean and continuous values, include age, number of sexual partners, number of pregnancies, smokes, use of hormonal contraceptives, STDs, previous cancer diagnoses, HPV diagnosis, Hinselmann colonoscopy, and others. The target attribute is biopsy, which was determined by extraction and examination of cervix tissue to determine the presence of cervical cancer.

3 METHODS

The code was written in Python and developed on Windows OS machines but should be cross-platform compatible. The following Python modules and versions, which are available in the requirements.txt file, were used in the analysis and preprocessing steps: scikit-learn=0.24.1, xgboost=0.90, numpy=1.20.0, seaborn=0.11.1, matplotlib=3.3.4, and pandas=1.2.1.

Preprocessing started with cleaning the data and replacing missing values. First, the missing data, which was marked with a '?', was replaced with blank values. Next, each Boolean and float STD-related attribute was set to False and 0, respectively, if the `STD: Number of diagnosis` attribute was 0 for that sample. The samples with missing smoking, hormonal contraceptive, and IUD-related fields were dropped. Additionally, the `Dx` attribute was dropped since no descriptive data could be found for it. This brought the total number of samples and attributes to 730 and 34, respectively, with the target class value being positive for around 10% of the samples. For the remaining float attribute columns with missing values, the median of each column was inserted. Lastly, True and False values were replaced with 1 and 0, respectively, and the class target values, attribute values, and attribute names were extracted from the dataset for visualization and classification.

The remaining samples and the three most correlated attributes to the target class (the screening test – Hinselmann, Schiller, Citology) and the target class (Biopsy) were visualized using t-distributed Stochastic Neighbor Embedding (t-SNE) for the scikit-learn module. The visualization projects the high-dimensional data to a low-dimensional space while preserving the inter-sample distance seen in the higher-dimension.

Next, both univariate (filter method) and tree-based (embedded method) feature selection were performed to select the top 10 attributes that would best predict the target attribute values. The univariate method used a χ^2 test statistic, and the tree-based method used a Random Forest with 100 trees.

To streamline the model evaluation, a classification evaluation metric function was created. The function arguments included a classification model, matrix of input values for each samples and attribute, true target class values, a cross-validation function, and an identifier of type of model. It then generated predicted target values from a repeated stratified k-fold cross-validation with 5 splits and 10 repeats. Using the true and predicted values, a confusion matrix was created, and the true positive, true negative, false positive, and false negative values were extracted. A confusion matrix plot was also generated with the appropriate 'Cancer' and 'No Cancer' labels. Lastly, the evaluation metrics for TNR (specificity), TPR (sensitivity/recall), PPV (precision), error rate, accuracy, F1 Score, and AUC-ROC were calculated and averaged across all cross-validation sets.

For classification, Random Forest, XGBoost, neural network models were run on each subset of attributes chosen by the feature selection models. The Random Forest and XGBoost models were run with the default parameters provided by the scikit-learn and xgboost modules, respectively, and the neural network was initiated as a scikit-learn multi-layer perceptron (MLP) classifier with 5 hidden units and 1 hidden layer.

4 EXPERIMENTS AND RESULTS

For the t-SNE visualizations, shown in Error! Reference source not found., there are no discernable patterns or splits in the data for any of the screening tests or the class target value, meaning there are likely multiple factors needed to make an accurate prediction of the target value. The silhouette scores for each plot, shown in the titles, were all nearly 0 which indicates that the samples are on or very close to the decision boundary between two neighboring clusters when split by each of the individual attributes.

