Cervical Cancer Risk Factors

Angelina Poole

Santa Clara University
Department of Computer Engineering
COEN 140: Machine Learning and Data Mining
Final Report

Abstract **–** Using machine learning techniques, I attempted to classify a patient if a patient had a high risk of cervical cancer based on their risk factors. Using the Cervical Cancer Risk Factors dataset from the UCI Machine Learning Repository, I used classification algorithms such as Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, and Gaussian Naïve Bayes to make my predictions.

I. Introduction

A. Background/Motivation

According to CancerIndex, cervical cancer accounts for around six percent of all cancers found in women. Cervical cancer is a disease where cancerous cells develop in the uterine cervix. The principal cause of most cervical cancers is the human papillomavirus (HPV). The peak incidence of cervical cancer is between the ages of 40-50 years old [1].

Although it is the most preventable type of cancer, each year cervical cancer kills about 300,000 women worldwide. Knowledge and access to screening techniques to detect cervical cancer in its early treatable stages is essential.

According to the ICO/IARC Information Centre on HPV and Cancer, in Venezuela cervical cancer is responsible for more fatalities among women between 15-44 years than any other form of cancer [2]. Venezuela has a population of 11.34 million women who are 15 years and older who are at risk of developing cervical cancer [2]. Cervical cancer is extremely rare in women who are younger than 20. However, young women become infected with multiple types of HPV, which can then increase their risk of getting cervical cancer in the future. Roughly about 35.2% of women in the general population are estimated to harbor cervical HPV-16/18 infection at a given time, and 79.2% of invasive cervical cancers are attributed to HPVs 16 or 18 [2].

The motivation for this project stems from my interest in healthcare. With the vast amount of biological data that we have at our disposal, there is an increasing need to analyze, understand, and manipulate it. As a computer science and engineering major and a bioengineering minor, I am interested in applying computing technologies and machine learning techniques towards healthcare.

II. Related work

After finding this dataset, I did some additional research to learn more about cervical cancer. Additionally, I thought that it would be useful to see if any studies were done using a similar dataset with health information in order to get a better idea of what machine learning models would be best.

I ended up finding three papers that used this dataset in their research.

The first paper, "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," used this cervical cancer dataset with a technique called transfer learning. Using the risk factor dataset and transfer learning techniques, they developed a predictive model that they then used to selectively transfer knowledge from one dataset to another. The knowledge that they gained was used to predict cross-modality individual risk and cross-expert subjective quality assessment of colposcopy images [3].

The second paper, "Determining Cervical Cancer Possibilities Using Machine Learning Methods" used machine learning models like k-Nearest Neighbors, Multilayer Perceptron, and Bayes Net to classify the patient as likely to have a high risk for cervical cancer [4].

The third paper, "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches," applied SVM, SVM-RFE, and SVM-PCA to analyze the cervical cancer dataset. Additionally, the authors used the RFE algorithm and PCA algorithm to reduce the computation burden and extract highly correlated risk factors [5].

A. Transfer Learning with Partial Observability Applied to Cervical Cancer Screening

In this paper, the authors instantiated their proposed partial transfer technique of the sign-transfer method to predict the individual patient's risk when multiple screening strategies are available. They utilized the same cervical cancer database. In order to fill in the gaps of missing information, they chose to fill the missing values with the sample mean. The authors applied regularized linear regression for the risk prediction task and Support Vector Machines for the quality assessment task [3].

B. Determining Cervical Cancer Possibilities Using Machine Learning Methods

This paper was published in the *International Journal of Latest Research in Engineering and Technology (IJLRET).* In this paper, the authors utilize the same dataset and improve on its classification using machine learning methods like k-Nearest Neighor (kNN), Multilayer Perceptron (MLP), and BayesNet. In their study, they used Percentage of Correctly Classified Instances (PCCI) as the performance indicator for their classification models. They divided the results into four groups,

- Correctly Classified Class 0 Instances, also called the True Negative Class 0 (TNC0)
- Falsely Classified Class 0 Instances, also called False Positive Class 0 (FPC0)
- Falsely Classified Class 1 Instances, also called False Negative Class 1 (FNC1)
- Correctly Classified Class 1 Instances, also called True Positive Class (TPC1)

The authors executed their classification algorithms on WEKA, an open-source data mining software that incorporates machine learning algorithms [???]. Their k-Nearest Neighbor approach varied the number of neighbors from 1 to 90. Out of their test dataset of 292 patients, it showed that as 280 instances were correctly classified with 86 neighbors. The correctly classified instance percentage was 95.89%. Their Multilayer Perceptron approach determined that presenting how many neurons in the hidden layer would provide the best results. Likewise, 280 instances were correctly classified with 28 neurons in the hidden layer. The correctly classified instance percentage was also 95.89%. Lastly, their Bayes Net approach correctly classified 284 patients, leading to a correctly classified instance percentage of 97.26%. The authors concluded that although the best classification was shown with the Bayes Net approach, the issue of predicting cervical cancer based on risk factors means that it is important to account for the number of falsely classified instances. This is because the number of instances that are classified as falsely positive mean the number of patients who have cancer but are not warned. [???] The authors determined the False Negative (FNC1) value for each machine learning model. The falsely classified instance rates were 1.37%, 1.71%, and 2.05% for k-Nearest Neighbors, Multilayer Perceptron, and Bayes Net, respectively. Ultimately, in the authors' opinion, this problem is best determined with the k-Nearest Neighbors method [4].

C. Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches

This paper was published in the Special Section on Data-Driven Monitoring, Fault Diagnosis and Control of Cyber-Physical Systems in IEEE Access [5]. According to the authors, in recent years, many detection methods were proposed and applied in the field to provide timely diagnosis, including data-driven approaches like principal component analysis (PCA), particle swarm optimization (PSO), fuzzy positivistic C-means clustering, linear regression (LR), artificial neural network (ANN), and support vector machine (SVM) [5].

III. Dataset

The dataset used for training and testing the machine learning models was found on Kaggle. It is from the UCI Machine Learning Repository. It focuses on the prediction of indicators/diagnosis of cervical cancer. The dataset was collected at the University of Hospital of Caracas in Caracas, Venezuela. It is comprised of demographic information, habits, and historic medical records of 858 patients. A note is that several patients decided not to answer some of the questions because of privacy concerns. This meant that there were missing values in the dataset.

As shown in Table 1, there are 35 attributes in total, such as age, number of sexual partners, age of first sexual intercourse, Boolean values of if the patient had certain STDs like HIV, HPV, and AIDS, Boolean values of it patients had been previously diagnosed with cancer, CIN (Cervical Intraepithelial Neoplasia), and HPV.

TABLE I
FEATURES ACQUIRED IN THE RISK FACTORS DATASET

Feature	Type	Feature	Type
Age	int	IUD (years)	int
# sexual partners	bool × int	STDs	bool × bool
Age of 1st sexual intercourse	bool × int	STDs (how many?)	int
# of pregnancies	bool × int	Diagnosed STDs	categorical
Smokes?	bool × bool	STDs (years since first diag.)	int
Smokes? (years & packs)	int × int	STDs (years last diag.)	int
Hormonal Contraceptives?	bool	Has previous cervical diag.?	bool
Horm. Contr.? (years)	int	Prev. cervical diag. (years)	int
Intrauterine device? (IUD)	bool	Prev. cervical diagnosis	categorical

IV. Methodology

A. Technologies Used

My project was written in Python 2.7.11. I chose to write it in Python because I first became familiar with it after taking COEN 169: Web Information Management. It is a popular language to use for machine learning. Specifically, I utilized scikit-learn, which is a free machine learning library. It has a lot of useful APIs that I

used for this project for preprocessing, feature selection, and creating my classification models.