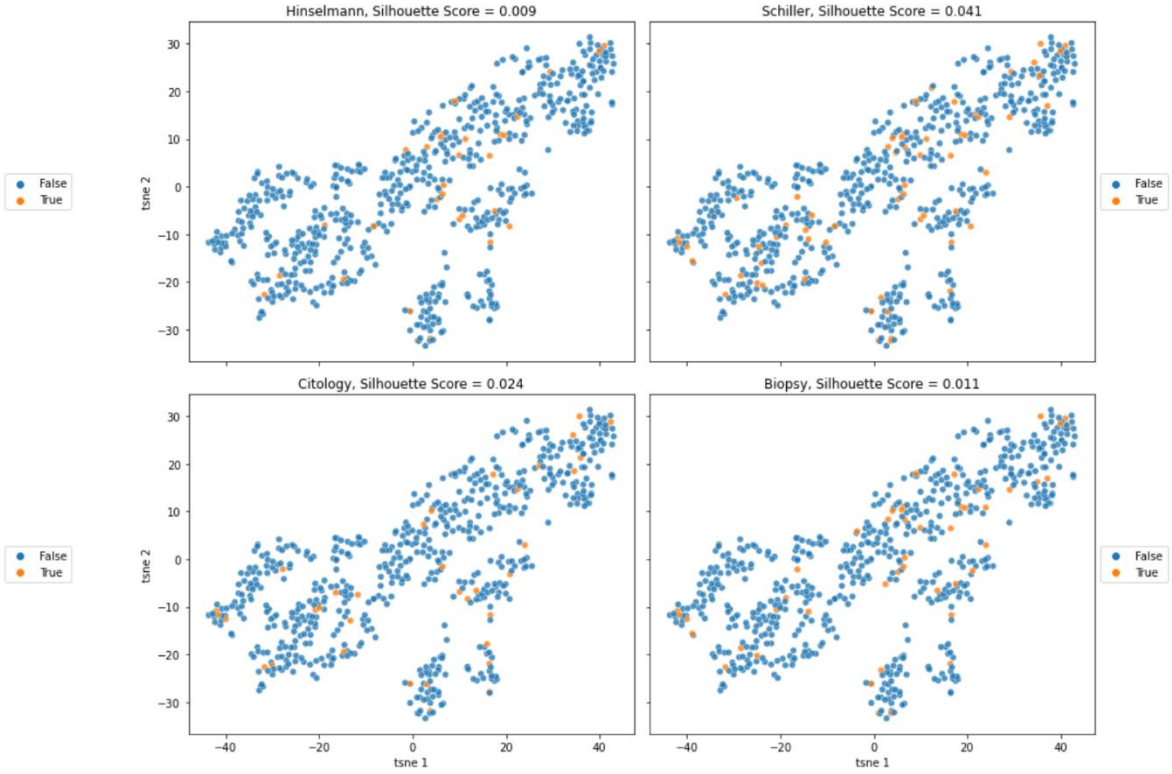

Figure 1. t-SNE Visualization of Hinselmann, Schiller, Citology, and Biopsy. A t-SNE visualization was created for the top 3 most correlated attributes from the preliminary analysis on the dataset (the screening tests - Hinselmann, Schiller, Citology) and the class target value, meaning there are likely multiple factors needed to make an accurate prediction of the target value.

The top 10 attributes that would best predict the target attribute values, as selected by the univariate and tree-based methods are shown in **Figure 3** and **Figure 2**, respectively. The univariate method used a χ^2 test statistic which maintained the sparse nature of the original data, and the tree-based method used an impurity-based ranking which captured the interactions between features in the dataset.

```
10 most important features:
---------------------------
1. Schiller
2. Hinselmann
3. Age
4. Hormonal Contraceptives (years)
5. First sexual intercourse
6. Citology
7. Number of sexual partners
8. Num of pregnancies
9. Smokes (years)
10. Dx:CIN
```

Figure 3. Tree-based feature selection. Top 10 most important attributes showing the best attributes to predict cervical cancer.

```
Top 10 attributes and k-best scores:
------------------------------------
                              Specs       Score
32                         Schiller  353.674831
31                       Hinselmann  199.711268
33                         Citology   77.635002
5                     Smokes (years)   51.967208
8   Hormonal Contraceptives (years)   43.009505
30                            Dx:HPV   22.925323
28                         Dx:Cancer   20.963675
12                     STDs (number)   13.688781
19               STDs:genital herpes   13.313725
27  STDs: Time since last diagnosis   11.478385
```

Figure 2. Univariate feature selection. Top 10 attributes and k-best scores showing the best attributes to predict cervical cancer.

Example results, shown in **Figure 4**, display the confusion matrix and average evaluation metrics for the XGBoost model using the tree-based selected features. The figure shows that 661 samples were true negative, 18 samples were false negative, 15 samples were false positive, and 36 samples were true positive. Little difference was seen between the model evaluation metrics for the tree-based and univariate selected feature subsets. The results were also close when isolated for different