B. Initial Steps

1) Target Variable Selection

Initially, I needed to manipulate my dataset. The first obstacle that I encountered was identifying the target variable. Amongst the papers I read, there were two different approaches to choosing the target variable. The first was to use the Boolean value of the biopsy screening results as the target variable. The second was to use the four screening methods, Hinselmann, Schiller, Cytology, and Biopsy, as the target variables. I chose to do the former, using the biopsy screening results as my target variable for classification. My rationale was that the biopsy screening would be an accurate indicator on whether or not the patient had a high risk of cervical cancer.

2) Missing Values in Dataset

In order to deal with the missing values in the dataset that the patients had omitted due to privacy concerns, I dealt with those by replacing all of the question marks with the column mean.

3) Preprocessing

I split my dataset up into 80% training and 20% testing. This seems to be the standard ratio of training data to testing data. Initially, I included the Hinselmann, Schiller, and Cytology attributes in my dataset. However, after performing some of the machine learning models, I noticed that my classification accuracy was 100%. This lead me to believe that Schiller, Hinselmann, and Cytology had the highest correlation with Biopsy. Thus, I removed those three screenings from my feature set, leading to a total of 32 features and 1 target variable. Then, I used a variety of preprocessing methods to observe if and how they would alter my results. I observed that in the dataset, while a majority of my attributes were Boolean, the integer values had a large range. So, I used feature scaling through standardization, also known as z-score normalization to rescale my features. Other preprocessing methods that I tried out were using MinMaxScaler, StandardScaler, Normalizer, and Binarizer to observe the changes.

For feature selection, to determine the features that had more importance, I used VarianceThreshold and SelectKBest.

B. Learning Techniques

1) Linear Discriminant Analysis (LDA)

The first technique that I implemented was linear discriminant analysis using the sklearn library. It is a classifier with a linear decision boundary that is generated by fitting class conditional densities to the data and using Bayes' rule with the assumption that the data has a Gaussian distribution. Also, it assumes that all classes share the same covariance matrix.

When using LDA, the classification accuracy score was consistently high. For LDA, I got the best accuracy score of 0.94767 by standardizing my dataset with zscore. I got the worst accuracy score of 0.18023 by binarizing my dataset. According to sklearn, binarized datasets are useful for downstream probabilistic estimators that make assumptions that the input data is distrubted according to a multi-variate Bernoulli distribution or are useful among the text processing community. My dataset fit neither of those criteria and thus was shown not to be a useful preprocessing methodology.

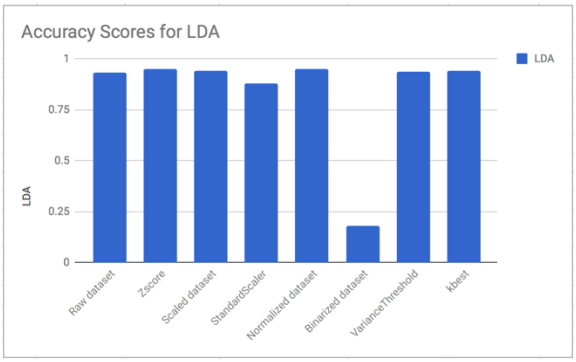

Fig. 1 Accuracy Scores for Linear Discriminant Analysis for Test Data

2) Quadratic Discriminant Analysis (QDA)

I implemented quadratic discriminant analysis using the sklearn library. It is a classifier with a quadratic decision boundary that is generated by fitting class conditional densities to the data and using Bayes' rule. It has the assumption that the data has a Gaussian distribution.

When compared to LDA, QDA had less consistent results overall, as shown in Figure 2. Its best classification accuracy score was 0.95349 through scaling the dataset with MinMaxScaler, StandardScaler, and featuring selecting the k best. However, the worst accuracies were shown to be 0.06395 when using the raw dataset, zscore preprocessing, and the normalized dataset.