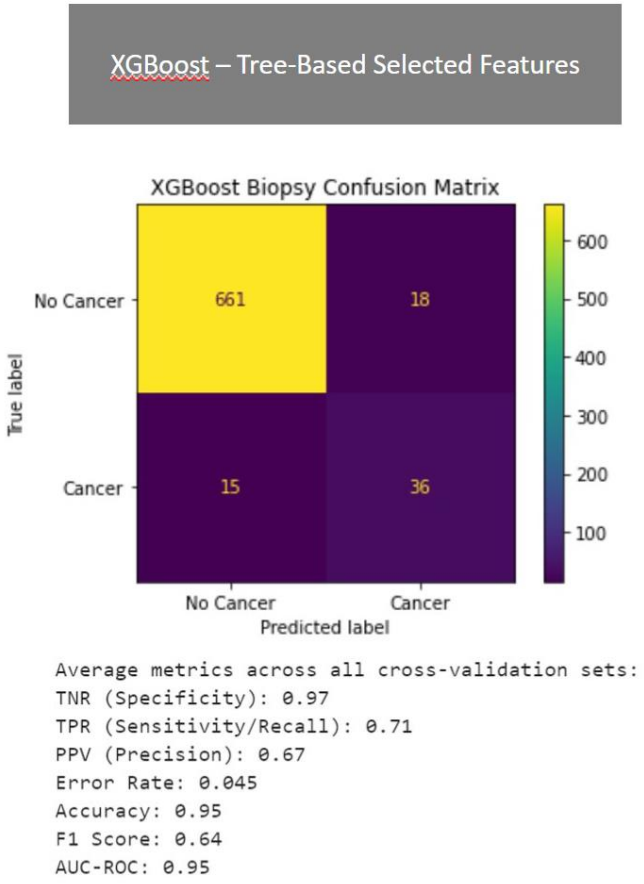

Figure 4. XGBoost classifier confusion matrix and model evaluation metrics. A confusion matrix of the predicted and true cancer labels of the biopsy target variable using the tree-based XGBoost method. 661 samples were true negative, 18 samples were false negative, 15 samples were false positive, and 36 samples were true positive. It had a TNR of 97%, TPR of 71%, PPV of 67%, error rate of 4.5%, accuracy of 95%, F1 score of 0.64, and an AUC-ROC value of 0.95.

classification approaches with less than a 5% difference between all metrics except TPR which was highest for XGBoost. On average, the TNR generally sat around 97% or 98%, the TPR was between 55% and 71%, PPV was between 58% and 73%, error rate ranged from 4.5% to 6%, accuracy was around 94% to 95%, F1 Score ranged from 0.62 to 0.65, and the AUC-ROC was between 0.91 and 0.95.

5 DISCUSSION

One of the main discussion points is the team's decision of evaluation metric between false positives and false negatives. False positives, also known as Type I error, can be defined as predicting that an individual will be diagnosed with cervical cancer when in fact they will not. The cost of Type I error is wasted resources since the individual will pursue more powerful diagnostic tests when they are not necessary. False negatives, also known as Type II error, can be defined as predicting that an individual will not be diagnosed with cervical cancer when in fact they will. The cost of Type II error is the loss of lives since an individual will not pursue further diagnostics and treatment when they in fact should. Type II error is far more dangerous in this case. The team thus made the decision to optimize the machine learning algorithm such that it maximizes the true negative rate. Six methodologies were run and compared – univariate and tree-based for random forest, XGBoost, and neural network. These methodologies had very similar true negative rates, while the XGBoost method had the best true positive rate.

There are a few limitations of this study. First, the dataset includes a very specific group of patients that have a particular risk and socioeconomic factors. Second, some samples had missing data. Finally, the classes had large imbalances in terms of numbers. This was mitigated by doing a shuffled, randomized, stratified k-fold cross-validation.

In terms of follow-up studies, the most beneficial would be to collect data on patients of a larger scope so that results can be stated with larger statistical significance and can be generalized to larger populations. It could also be interesting to build models based on subsets of the data. Since some data are missing, this would allow us to answer the question, "What happens if a patient doesn't have all screening options available?" Finally, an alternative methodology, such as grid search or other hyper-parameter optimization techniques, could be used to corroborate results found in this study.

6 REFERENCES

Basic Information about Cervical Cancer. (2021). Retrieved from CDC.Gov: https://www.cdc.gov/cancer/cervical/basic_info/

Cervical Cancer Overview | Guide To Cervical Cancer. (2021). Retrieved from Cancer.Org: https://www.cancer.org/cancer/cervical-cancer.html

Cervical Cancer Research. (2020). Retrieved from National Cancer Institute: https://www.cancer.gov/types/cervical/research

Sherris, J. (2001). Cervical Cancer In The Developing World. *Western Journal of Medicine, vol 175, no. 4*, 231-233.