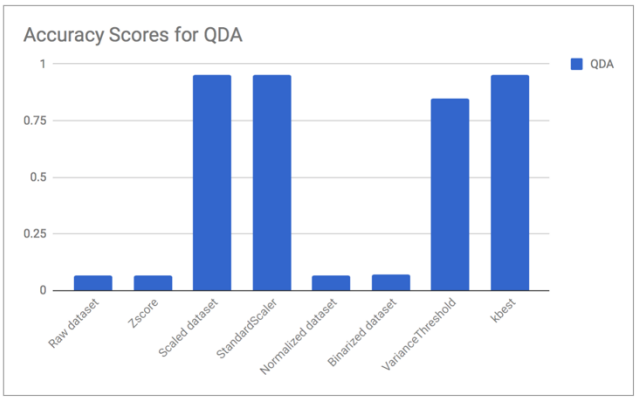

Fig. 2 Accuracy Scores for Quadratic Discriminant Analysis for Test Data

3) *Logistic Regression*

The third technique that I used was Logistic Regression. It is a classification algorithm that is used to assign observations to a discrete set of classes. Logistic regression transforms its output using the logistic sigmoid function to return a probability value, which can then be mapped to two or more discrete classes.

Logistic regression was shown to be the most consistent with its accuracy scores as depicted in Figure 3. The highest accuracy score was 0.95348 and the lowest was 0.94767.

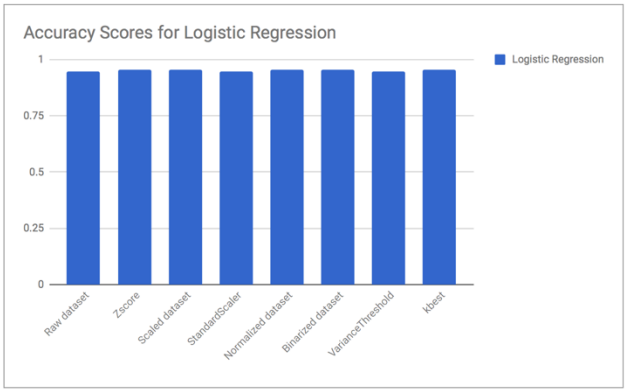

Fig. 3 Accuracy Scores for Logistic Regression for Test Data

4) *Gaussian Naïve Bayes*

Naïve Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naïve" assumption of independence between every pair of features. Gaussian Naïve Bayes have been famously used for document classification and spam filtering. Gaussian Naïve Bayes assumes that the dataset was generated through a Gaussian process with normal distribution.

Lastly, like QDA, Gaussian Naïve Bayes was inconsistent in its accuracy scores, as depicted in Figure 4. However, in contrast to QDA, its best accuracies were using preprocessing methods of zscore and StandardScaler and feature selection methods like VarianceThreshold, and kbest. Out of those, the best accuracy score of 0.95349 was achieved through StandardScaler. The worst accuracy was 0.18605 with normalized data.

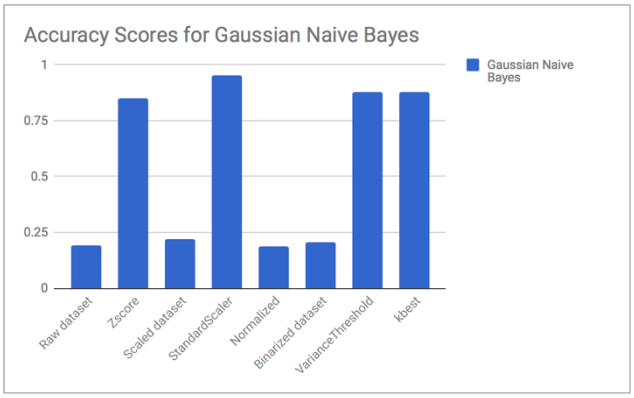

Fig. 4 Accuracy Scores for Gaussian Naïve Bayes for Test Data

IV. RESULTS

As shown in Table II, Logistic Regression and LDA had the most consistent accuracies. QDA and Naïve Bayes had the lowest accuracies. However, this could possibly be due to the scale of the data.

TABLE II
ACCURACY SCORE

Accuracy Score				
	LDA	**QDA**	**Logistic Regression**	**Gaussian Naive Bayes**
Raw dataset	0.930232558	0.0639534883721	0.947674418605	0.191860465116
Standardized dataset	0.94767441860	0.0639534883721	0.953488372093	0.848837209302
Scaled dataset	0.941860465116	0.953488372093	0.953488372093	0.220930232558
Normalized dataset	0.877906976744	0.953488372093	0.947674418605	0.953488372093
Feature selected dataset SelectKBest	0.941860465116	0.953488372093	0.953488372093	0.877906976744

For preprocessing, I found the most success over all of my models by standardizing my dataset using Standard Scalar, as shown in Figure 5. Out of all of the models using data that had been preprocessed with StandardScaler, LDA, shown in blue, had the lowest accuracy score of 0.8780 and Gaussian Naïve Bayes and QDA had the highest with 0.95349.

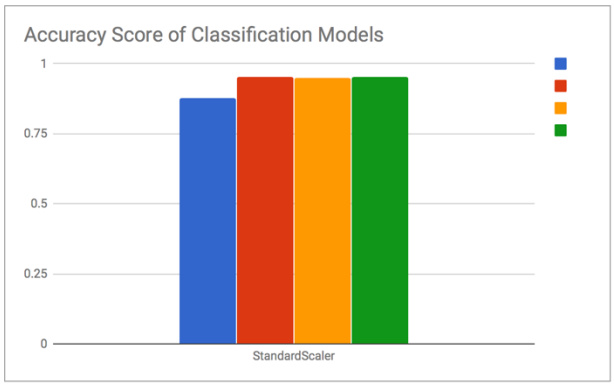

Fig. 5 Accuracy Scores for all of the Classification Models
(Blue is LDA, red is QDA, Yellow is Logistic Regression, and Green is Naïve Bayes)

V. Obstacles Encountered

As mentioned before, one of the obstacles that I encountered was figuring out how to deal with the missing values in the dataset. I implemented a fairly simple solution of replacing all of the "?" missing values with nan, finding the column mean, and replacing all of the nan values with the column mean.

The second major obstacle that I encountered was determining the target variable. Amongst the discussions on the Kaggle dataset page and the papers that I read, each study performed on the dataset used a different methodology and rationale behind determining the target variable. For example, one Kaggle user created a new target variable, CervicalCancer, and added the values of the four screening tests to represent the likelihood of cervical cancer.

V. Conclusion/Future Work

While some of the classification models provided some promising results, there is still work to be done on improving the classification accuracy and understanding the data.

First, I would like to test other machine learning models, specifically Support Vector Machine-based approaches, as well as use Principal Component Analysis to extract the risk factors that have the highest correlation.

Second, I would like to compare differences in classification had I used the four screening techniques, Hinselmann, Schiller, cytology, and biopsy, as my target variables instead of just the one. Admittedly, a single exam isn't enough to determine if a person has cancer or not. More medical exams and tests are required to determine the final diagnosis, since exams can yield false positives and false negatives.

Third, I would like to do more feature extraction to determine which features are the most important. Based on my research on cervical cancer, I anticipated that since the human papillomavirus was the principal cause of cervical cancer that it would be shown in my dataset as having a high correlation to cervical cancer. I would like to further explore this relationship.

References

[1] Anon. Key Statistics for Cervical Cancer. Retrieved March 24, 2018 from https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html

[2] Anon. 2017. Venezuela Human Papillomavirus and Related Cancers, Fact Sheet 2017. (2017). http://www.hpvcentre.net/statistics/reports/VEN_FS.pdf

[3] 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.

[4] Muhammed Fahri Ünlerşen and Kadir Sabanci. Determining Cervical Cancer Possibility by Using Machine Learning Methods. (December 2017). Retrieved March 23, 2018 from https://www.researchgate.net/publication/322233711_Determining_Cervical_Cancer_ Possibility_by_Using_Machine_Learning_Methods

[5] W. Wu and H. Zhou, "Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches," in *IEEE Access*, vol. 5, pp. 25189-25195, 2017. doi: 10.1109/ACCESS.2017.2763984

[6] L.H. Witten, I.H., E. Frank, and M.A. Hall, Data mining: practical machine learning tools and techniques. 2011, London: Elsevier.